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

The scientific community is faced with the challenge of managing large quantities of environmental and genomic data. Ontologies represent expert scientific knowledge in human machine readable formats. Ontology terms can be used to annotate interdisciplinary datasets stored in machine-actionable linked open data formats. Such practices allow data to be findable, accessible, interoperable and reusable to both humans and machine agents. Openly published environmental and genomic data often lack machine-focused accessibility, precluding integrated analyses from being performed. Here I show that ontologies are fit for purpose to address data management challenges by interconnecting disparate datasets with interoperable terminology. I demonstrated how datasets annotated with ontology terms can be mobilized via semantic querying to facilitate the assembly of data upon which to perform ecological analyses, however more open linked data is required to do so. I investigated the network properties of an example ontological knowledge graph. Although I found the upper level semantic model BFO provides a well-connected hierarchical structure, a lack of connectivity pervades the majority of ontology terms. This hinders ontology knowledge graphs from guiding users to novel information. Finally, I demonstrated how interoperable ontology terms can be leveraged to generate bioinformatic hypotheses from the comparison of environmentally annotated omics datasets. I anticipate these results will be a starting point for further analysis of how ontology terms can be used to interconnect environmental and genomic data. As well use ontology terms to improve the accessibility and reuse of published data.

Abbreviations

Acronyms

ANY23 Anything To Triples

AWI Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research

BFO Basic Formal Ontology DOI Digital Object Identifier

ECOCORE Ontology of Core Ecological Entities

ENVO Environment Ontology FRAM FRontiers in Arctic marine Monitoring

FAIR Findable Accessible Interoperable and Reusable

IACS International Association of Cryospheric Sciences

IAO Information Artifact Ontology

MIxS Minimum Information about any (x) Sequence NASA National Aeronautics and Space Administration

NOAA National Oceanic and Atmospheric Administration

OBCS Ontology of Biological and Clinical Statistics

OBO Open Biological and Biomedical Ontology

ORCID Open Researcher and Contributor ID

ORF Open Reading Frame

OWL Web Ontology Language

PATO Phenotypic Quality Ontology

PCA Principal Component Analysis

PCoA Principal COordinate Analysis

PCO Population and Community Ontology

RDF Resource Description Framework

RO Relations Ontology

SPARQL SPARQL Protocol and RDF Query Language

SWEET Semantic Web for Earth and Environmental Technology

UNESCO United Nations Educational Scientific and Cultural Organization

URI Uniform Resource Identifier

python scripts

py.1 create_rdf_triples_from_csv_files.py

py.2 histogram_median.py

py.3 merge_triples_to_datastore.py

py.4 network_distribution_stats.py

py.5 query_annotation_of_data_files_data_or_columns_about_input.py

py.6 query_data_set_references.py

py.7 query_for_class.py

py.8 query_for_classes_linked_by_input_classes_and_input_properties.py

py.9 query_for_created_by.py

py.10 query_for_data_about_exclusive_and.py

py.11 query_for_parts_associated_with_input_class.py

py.12 query_for_property.py

py.13 query_for_subclasses_of_input_purl.py

py.14 query_for_subproperties_of_input_purl.py

py.15 query_for_term_editor.py

py.16 query_GO_annotation_of_data_files_csv_annotations_columns.py

R scripts

R.1 data_set_references.r

R.2 degree_calculation.r

R.3 pca_analysis.r

R.4 pcoa_analysis.r

R.5 querying_exclusive_AND_annotations.r

R.6 query_parts_of_annotation.r

1. Introduction

The Earth system is facing unprecedented anthropogenic pressures which have the potential to destabilize critical biophysical systems, triggering irreversible environmental changes [1]. To address the broad scope of climate change interdisciplinary scientific action is required of the scientific community [2], unfortunately interdisciplinarity is not a prominent feature of climate research [3]. Hence, strategies to link complex and differentially structured data generated from interdisciplinary sources are needed [4]. In order to prevent large quantities of information and data from contributing to uncertainty rather than reducing it [5], strategies are needed to store environmental data in a way which captures subtleties of the data's structure, content and inter-relationships [4].

problem statment deluge of data incoming

Over the next decade, it is likely that science and engineering research will produce more scientific data than has been created over the whole of human history [6]

Currently scientific knowledge, information and data about environmental systems are published in wide variety of heterogeneous formats [7], which could be metaphorically described as a wild and untamed wilderness of knowledge. Scientific knowledge, information and data contained in this knowledge wilderness are commonly not interoperable, i.e. they lack the capacity to be used together [8]. Interoperability is a crucial component in the FAIR guiding principles for scientific data management and stewardship, and is required to make published data reusable to both human and machine agents [8]. In this work, I address the question of what can be done to integrate and prepare interdisciplinary earth systems knowledge, information and data to be machine-actionable, i.e. processable by computational systems in an automated fashion [9].

Semantics, i.e. the meanings of words or digital objects [8], are needed to make knowledge, information and data work interoperably at a machine-accessible level. Interoperable semantics can be used by machine agents to retrieve and analyze data relevant to a task of study [8], facilitating the process of knowledge synthesis. Knowledge representation a field of artificial intelligence makes use of machine-actionable knowledge, information and data, by integrating it into computer system intended to solve complicated tasks [10]. Knowledge representation systems have been successful used in medical domains to perform tasks such as computer-aided diagnosis of medical conditions [11].

Knowledge representation models can be constructed from expert knowledge which is represented in ontologies. An ontology is a hierarchically structured, machine and human readable semantic representation of the knowledge used by experts to describe entities in the universe, and capture the relationships between them [12]. 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).

In medical domains, ontologies have been used to interconnect disparate data and information, to enable computational interrogation of models to reveal underlying relationships. 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 [13].

Analogously to what has been done in medical domains, the use of federated ontology semantics have been discussed as having the potential to link data about phenotypes with environments [14], as well as environmental and genomic datasets [15]. This would allow for users to leverage knowledge, information and data connected by ontologies to ask questions such as "Which crop varieties are expected to do well in a particular location over the next century?" [14], or "Can we gather all metagenomes collected from insects found in soil?" [15].

Working toward the objective of using ontologies to interlink environmental knowledge, information and data to conduct future machine-focused ecological analyses, I evaluated the fitness for purpose of using ontologies to integrate interdisciplinary data.

In order to do this I created a model datastore consisting of various types of interdisciplinary polar and microbial omics data. The data in this datastore is annotated with ontology semantics and stored in a machine-accessible format to simulate a linked open data environment, i.e. a collection of data which can be accessed by machine agents. I made use of Polar data as the subject of the interdisciplinary data integration analysis, as polar systems are particularly vulnerable to the effects of climate change [16]. Subject to the effects of polar amplification, in which the effects of global warming produce larger temperature changes at the poles [17], polar regions are predicted to be free ice within 20 to 50 years [16]. In addition to the polar environmental data I also included environmental microbial omics data in my analysis to test of ontology knowledge-representation artifacts are fit for purpose to integrate interdisciplinary data.

To explore how ontology knowledge representations can fill the niche of improving informatics interactions in the wilderness of scientific data, I developed methods to test if ontologies are fit to interconnect interdisciplinary knowledge, information and data. To do so I developed competency questions, which are analogous to the questions asked by employers during job interviews to judge the competence of prospective employees [18]. These competency questions are specific, targeted and directed questions intended to evaluate how well ontology knowledge representation systems performs essential knowledge-representation and data-mobilization related tasks. Competency questions referred to in the text are abbreviated as CQ and are followed by the question number in parenthesis.

//in flowing prose and not a shopping list - introduce each competency question with a few lines describing why it's needed.

In semantic research, competency questions serve as testable hypotheses, by which to asses knowledge-representation models from a variety of angles.

… //maybe finish this up later??

Following the convention set by Buttigieg et al. (2013) ontology terms (alternatively referred to as classes) are written in italics with a prefix denoting the ontology namespace from which the class was sourced and in brackets the unique identifier of each term's OBO Foundry Uniform Resource Identifier following the term label [19]. For example the Environment Ontology term *sea ice formation process* is referred to as 'ENVO:*sea ice formation process* [ENVO_03000044]'.

//fix the first steps and mabye include here maybe not. The first steps toward using ontologies to integrate environment and genomic data have been undertaken via the creation of the Earth Microbiome Project application Ontology (EMPO). Built as subset of the Environment Ontology, EMPO provides the semantics by which to map environmental 16S tag sequence data to descriptions of environmental features [20].

The first and possibly most important task which I tested the competency of ontologies to perform was to see if they could be used to interconnect environmental and omics data. Linking environmental and omics data to uncover interactions between genomes and environments is a topical challenge in genomic and medical disciplines [21]. In a recent study linking environmental and omics data, Favé et al. (2018) were able to determine that environmental factors (air pollution) were responsible for disease risk phenotype outcomes rather than ancestral genotypes [21].

Using ontologies to interconnect genomic and environmental data would allow for machine agents to act upon large environmental genomic datasets such as the TARA Oceans project [22], the Global Ocean Sampling Expedition [23] and the Hawaii Ocean Time-series program [24]. Systematic machine operation over such a wealth of knowledge has the potential to help answer future ecological questions.

In principal ontologies should be able to interconnect environmental and genomic data. One of the most widely used ontologies in biomedical domains is the Gene Ontology (GO), which provides semantic representations describing the roles of genes and gene products [25]. The Gene Ontology along with other ontologies are part of the Open Biological and Biomedical Ontology (OBO) Foundry and Library [12]. These ontologies use common upper level semantic models and shared relations so they can work interoperably, creating a unified multidisciplinary knowledge representation model [15]. One such OBO ontology the Environment Ontology (ENVO) provides semantic descriptions of environments [19][26]. Although ENVO and GO make use of interoperable semantics, efforts to use these ontologies in combination, have infrequently been attempted [27]. Assessing if ontologies can be used to interconnect environmental and genomic data in order to compare genetic differences between environments, I formulated my first competency question:

> "What are the relative proportions of oxidation-reduction process genes in various types of marine biomes?"
>
> $$CQ(1)$$

This question evaluates a variety of ontology competencies. First it tests if it is possible to use ontologies to interconnect genomic and environmental data, retrieving data which is about a GO:*oxidation-reduction process* [GO_0055114 as well as an ENVO:*marine biome* [ENVO_00000447]. The question doesn't just test if there is a specific type of GO data about ENVO:*marine biome* [ENVO_00000447], but it tests if the knowledge contained in the knowledge hierarchy about different types of ENVO:*marine biome*s [ENVO_00000447] can be leveraged to determine if any of those subclass have information about this specific GO term. Simulating the kinds of questions a microbial ecologist would ask about gene which differentiate various ENVO:*marine biome*s [ENVO_00000447] I asked:

> "What are the relative proportions of vitamin biosynthetic process genes in various types of marine

biomes?"

$$CQ(2)$$

This question is intended to further explore if ontologies can be used to facilitate the collection of data by which to perform comparative genomic analyses. The results of the second question prompted further analysis leading me to ask:

> "What genomic features may help to explain the differences in riboflavin abundances between deep and shallow marine benthic biomes?"

$$CQ(3)$$

To answer this I explored terms from higher level GO hierarchies of GO*molecular_function*s [GO_0003674] and GO:*biological_process* [GO_0008150] to find information relevant to transition metal binding and transport functions.

I wanted to test what other kinds of expert knowledge were present in the higher level GO ontology GO:*biological_process* [GO_0008150] hierarchy which may help to differentiate ENVO:*marine biome*s [ENVO_00000447] samples. Hence I created a competency question asking:

> "What biological processes differentiate various types of marine benthic biomes?"

$$CQ(4)$$

As there is quite a lot of expert knowledge contained in the upper level GO:*biological_process* [GO_0008150] hierarchy I wanted to ask a more specific question about a subclass of GO:*biological_process* [GO_0008150], GO:*cellular amino acid biosynthetic process* [GO_0008652]:

> "What cellular amino acid biosynthetic processes differentiate various types of marine benthic biomes?"

$$CQ(5)$$

Drawing from the results of CQ (5) to further leverage the GO biological processes hierarchy to see if specific GO terms could be found which are are completely different in ENVO:*marine biome*s [ENVO_00000447] samples, I asked:

> "What serine family amino acid biosynthetic processes differentiate various types of marine benthic biomes?"

$$CQ(6)$$

Although the interconnection of genomic and environmental data is an important competency by which to assess ontologies for their ability to interconnect interdisciplinary data, it is far from being the only use case for ontologies to interconnect knowledge, information and data. I created the following question to assess the fitness

of ontologies to be able to navigate machine-accessible information to discover data relevant to a phenomena of interest. I tested this with the competency question asking:

> What data about environmental factors with the potential to influence the dynamics of a sea-ice associated phytoplankton community could be collected using an ontology?

$$CQ(7)$$

Discovery of data is an important task for ontologies to be able to perform, similarity I wanted to assess if ontologies can also be used by scientists to discover new knowledge from a knowledge graph which is related to stated input knowledge. Operating on the envoPolar subset of ENVO I asked:

> "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?"

$$CQ(8)$$

Ontology-guided knowledge-discovery is an important competency by which to assess the fitness for purpose of ontologies to interconnect interdisciplinary knowledge.

The following questions address if ontologies are fit for purpose to identify the provenance of knowledge data or other information:

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

$$CQ(9)$$

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

$$CQ(10)$$

tracking provenance is good …, are the ontology knowledge models also resilient to changing semantic models.

> "Are ontology knowledge representations resilient to the input of new information which may contain definitional discrepancies"

$$CQ(11)$$

I created the following question to assess the resilience of semantic data annotation models, asking:

> "What percentages of data items discovered to be about participants in sea ice formation processes would be retrievable if changes were to be made to the underlying semantic models used by OBO

ontologies, such as not using hierarchical subclass structures to represent knowledge, or not using structured relations from the Relations Ontology?"

$$CQ(12)$$

The following question addresses the extent to which semantically annotated data is discoverable to general users.

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

$$CQ(13)$$

Finally, to assess if the process of encoding expert ecological knowledge into an ontology knowledge graph aids to clarify the understanding of ecological phenomena, I asked:

"Does the inclusion of novel expert knowledge about phenomena relating to plankton ecology into the ENVO knowledge graph aid to better understand the interconnections of such phenomena?

$$CQ(14)$$

Unprecedented quantities of data and information generated as a result of the information revolution have arguably made the world more uncertain complex and ambiguous than less [5]. As a consequence, challenges we face tend to be dilemmas requiring management, rather than easily defined problems to solve [28]. Coming technological advances will further accelerate the rate of ecological data capture [5]. Lacking well-established repositories or standard protocols for the management of ecological data [4], such influxes of data have the potential to contribute rather than detract from uncertainty about the natural world. The Earth system is facing unprecedented anthropogenic pressures which have the potential to destabilize critical biophysical systems, triggering irreversible environmental changes [1]. Needed are strategies to store ecological data in a way which captures subtleties of the data's structure, content and inter-relationships [4]. This holds especially true for vulnerable and rapidly changing environments, such as polar systems which have been predicted to be free ice within 20-50 years [16]. Despite the existence of numerous environmental monitoring efforts such as AtlantOS [29], the Fixed point Open Ocean Observatory network (FixO3) [30], or the Hausgarten Long-Term Arctic Observatory [31]. The management and integration of generated data remains a major obstacle, precluding integrated analysis from being performed on interdisciplinary data. Working toward the improvement of infrastructure for scholarly data publication, Wilkinson et al. (2016) have proposed the FAIR data guiding principles of data findability, accessibility, interoperability and reusability. These principles aim to promote the publication of data which is accessible to both humans, as well as machines [8]. Although many ecological datasets, such as those generated by observatories, have been published to openly accessible repositories such as PANGAEA [32] or the Biological and Chemical Oceanography Data Management Office (BCO-DMO) [33]. Data contained within such repositories are typically not machine-readable, or interoperable. In order to make annotated data work interoperably, annotations need to make use of controlled, universally shared, and machine accessible vocabularies. Such annotation terms can be provided by

ontologies.

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 [12]. 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 typically developed to cover the terminological needs of a particular domain of interest. In order to interconnect ontologies representing scientific knowledge from different domains, as well as coordinate their development, the Open Biological and Biomedical Ontology (OBO) Foundry and Library was created [12]. OBO Foundry ontologies share common upper level semantic models. Notable the Basic Formal Ontology (BFO) [34][35] providing common hierarchical structures by which to characterize knowledge, and the Relations Ontology (RO) [36], to standardize the connections between represented knowledge. Existing OBO ontologies such as the Environment and Gene Ontologies, for the description of environments [19][26] and genetic functions [25] are designed to work interoperably. Efforts to use these ontologies in combination, however, have infrequently been attempted [27].

Ontologies make use of Semantic Web technologies to interconnect information. The Semantic Web is an extension of the World Wide Web, which provide common data formats and exchange protocols on the Web [37]. The Semantic Web stores data in a format by which the meaning of the data is processable by computers, referred to as machine-readable data [37]. Data is rendered machine-readable by converting it into a format compliant with the Resource Description Framework (RDF) [38]. Such formatting allows data to be linked to other data, stored in open formats. Linked data can be accessed through semantic queries, performed using the SPARQL Protocol and RDF Query Language (SPARQL) [39].

Within medical domains ontologies have been used to interconnect disparate data and information, enabling computational interrogation of models to reveal underlying relationships. For example, 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 [13]. Although ontologies have not yet been utilized to fully integrate environmental and omics data, efforts to such extent are underway [40]. Analogously to what has been done in medical domains, the use of federated ontology semantics have been discussed as having the potential to link data about phenotypes with environment [14], as well as environmental and genomic datasets [26]. This would allow for users to submit queries of linked data asking questions such as "Which crop varieties are expected to do well in a particular location over the next century?" [14], or "Can we gather all metagenomes collected from insects found in soil?" [15].

The first steps toward using ontologies to integrate environment and genomic data have been undertaken via the creation of the Earth Microbiome Project application Ontology (EMPO). Built as subset of the Environment Ontology, EMPO provides the semantics by which to map environmental 16S tag sequence data to descriptions of environmental features [20]. Future work is required to integrate large publicity available genomic sequence datasets with consistently structured accompanying environmental metadata. Examples of such datasets include the TARA Oceans project [22], the Global Ocean Sampling Expedition [23] and the Hawaii Ocean Time-series

(HOT) program [24]. Unfortunately, the data publication formats of such datasets often preclude them from being semantically querable. Efforts to make such environment sequencing data semantically querable are underway by research groups such as the Cyberinfrastructure for Microbial Ecology [41], who have developed tools such as the Ocean Cloud Commons (OCC), which allows for researchers to query the Tara Oceans Expedition Data [42]. With future projects aiming to integrate publicly available genomic and oceanographic data [43]. Such efforts would benefit from the use of well-structured ontology semantics by which to mobilize and interconnect environmental and genomic data.

In as far as integrating environmental observatory data, 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 [44]. These efforts are relevant to the upcoming United Nations decade of ocean science for sustainable development 2021-2030 [45]. Which will bring 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 [46]. The ongoing efforts of the Open Geospatial Consortium (OGC) have seen the creation and development of community consensus standards for the Sensor Web [47]. 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 [48]. SWE standards enable developers to make all types of sensors, transducers and sensor data repositories discoverable, accessible and usable via the Internet. An example SWE standard is the Sensor Observation Services (SOS), which provides interoperable management of sensor data, allowing users to query the Sensor Web for real-time sensor data [49]. Extending the SWE with Semantic Web technologies, is the Semantic Sensor Web (SSW) providing enhanced descriptions and access to sensor data using the Semantic Sensor Network Ontology (SSN) [50]. Ontologies such as the SSN will be important for mobilizing the coming large data generated by sensor networks. Especially if the SSN ontology is made to work interoperably with other ontologies describing environmental phenomena such as the Environment Ontology [19][26] and the suite of Semantic Web for Earth and Environmental Technology (SWEET) ontologies [51].

With future influxes of data from sensor monitoring endeavors along with the ever increasing quantities of genomic data, there is a clear need to find ways to facilitate the management incoming and existing data. This work is motivated by the need to need to overcome barriers preventing the reuse of published data [52], as well as facilitate the publication if data which can be used interoperably.

As well-structured semantics, such as those provided by ontologies benefit the reuse and interoperation of published data, I focused my research on the study of semantics. In semantic research knowledge is treated as the subject of study. In my work, I have applied the scientific method to conduct semantic research. I begin by formulating hypotheses about the representation of knowledge. I have done so by formulating competency questions, as is done in interviews to judge the competence of prospective employees [18]. Competency questions are specific and targeted questions which are intended to evaluate how well a person or a system is able to perform certain essential functions. In this work I have created a set of competency questions to judge the fitness for purpose of using ontologies to interconnect interdisciplinary knowledge and data from environmental and genomic domains.

Following the scientific method, competency questions are created to serve as testable hypotheses. These competency question based hypotheses can be tested by using ontologies. The results derived from testing competency questions can provide insights by which to formulate and test additional competency question hypotheses. For a description of the adaptation of semantic science and linked open data to facilitate the scientific method, see **Figure A1**.

Ontologies can be used directly to answer competency questions, by searching for relevant knowledge encoded within ontology knowledge graphs. Alternatively, we can test competency questions by using ontologies to gather data and other machine-accessible information from open linked data. This is made possible by setting up linked open data repositories where data is annotated using ontology terms. These open linked ontology annotated data repositories allow for data to be accessed via semantic queries for ontology terms of interest. To be able to ask and answer such competency questions, I have created an example open linked datastore containing environmental and genomic data. For the semantic annotation of such data I made use of ontology terms from the OBO Foundry and Library, which offers a multidisciplinary unified knowledge model spanning various scientific domains [15]. By using the BFO [34] and RO [36] 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.

To interconnect my example data as well as access information within the OBO knowledge graph, I developed scripts which are able to retrieve classes pertaining to a phenomenon 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 the interdisciplinary datastore. In this work I employed my semantically-annotated datastore and ontology querying scripts to answer the competency questions I created. These competency questions are intended to assess the fitness for purpose of ontologies to interconnect interdisciplinary environmental and genomic datasets. The following is a description of the competency questions I asked and attempted to answer in the course of this work.

2. Materials and Methods

2.1 Model polar datastore creation

I created

I assembled an example polar themed datastore from freely available Alfred Wegener Institute for Polar and Marine Research datasets hosted by the data publisher PANGAEA [32], as well as data from unpublished metagenomes and metatranscriptomes. I selected data sets primary from the FRAM [53], and Hausgarten [31] observatory projects and programs. I additionally made use of 16S taxonomic and Pfam2GO annotation tables provided courtesy of Josephine Z Rapp, David Probandt, and Matthew Schechter. Such metagenomes and metatranscriptomes samples had previously been processed using the BBDuk tool (Version 35.68 for metagenomes, Version 36.92 for metatranscriptomes) from the BBTools suite for read quality control [54], RNA filtered using SortMeRNA (Version 2.0, 29/11/2014) [55], taxonomically classified for 16S ribosomal RNA with the SINA alignment tool (Version 1.2.11, revision 21227) [56], ORF-predicted via FragGeneScan tool (Version 1.20) [57], ORF-annotated with PFAM domains via the Ultra-fast Protein domain Classification tool (UProC) (1.2.0) [58], and mapped to GO terms using the Pfam2GO annotations page (January 2016 version) available from http://www.geneontology.org/external2go/pfam2go [59]. A list of data sets included in the data store is provided in the appendix **A.2 Model polar datastore creation**. The example datastore was created by converting comma separated value (csv) files into the RDF specification [60] turtle format [38], using the script py.1, which makes use of the Apache Anything To Triples (any23) command line *rover* tool (Version 2.1) [61]. Turtle formatted data along with semantic data annotation files were merged into a single datastore file using the script py.3.

2.2 Semantic data annotation

Semantic annotation of the example data was conducted in the RDF serialization turtle facilitate scripting Web Ontology Language (OWL) [62] code in RDF. Annotations make use ontology terms from the Open Biomedical Ontologies (OBO) Foundry and Library [12]. Ontology terms were accesses using the Ontobee linked data server for ontologies and their terms [63]. 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* [64], *Polar express* [65], and *Hot tub time machine* [66] ENVO releases.

Figure 2 Shows an example of the post-compositional data annotation of datasets using ontology terms. The example is of a data column which is about a *chlorophyll a concentration* in *sea water*. Expressed as an ontology axiom, we have a data column: which is BFO:*part of* [BFO_0000050] some OBCS:*data matrix* [OBCS_0000120] and IAO:*is about* [IAO_0000136] some PATO:*concentration of* [PATO_0000033] and RO:*inheres in* [RO_0000052] some CHEBI:*chlorophyll a* [CHEBI_18230] and BFO:*part of* [BFO_0000050] some ENVO:*sea water* [ENVO_00002149].

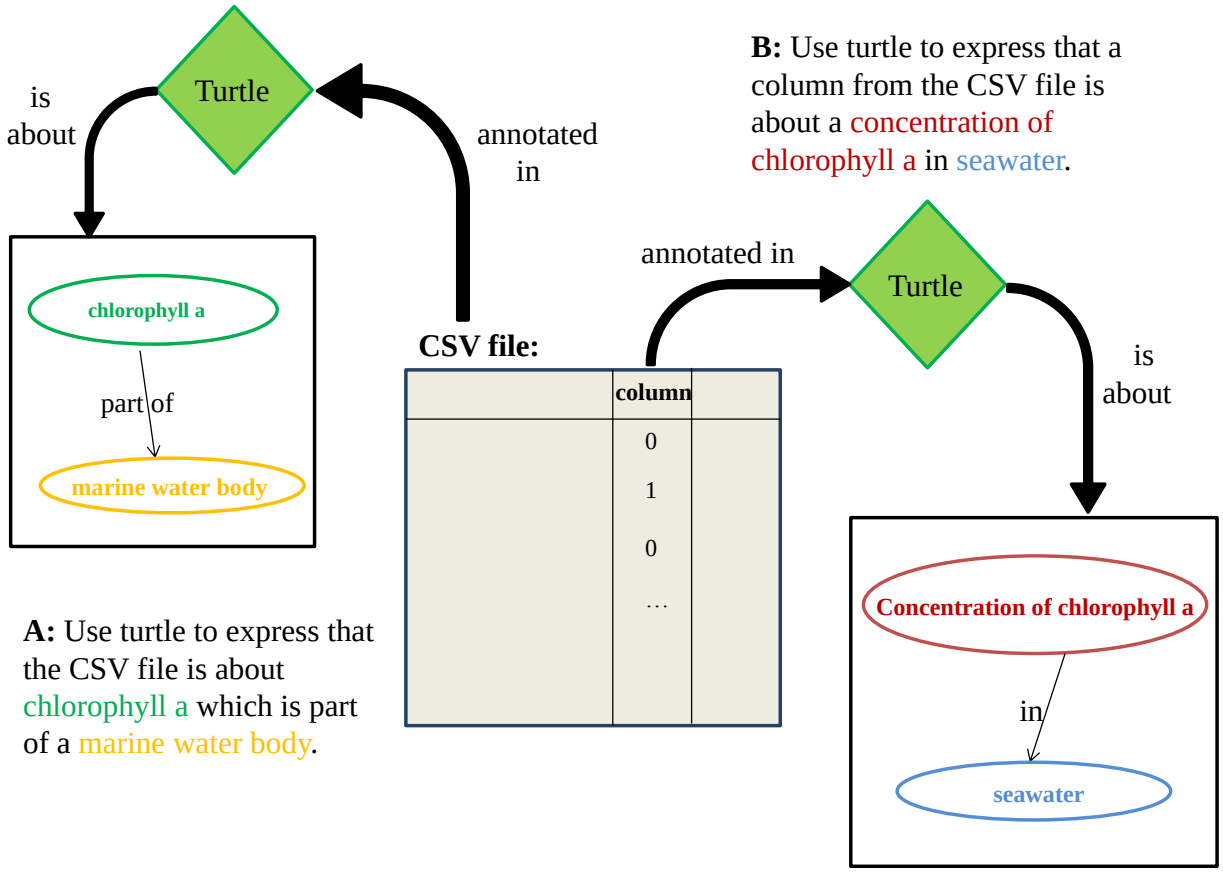

Figure 1 Shows content from a CSV file which has been semantically annotated in the RDF serialization turtle. Item **A** shows how the CSV file is annotated in turtle as being about chlorophyll a which is part of a marine water body. Item **B** shows how a column from the CSV files is annotated in turtle as being about a 'concentration of chlorophyll a in seawater'.

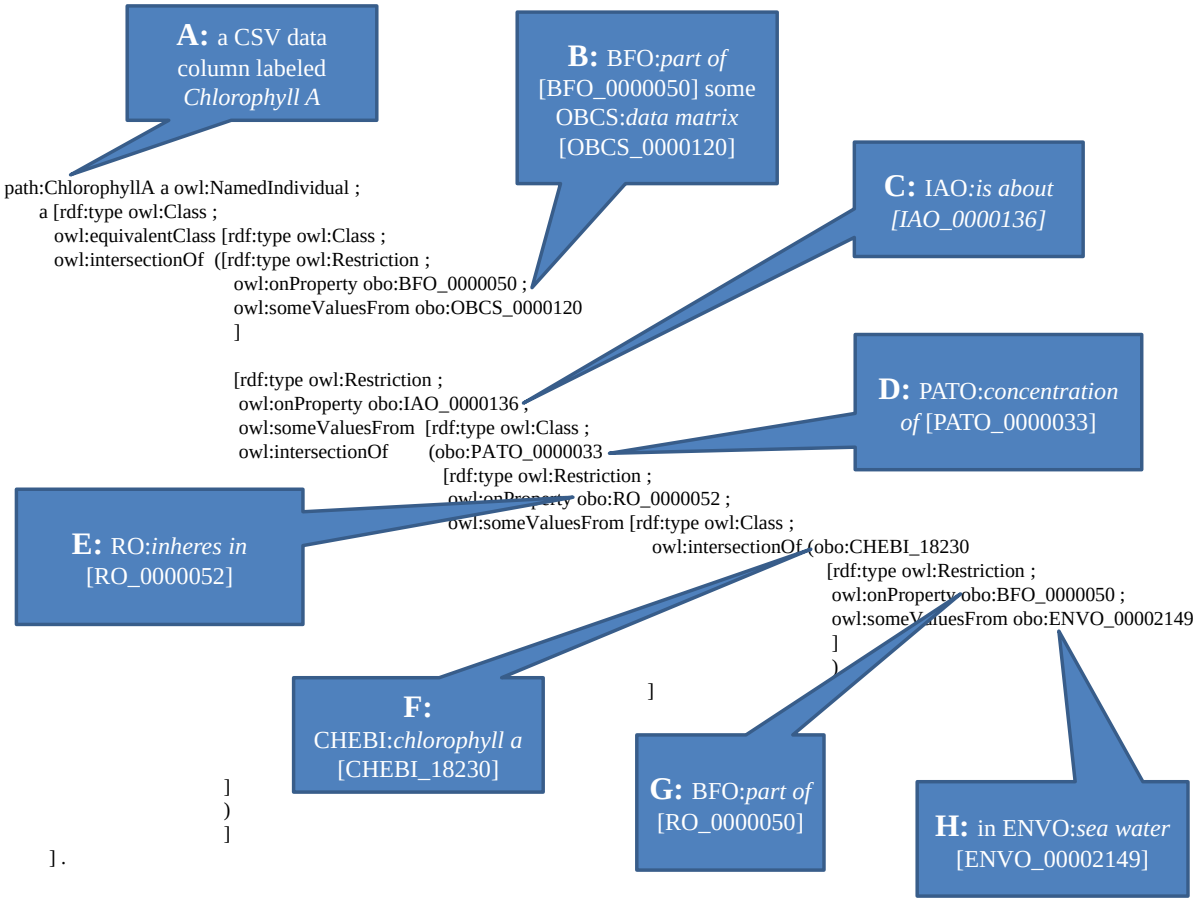

Figure 2 shows the RDF specification turtle code to annotate a data column from a csv file. Item **A** shows the csv file data column is labeled chlorophyll A. Item **B** shows the data column is BFO:*part of* [BFO_0000050] some OBCS:*data matrix* [OBCS_0000120]. Item **C** shows the data matrix column IAO:*is about* [IAO_0000136] more a list of annotation terms. Item **D** shows object is about a PATO:*concentration of* [PATO_0000033]. Item **E** shows the concentration measured is of CHEBI:*chlorophyll a* [CHEBI_18230]. Item **F** shows the *chlorophyll a* is BFO:*part of* [BFO_0000050] some ENVO:*sea water* [ENVO_00002149].

2.3 SPARQL querying

Scripts were written in Python (Version 2.7.12) [67] and make use of the rdflib module (Version: 4.2.2) an RDF parsing python library [68] 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/ [63]. A demonstration of the process by which a SPARQL query is executed is shown in **Figure 3**. Item **A** shows a SPARQL query, which is assembled in a python script based on some user input class (shown in green). This query selects some unknown thing called ?X (shown in red), where ?X is a subclass of some input class, and ?X is part of some class Y (shown in orange). Item **B** shows a knowledge graph of ontology terms and the relations connecting them. The knowledge graph could either be a local copy of an ontology such as ENVO or a web accessible ontology knowledge graph such as the Ontobee programmatic SPARQL endpoint. Item **C** shows how the SPARQL query navigates the relations in the knowledge graph to find a term which satisfies the input conditions. Item **D** shows the successful retrieval of the ?X term, which was discovered from the knowledge graph.

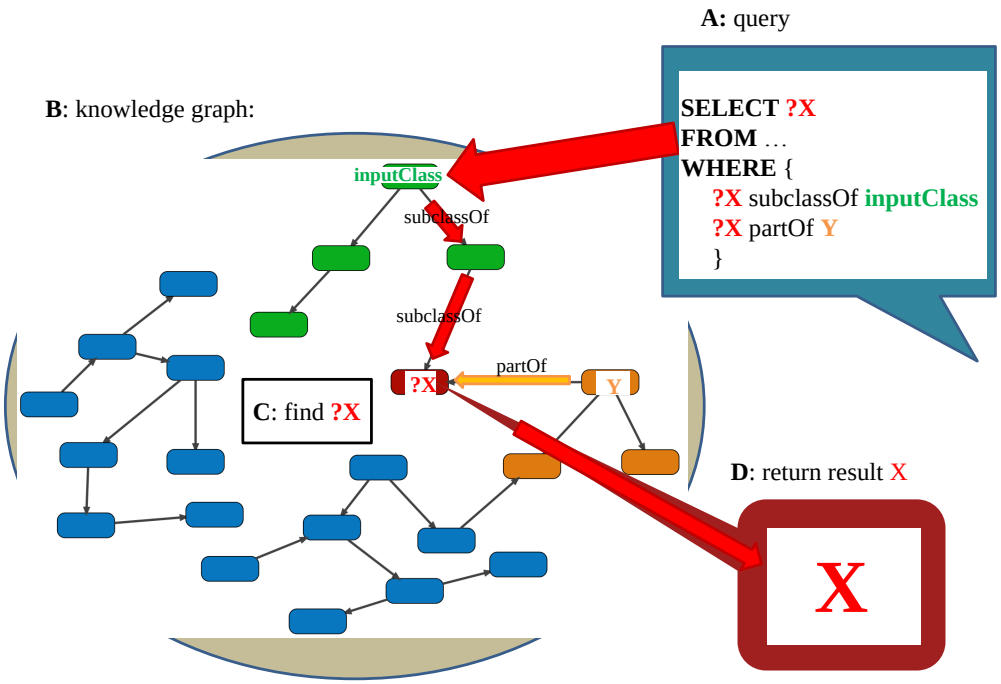

Figure 3 Shows a demonstration SPARQL query. In item **A** a SPARQL query is assembled. Item **B** represents a knowledge graph of ontology terms and their interconnecting relations. Item **C** shows the process by which the graph is navigated to find some ?X unknown item which is both a subclass of some input class and part of some Y. Item **D** Shows the end result of the query, returning the ?X class of interest having discovered it from the knowledge graph.

2.4 Interconnecting genomic and environmental data via ontologies

Retrieval of GO relative abundance data from ontology term annotated data was conducted using the script py.16. The script makes 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 script py.13. The script py.16 queries the local datastore for data matrices 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) [69]. The script returns a data table with rows corresponding to samples, and columns corresponding to relative abundances of GO terms. 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, GO:*biological process* [GO_0008150] or GO:*molecular function* [GO_0003674], then dividing the relative abundances of individual terms by the sum of the total abundances of biological process or molecular function terms. Principal coordinate analysis (PCoA) on a data matrices of relative GO term abundances in ENVO annotated samples was conducted in R [70] (Version 3.3.2), in the script R.4, which uses the vegan package [71] version 2.4-3. Prior to PCoA analysis data was standardized using the decostand function making use of a Hellinger tranformation, then converted into a Bray-Curtis dissimilarity matrix using the vegdist function with standard parameters.

2.5 Ontology guided data assembly for ecological analysis

Subclasses of ontology terms included in the axioms of the hypothetical ENVO:*environment determined by a phytoplankton community associated with sea-ice* term, were assembled using the py.13 script, and concatenated together. Data annotated as being about this assembled list of subclasses terms was retrieved from the datastore using the py.5 script. In the script R.3 a principal component analysis (PCA) was performed on the retrieved data using the vegan package [71] version 2.4-3. Prior to PCA analysis, data was z-score standardized using the scale function with parameters: *center* and *scale* being set to *TRUE*. The hypothetical term ENVO:*environment determined by a phytoplankton community associated with sea-ice* follows the design pattern of several pre-existing terms in the Environment Ontology such as ENVO:*environment determined by a biofilm on a fungal surface* [ENVO_01001035] within the ENVO:*environmental system* [ENVO_01000254] hierarchy.

2.6 Connecting ontology term information with term authors

Percentages of ENVO and envoPolar ontology terms annotated with a IAO:*term editor* [IAO_0000117] or oboInOwl:*created_by* [created_by] relation referencing an ORCID were calculated using the following workflow. Local turtle specification versions of the ENVO and envoPolar v2017-08-22 *Planetary ecology* release [72] ontologies were exported to a local version formatted in the RDF turtle serialization using the standard Protégé [73][74] `export as` function. Python scrips py.7, py.9 and py.15 were used to perform SPARQL queries upon the local turtle ontology versions, for total numbers of ENVO terms and numbers of terms annotated with an IAO:*term editor* [IAO_0000117] or oboInOwl:*created_by* [created_by] relation an ORCID. Percentages of

ontology terms with IAO:*term editor* [IAO_0000117] or oboInOwl:*created_by* [created_by] annotations were calculated from retrieved counts.

The demonstration of how to retrieve the current author contact information from an ORCID, was conducted as follows. A two-legged OAuth authorization request was sent to the ORCID application programming interface sandbox demonstration contact information retrieval service, issuing a token request for the information associated with the demonstration ORCID client identifier `APP-NPXKK6HFN6TJ4YYI`. Such request was sent from the linux command line using the curl data transfer tool (Version 7.47.0) using the following line of code:

```
curl "Accept: application/json" -d "client_id=APP-NPXKK6HFN6TJ4YYI" -d "
   client_secret=060c36f2-cce2-4f74-bde0-a17d8bb30a97" -d "scope=/read-public
   " -d "grant_type=client_credentials" "https://sandbox.orcid.org/oauth/
   token"
```

Making use of the returned authentication token `2bd6d6b7-9438-4a5a-8f87-7e43d6eaac25`, a request for the email contact information associated with the demonstration ORCID client identifier was sent with curl (Version 7.47.0) using the follows line of code:

```
curl -i -H "Accept: application/vnd.orcid+xml" -H 'Authorization: Bearer 2
   bd6d6b7-9438-4a5a-8f87-7e43d6eaac25' 'https://api.sandbox.orcid.org/v2
   .0/0000-0002-9227-8514/email'
```

From the returned XML code block, the email address associated with the demonstration ORCID client identifier was parsed using the GNU grep command line tool (Version 2.25), resulting in the retrieval of the email address associated with the demonstration ORCID client identifier.

2.7 Connecting datasets and publications about an ontology term

The following was used to retrieve digital object identifiers (DOIs) of publications associated with datasets about parts of a marine biome. The python script py.11 was used to query for ontology classes which are parts of or have parts which are an ENVO:*marine biome* [ENVO_00000447]. The resulting list of ontology terms which are parts associated with ENVO:*marine biome* [ENVO_00000447] were queried against the local datastore for all data from data matrix columns which are annotated with an oboInOwl:*hasDbXref* [hasDbXref] database cross reference using the py.6 python script. Results were processed in script R.1.

2.8 Glacial semantics community consultation and creation of proposed ontology terms

Participation in the Feb 2, 2018 VoCamp Glacier Ontology Hackathon community glacial-consultation session [75] was an opportunity to consult with polar-related domain experts to improve the semantic representation of glacial-related knowledge. Drawing upon knowledge sourced from this event, as well as a variety of scientific publications from AWI polar domain experts [76][77], I created ontology terms which I have proposed to be

added the ENVO [19][26] PATO [78] ECOCORE [79] and PCO [80] ontologies. These potential ontology terms were created following the best practices for BFO ontology development as outlined by Arp et al. (2015) [34], making use of the BFO to provide common hierarchical structures by which to characterize knowledge, and the RO [36], to standardize the connections between represented knowledge.

2.9 Classes linked by subclasses and subproperties

Retrieval of classes participating in subclasses of sea ice formation processes utilizing upper level semantic models was conducted as follows. Subclasses the term ENVO:*sea ice formation process* [ENVO_03000044] were assembled using the py.13 script. Subproperties of RO:*has participant* [RO_0000057] were assembled using the py.14 script. The results of classes which participate in sea ice formation processes, were discovered using the py.8 script with the assembled subclasses of ENVO:*sea ice formation process* [ENVO_03000044] and subproperties of RO:*has participant* [RO_0000057]. This workflow of querying for subclasses, subproperties and using them to search the knowledge graph for new classes of interest is documented in **Figure 4**.

This workflow was used to simulate what would happen if my example datastore didn't use the BFO and RO upper level semantic models. These simulations were evaluated for the extent to which annotated data would be retrieved by querying the example polar datastore in the absence of either BFO or RO. Lacking the RO the script py.13 was used to find subclasses of ENVO:*sea ice formation process* [ENVO_03000044], and the results of which were directly passed to script py.5 to get data about these subclasses, without using script py.8 using RO relations to discover new classes. The simulation for the retrieval of data lacking the BFO upper level semantic model was conducted using the same methods as the main workflow, minus the query for subclasses of ENVO:*sea ice formation process* [ENVO_03000044] with only the term itself passed to the py.8 script. The simulation for the retrieval of data lacking the Relations Ontology upper level semantic model was conducted without the py.8 script. In which the subclasses of ENVO:*sea ice formation process* [ENVO_03000044] from the py.13 script were directly passed to the py.5 script.

2.10 EnvoPolar network creation

The following was used to create a network from the ENVO polar subset. The envoPolar subset from the ENVO v2017-08-22 *Planetary ecology* release [72] was used. The envoPolar owl file was exported to the RDF specification turtle using the software Protégé's standard `export as` function [73][74]. Python script py.7 was used to query for all classes and python script py.12 was used to query for all property terms in the envoPolar subset. The results of which were used as inputs for the py.8 script to obtain the connections between all classes in the ontology. The resulting output consisting of subject classes, properties linking subject classes to target classes, and target classes, which was used to create a network in the program cytoscape [81].

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.

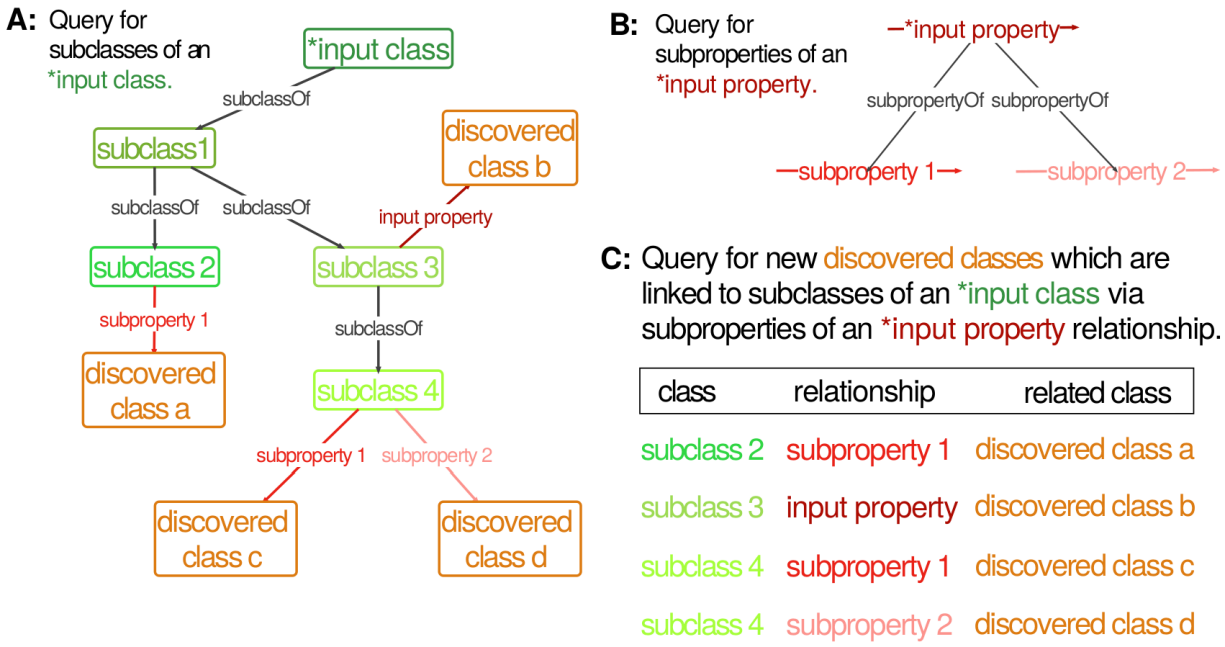

Figure 4 Shows a demonstration of the workflow pertaining to script py.8. Item **A** shows an ontology knowledge graph, in this case the merged Ontobee programmatic SPARQL endpoint. In item **B** the py.13 script is executed to retrieve subclasses of an input class. In item **C** the py.14 script is executed to retrieve subproperties of an input property. In item **D** the py.8 script takes the list of subclasses generated in script py.13, the list of subproperties generated by script py.14. Searching the Ontobee knowledge graph script py.8 discovers new classes which are linked to subclasses of interest by the subproperties relations of interest.

Calculations of mean and median values for the in-degree, out-degree and shortest path length distributions were conducted in python the python script py.2 using the statistics library Version: 1.0.3.5 [82] from distribution data output by the cytoscape network analysis. Figures of the in-degree, out-degree and shortest path length distributions were created in the R script R.1.

Nodes of highest and lowest in-degree value were extracted manually from the envoPolar network in cytoscape.

2.11 Predicted user data mobilization

In this thesis, I lacked the scope to be able to conduct an experiment on a study group of scientists with various proficiencies for performing semantic queries to retrieve data from the example datastore. In its place I estimated the performance of predicted user stories, as is done in agile software development [83]. I created three categories of predicted user stories, for users of various SPARQL querying proficiencies. The predicted users were modeled to have basic, intermediate or advanced querying expertise. I evaluated the performance of these predicted user stories based on the percentage of total data each would be able to retrieve from our example datastore, when performing a variety of queries.

The basic user story was programed to only have a very limited understanding of how to perform SPARQL queries. The basic users were programed not to be able to perform queries of any axiomatic depth. They were limited to direct queries. The three querying cases basic users stores were programmed to execute are for 1) data about a class directly, 2) data about a class **and** other classes, 3) data about a class **or** other classes.

The intermediate user story was programed to be able to have moderate understanding of property path relationships and is modeled to produce queries with a moderate amount of depth. For example queries about some class and *some property* another class.

The advanced user story was programmed to be able to have a fairly deep understanding of possible query paths. For example queries about some class 1 and *some property* some class2, and *another property* some class 3. Although the advanced user story can handle queries of sufficient, they were programmed to only make use of regular pattern. They were not programmed to handle all possible data annotation property path irregularities present in the model datastore.

Two user story experiments were modeled. The first for mobilizing data annotated with an exclusive *and* intersections of two ontology terms was conducted as follows. Eight combinations of terms were used to query for data. For each combination, the py.13 script was used to query for subclasses of the first term in the intersection combination. Modified versions of the py.10 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 second user story experiments involved querying for 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 py.11 script was used to query for terms corresponding to associated parts of the

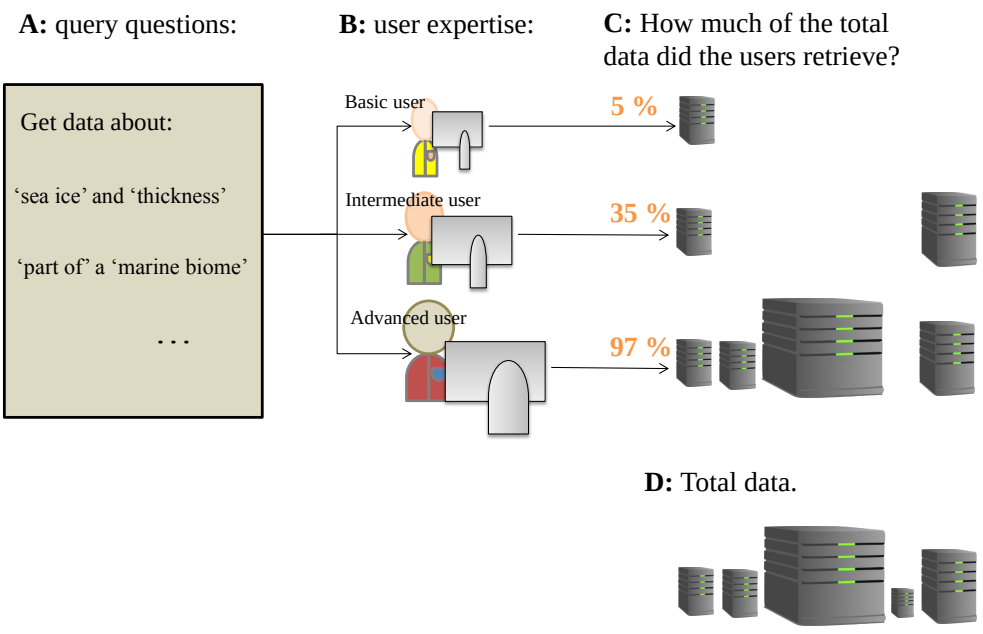

Figure 5 Shows the process by which I simulated user stories of people with basic, intermediate and advanced querying expertise, in order to evaluate how much of the total data they would retrieve from the example datastore. Item **A** shows examples of querying scenario questions which the simulated users stories were asked to get data about, for example data about 'sea ice thickness' or data about 'part of' a 'marine biome'. Item **B** shows the three categories of simulated user stories which had basic, intermediate or advanced expertise for performing SPARQL queries. Item **C** shows the percentages of data which each simulated user story was able to retrieve for a given querying scenario. Item **D** shows all of the data in the datastore which the querying scenarios were attempting to retrieve.

input term. Modified versions of the py.10 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 user story experiments, percentages of retrieved data matrix columns, annotations, points and data matrices were calculated compared with the results derived from an unmodified version of the py.10 or py.11 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.

3. Results

3.1 Leveraging interoperable GO and ENVO semantics

Making use of the interoperable Gene and Environment ontology semantics, I mobilized genomic and environmental data comparative analysis of various types of ENVO:*marine biome* [ENVO_00000447] samples.

I began by asking the competency question:

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

The results of querying the example datastore for the relative genomic and transcriptomic abundance of sequences matching GO:*oxidation-reduction process* [GO_0055114] genes, in various ENVO:*marine biome* [ENVO_00000447], are shown in **Table 1**.

Table 1 Shows the results of a query for the relative genomic and transcriptomic abundances of GO:*oxidation-reduction process* [GO_0055114] genes, in various ENVO:*marine biome*s [ENVO_00000447].

term	ENVO:*marine abyssal zone biome* [ENVO_01000027]	ENVO:*marine bathyal zone biome* [ENVO_01000026]	ENVO:*marine neritic benthic zone biome* [ENVO_01000025]
GO:*oxidation-reduction process* [GO_0055114]	18.15	18.39	9.36
GO:*aerobic respiration* [GO_0009060]	0.23	0.26	0.87
GO:*methanogenesis* [GO_0015948]	0.11	0.12	0.06
GO:*ATP synthesis coupled electron transport* [GO_0042773]	0.06	0.06	0.04
GO:*L-lysine catabolic process to acetate* [GO_0019475]	0.06	0.07	0.01
GO:*respiratory electron transport chain* [GO_0022904]	0.03	0.03	0.13
GO:*mitochondrial electron transport, NADH to ubiquinone* [GO_0006120]	0.02	0.02	0.01
GO:*electron transport chain* [GO_0022900]	0.02	0.02	0.05

term	ENVO:*marine abyssal zone biome* [ENVO_01000027]	ENVO:*marine bathyal zone biome* [ENVO_01000026]	ENVO:*marine neritic benthic zone biome* [ENVO_01000025]
GO:*fatty acid beta-oxidation using acyl-CoA dehydrogenase* [GO_0033539]	0.02	0.02	0.01
GO:*anaerobic electron transport chain* [GO_0019645]	0.01	0.01	0.00
GO:*glycogen biosynthetic process* [GO_0005978]	0.00	0.00	0.01
GO:*aerobic electron transport chain* [GO_0019646]	0.00	0.00	0.00
GO:*methanogenesis, from acetate* [GO_0019385]	0.00	0.00	0.00
GO:*anaerobic glutamate catabolic process* [GO_0019670]	0.00	0.00	0.00
GO:*fatty acid beta-oxidation* [GO_0006635]	0.00	0.00	0.00
GO:*photosynthetic electron transport in photosystem II* [GO_0009772]	0.00	0.00	16.08
GO:*heme oxidation* [GO_0006788]	0.00	0.00	0.00
GO:*photosynthetic electron transport chain* [GO_0009767]	0.00	0.00	1.38
GO:*mitochondrial electron transport, ubiquinol to cytochrome c* [GO_0006122]	0.00	0.00	0.00

Analysis of these results show deep biome samples ENVO:*marine abyssal zone biome* [ENVO_01000027] and ENVO:*marine bathyal zone biome* [ENVO_01000026], had double the relative abundances of non-specific annotations to general oxidation-reduction reduction processes ~18%, relative to ENVO:*marine neritic benthic zone biome* [ENVO_01000025] samples. In contrast, ENVO:*marine neritic benthic zone biome* [ENVO_01000025] samples had three fold increases in GO:*aerobic respiration* [GO_0009060] gene abundances relative to the deep samples.

Deep samples had nearly double the GO:*methanogenesis* [GO_0015948] gene abundances than those of neritic samples, while neritic samples had much greater relative GO:*respiratory electron transport chain*

[GO_0022904] abundances than deep samples.

Neritic samples had elevated abundances of photosynthetic related genes, 16% GO:*photosynthetic electron transport in photosystem II* [GO_0009772] genes and 1.4% GO:*photosynthetic electron transport chain* [GO_0009767] genes, contrasting with the 0.00% abundances of such genes in the deep benthic samples.

Comparisons of GO:*vitamin biosynthetic process* [GO_0009110] genes of which are summarized in **Table 2**.

Table 2 Shows the relative abundance of GO:*vitamin biosynthetic process* [GO_0009110] genes in various types of ENVO:*marine benthic biomes* [ENVO_01000024].

term	ENVO:*marine abyssal zone biome* [ENVO_01000027]	ENVO:*marine bathyal zone biome* [ENVO_01000026]	ENVO:*marine neritic benthic zone biome* [ENVO_01000025]
GO:*riboflavin biosynthetic process* [GO_0009231]	0.25	0.25	0.07
GO:*cobalamin biosynthetic process* [GO_0009236]	0.19	0.19	0.03
GO:*pantothenate biosynthetic process* [GO_0015940]	0.13	0.12	0.04
GO:*thiamine biosynthetic process* [GO_0009228]	0.10	0.11	0.04
GO:*pyridoxine biosynthetic process* [GO_0008615]	0.10	0.09	0.02
GO:*vitamin B6 biosynthetic process* [GO_0042819]	0.05	0.05	0.02
GO:*pyridoxal phosphate biosynthetic process* [GO_0042823]	0.05	0.05	0.02
GO:*pyrroloquinoline quinone biosynthetic process* [GO_0018189]	0.00	0.00	0.00
GO:*anaerobic cobalamin biosynthetic process* [GO_0019251]	0.00	0.00	0.00

From this table I noted that in the deep ENVO:*marine abyssal zone biome* [ENVO_01000027] and ENVO:*marine bathyal zone biome* [ENVO_01000026] samples, there is a general trend that relative gene abundance of GO:*vitamin biosynthetic process* [GO_0009110] genes are higher than the relative transcriptomic abundance of ENVO:*marine neritic benthic zone biome* [ENVO_01000025] samples. For example the relative

gene abundance of GO:*riboflavin biosynthetic process* [GO_0009231] genes was approximately 3.5 times greater in deep biome sample genomes than neritic sample transcriptomes.

These results prompted me to ask the question:

> What genomic features may help to explain the differences in riboflavin abundances between deep and shallow marine benthic biomes?

To investigate other genomic features which may help to explain the differences in riboflavin abundance, I investigated GO:*transition metal ion binding* [GO_0046914] and GO:*transition metal ion transport* [GO_0000041] 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 [84][85]. For full results of relative abundances of metal ion binding, and metal ion transport subclasses see **Tables A1** and **A2**. Querying for subclasses of GO:*transition metal ion binding* [GO_0046914] I found that GO:*ferrous iron binding* [GO_0008198] gene abundance is 0.02-0.03% in deep samples vs. 0.00% in neretic. Furthermore deep sample GO:*ferrous iron transport* [GO_0015684] gene abundance is double that of neritic sample, 0.04% vs. 0.02%.

I next investigated the question of:

> "What biological processes differentiate various types of marine benthic biomes?"

Results of this investigation into GO:*biological process* [GO_0008150] differentiating marine benthic samples are as follows. A PcoA analysis was conducted on all the subclasses of GO:*biological process* [GO_0008150] found in various ENVO:*marine benthic biomes* [ENVO_01000024]. The results of which are shown in **Figure 6** which shows the relative gene and transcript abundances of datasets annotated as various types of ENVO:*marine benthic biomes* [ENVO_01000024]. Together PcoA axes 1 and 2 explain 97.9% of the total variance, with axis 1 explaining 96.1% and axis 2 explaining 1.8%. From the figure we see a differentiation between ENVO:*marine abyssal zone biome* [ENVO_01000027] and ENVO:*marine bathyal zone biome* [ENVO_01000026] samples from those of ENVO:*marine neritic benthic zone biome*s [ENVO_01000025]. We see deep samples ordinated toward the positive values of PcoA dimension 1, and shallow samples toward negative PcoA dimension 1 values.

Table 3 Shows the top ten GO term species which were ordinated closest the to average of the ENVO:*marine neritic benthic zone biome* [ENVO_01000025] sites in the PCoA analysis conducted on the subclasses of GO:*biological process* [GO_0008150] from various ENVO:*marine benthic biomes* [ENVO_01000024].

term	distance between GO site and ENVO species average
GO:*respiratory electron transport chain* [GO_0022904]	0.0090311336
GO:*intracellular protein transport* [GO_0006886]	0.0099281303
GO:*response to arsenic-containing substance* [GO_0046685]	0.0223422138
GO:*regulation of cell division* [GO_0051302]	0.0241089607
GO:*vesicle docking involved in exocytosis* [GO_0006904]	0.0246772626
GO:*spermatogenesis* [GO_0007283]	0.0269553554
GO:*RNA metabolic process* [GO_0016070]	0.027765135

term	distance between GO site and ENVO species average
GO:*tRNA wobble position uridine thiolation* [GO_0002143]	0.0278909429
GO:*cellulose catabolic process* [GO_0030245]	0.0285254331
GO:*inosine salvage* [GO_0006190]	0.0304904183

Table 4 Shows the top ten GO term species which were ordinated closest the to average of the deep ENVO:*marine benthic biomes* [ENVO_01000024] sites in the PCoA analysis conducted on the subclasses of GO:*biological process* [GO_0008150] from various ENVO:*marine benthic biomes* [ENVO_01000024].

term	distance between GO site and ENVO species average
GO:*tyrosine biosynthetic process* [GO_0006571]	0.0021397886
GO:*2'-deoxyribonucleotide metabolic process* [GO_0009394]	0.0026369643
GO:*organic phosphonate metabolic process* [GO_0019634]	0.0033619593
GO:*biotin transport* [GO_0015878]	0.0035475566
GO:*cellular aromatic compound metabolic process* [GO_0006725]	0.0038546443
GO:*tetrahydrobiopterin biosynthetic process* [GO_0006729]	0.0044421841
GO:*riboflavin biosynthetic process* [GO_0009231]	0.0045954664
GO:*peptidyl-lysine modification to peptidyl-hypusine* [GO_0008612]	0.0050485454
GO:*tetrapyrrole biosynthetic process* [GO_0033014]	0.005543313
GO:*glycine biosynthetic process* [GO_0006545]	0.005644747

Using the GO biological processes hierarchy, shown in **Figure A2**, to find drill down to a more specific term by which to investigate the differentiation of these environmental omic data I asked the question:

> "What cellular amino acid biosynthetic processes differentiate various types of marine benthic biomes?"

I performed another PcoA analysis this time on the subclasses of GO:*cellular amino acid biosynthetic process* [GO_0008652]. The results of which are shown in **Figure 7**. Here we can more clearly see classes which separate deep and shallow biome samples. Together the first two PcoA axes explain 94.2% or total variance, with axis 1 explaining 81.7 and axis 2 12.5%.

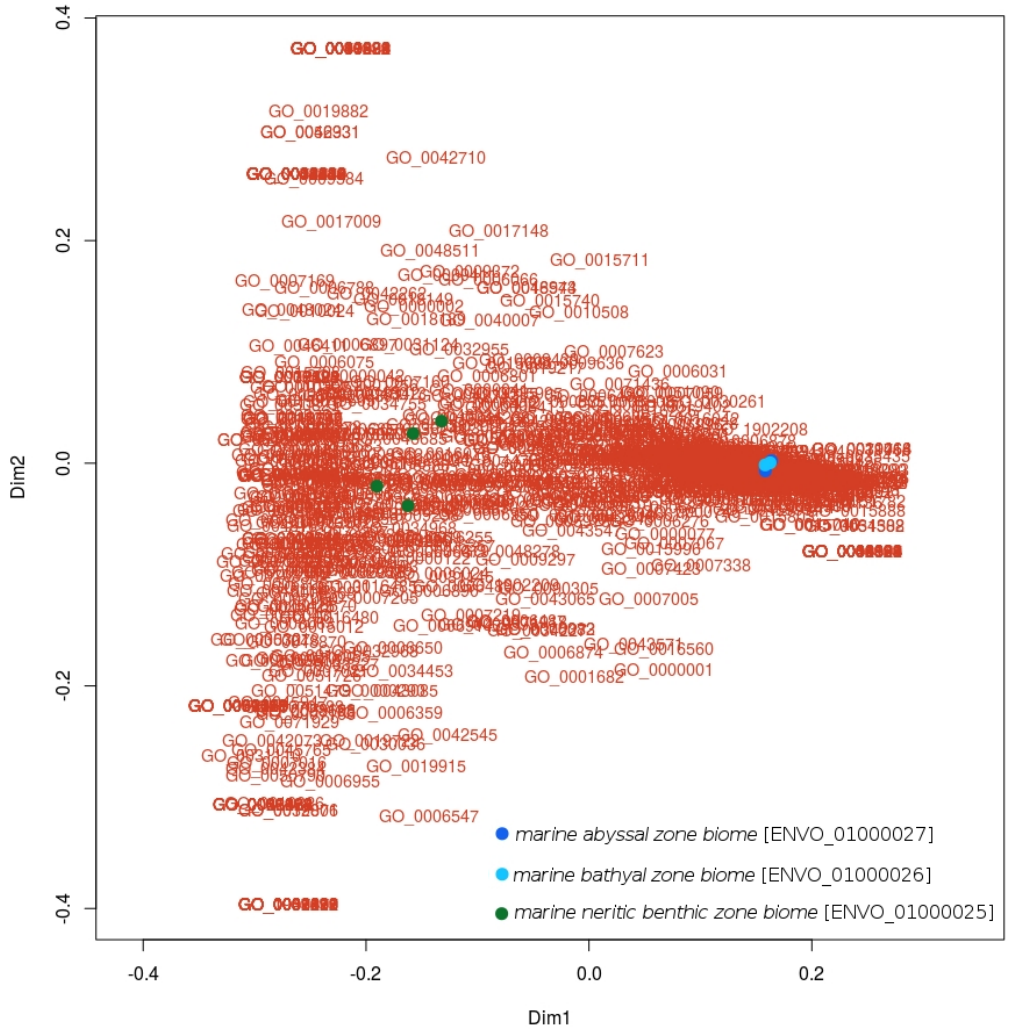

Figure 6 Shows a principal coordinate analyses plot of relative genomic abundance of subclasses of GO:*biological process* [GO_0008150] terms in various marine benthic biomes.

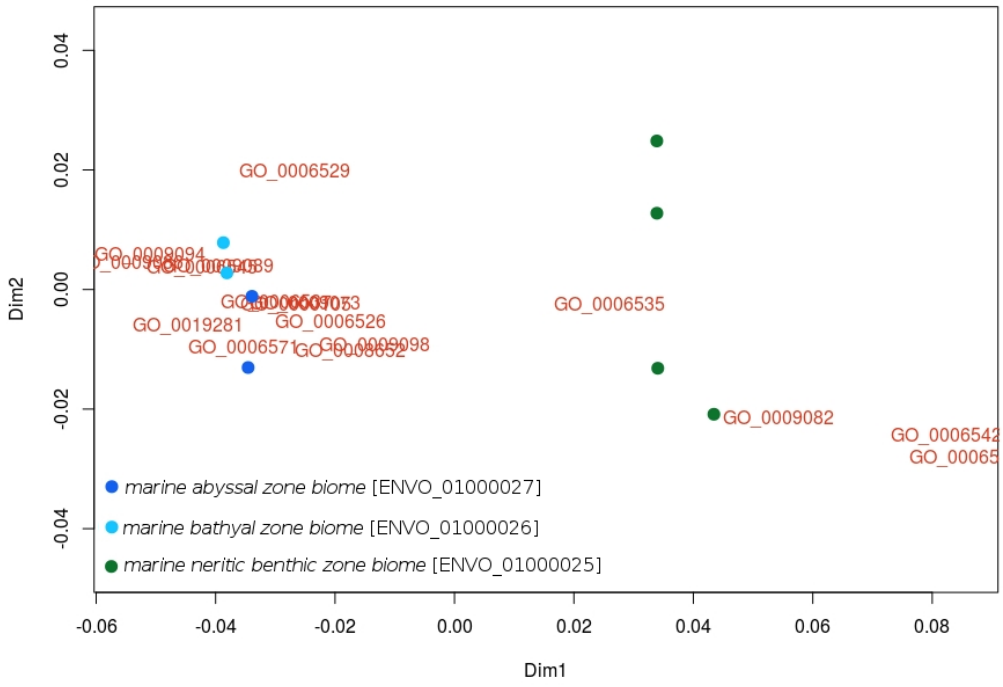

Figure 7 Principal coordinate analyses plot of relative genomic abundance of subclasses of GO:*cellular amino acid biosynthetic process*es [GO_0008652] in various ENVO:*marine benthic biomes* [ENVO_01000024].

Table 5 Shows four GO term species which were ordinated closest the to average of the ENVO:*marine neritic benthic zone biome* [ENVO_01000025] sites in the PCoA analysis conducted on the subclasses of GO:*cellular amino acid biosynthetic process* [GO_0008652] from various ENVO:*marine benthic biomes* [ENVO_01000024].

term	distance between site and species average
GO:*cysteine biosynthetic process from serine* [GO_0006535]	0.0109862681
GO:*branched-chain amino acid biosynthetic process* [GO_0009082]	0.0287599861
GO:*leucine biosynthetic process* [GO_0009098]	0.0508556504
GO:*glutamine biosynthetic process* [GO_0006542]	0.0525977927

Table 6 Shows the top ten GO term species which were ordinated closest the to average of the ENVO:*marine benthic biomes* [ENVO_01000024] sites in the PCoA analysis conducted on the subclasses of GO:*cellular amino acid biosynthetic process* [GO_0008652] from various ENVO:*marine benthic biomes* [ENVO_01000024].

term	distance between site and species average
GO:*lysine biosynthetic process via diaminopimelate* [GO_0009089]	0.005673472
GO:*glutamate biosynthetic process* [GO_0006537]	0.0066474342
GO:*glycine biosynthetic process* [GO_0006545]	0.0074418762
GO:*tyrosine biosynthetic process* [GO_0006571]	0.0088699732
GO:*histidine biosynthetic process* [GO_0000105]	0.0098166072
GO:*L-methionine biosynthetic process from homoserine via O-succinyl-L-homoserine and cystathionine* [GO_0019281]	0.0098643686
GO:*aromatic amino acid family biosynthetic process* [GO_0009073]	0.0114389565
GO:*L-phenylalanine biosynthetic process* [GO_0009094]	0.0160432659
GO:*arginine biosynthetic process* [GO_0006526]	0.0162479534
GO:*methionine biosynthetic process* [GO_0009086]	0.0191591849

Examining the GO hierarchy for subclasses of GO:*cellular amino acid biosynthetic process* [GO_0008652], shown in **Figure A2**, we can drill down into even more specific terms to further differentiate the ENVO:*marine benthic biomes* [ENVO_01000024] annotated samples. I noted from this analysis that the term GO:*cysteine biosynthetic process from serine* [GO_0006535], is ordinated in close proximity with the ENVO:*marine neritic benthic zone biome* [ENVO_01000025] samples.

I further investigated the differences between subclasses of GO:*serine family amino acid biosynthetic processes* [GO_0009070]. The results of which are shown in **Figure 8**. In this analysis, PcoA axis 1 explains 100% of the variance. From this PcoA analysis of subclasses of GO:*serine family amino acid biosynthetic processes* [GO_0009070] we see a clear differentiation of ENVO:*marine neritic benthic zone biome* [ENVO_01000025] vs deep ENVO:*marine abyssal zone biome* [ENVO_01000027] and ENVO:*marine bathyal zone biome* [ENVO_01000026] samples. From this plot we learn that the relative gene abundance of GO:*glycine biosynthetic process*es [GO_0006545] is more abundant in the deep samples, while GO:*cysteine biosynthetic process from serine* [GO_0006535] are more abundant in the ENVO:*marine abyssal zone biome* [ENVO_01000027] samples.

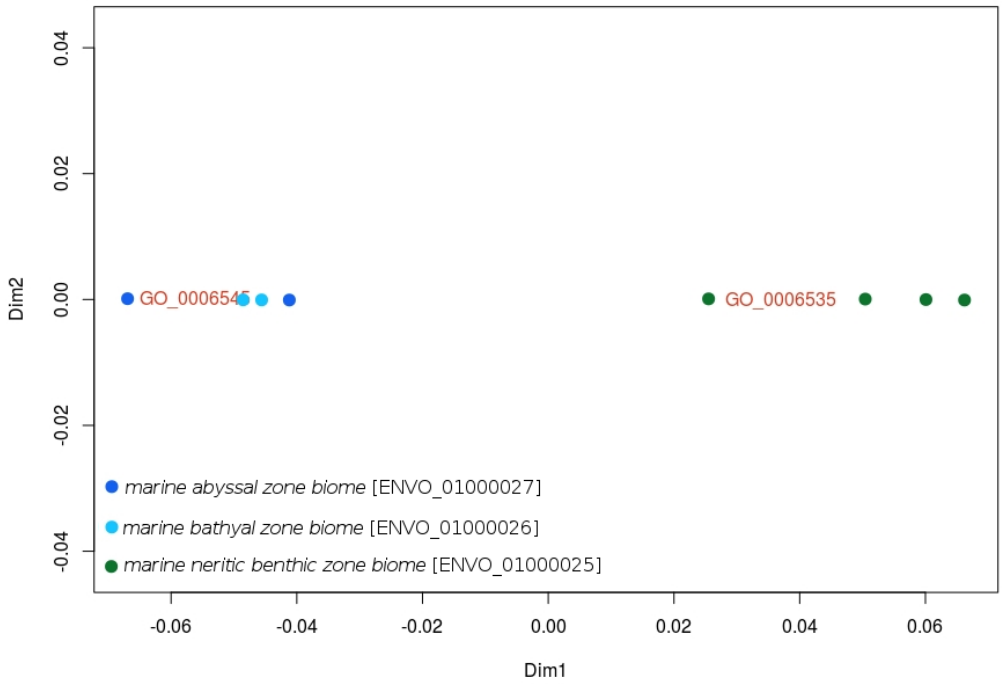

Figure 8 shows a principal coordinate analyses of relative genomic abundance of subclasses of GO:*serine family amino acid biosynthetic processes* [GO_0009070] in various ENVO:*marine benthic biomes* [ENVO_01000024].

3.2 Data pertinent to an environment determined by a phytoplankton community associated with sea-ice

To test the fitness for purpose of using ontologies to collect interdisciplinary data relevant to an ecological question as in (CQ X), I made use of a hypothetical ontology term ENVO:*environment determined by a phytoplankton community associated with sea-ice*. I defined this term 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 include the following subclass axioms: ENVO:*environmental system determined by a community* [URI pending], ENVO:*determined by* [ENVO_2100001] some PCO:*phytoplankton community* [URI pending], RO:*located in* [RO_0001025] some (ENVO:seawater* [ENVO_00002149] and (RO:*part of* [BFO_0000050] some ENVO:*marine water body* [ENVO_00001999])), and finally RO:*adjacent to* [RO_0002220] some ENVO:*sea ice* [ENVO_00002200].

By searching the datastore for data annotated with terms which are present in the axioms of the hypothetical ENVO:*environmental system determined by a community* [URI pending] term, or their subclasses, I was able to collect the following data columns shown in **Table 7**.

Table 7 shows the data columns collected with the assistance of an ontology which may be of relevance to an environment determined by a phytoplankton community associated with sea-ice. The data column label is shown in the first column. The ontology-annotation term which allowed for the retrieval of the collected data column is shown in the second column. Finally the dataset from which the retrieved column originated is shown in the third column.

data column label	annotation term	dataset of origin
phosphate	ENVO:*sea water* [ENVO_00002149]	Inorganic nutrients measured on water bottle samples at AWI HAUSGARTEN during POLARSTERN cruise MSM29.
nitrate	ENVO:*sea water* [ENVO_00002149]	Inorganic nutrients measured on water bottle samples at AWI HAUSGARTEN during POLARSTERN cruise MSM29.
ice or snow temperature	ENVO:*multiyear ice* [ENVO_03000073]	Ice-algal chlorophyll a and physical properties of multi-year and first-year sea ice of core CASIMBO-CORE-2_11.
sea ice thickness	ENVO:*sea ice* [ENVO_00002200]	Influence of snow depth and surface flooding on light transmission through Antarctic pack ice, supplementary data.
signal strength	ENVO:*sea ice* [ENVO_00002200]	Influence of snow depth and surface flooding on light transmission through Antarctic pack ice, supplementary data.
oxygen	ENVO:*sea water* [ENVO_00002149]	Physical oceanography and current meter data from mooring TD-2014-LT.
salinity	ENVO:*sea water* [ENVO_00002149]	Physical oceanography and current meter data from mooring TD-2014-LT.

A demonstration of using such data to perform a mock ecological analysis is included in appendix A.4 Ecological analysis of ontology-collected environmental data. For the references of the original datasets see appendix A2.

3.3 Identifying term author provenance

The following questions address if ontologies are fit for purpose to identify the provenance of ontology term authors:

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

To evaluate such a question, I performed queries to calculate the proportions of ENVO and envoPolar terms which are annotated with an IAO:*term editor* [IAO_0000117] or oboInOwl:*created_by* [created_by] relation which reference an ORCID. ORCID is a unique identifier for scientific and other academic authors and contributors [86]. The results of this analysis are summarized in the **Table 8**.

Table 8 Shows the percentage of ENVO and envoPolar terms annotated with an IAO:*term editor* [IAO_0000117] or oboInOwl:*created_by* [created_by] relation.

ontology	% terms with oboInOwl:*created_by* [created_by]	% terms with IAO:*term editor* [IAO_0000117]
ENVO	14.5	4.2
envoPolar	17.2	31.4

Examining these results I found that 4.2% of ENVO terms have an IAO:*term editor* [IAO_0000117] annotation, whereas 31.4% of envoPolar terms are annotated with an IAO:*term editor* [IAO_0000117]. Terms related by an oboInOwl:*created_by* [created_by] relation account for 14.5% of ENVO terms, wheres they are found in 17.2% of terms from the envoPolar subset. Altogether approximately 20% of ENVO and nearly 50% of envoPolar terms are annotated with a IAO:*term editor* [IAO_0000117] or oboInOwl:*created_by* [created_by] relation.

I also examined how the ORCID application programming interface can be used to retrieve the current contact information from an input ORCID client identifier. Using the publicly available demonstration ORCID client identifier `APP-NPXKK6HFN6TJ4YYI`, I was able to retrieve the associated contact information and obtain the email address `s.garcia@orcid.org`.

3.4 Retrieving primary literature via nodes in a knowledge graph

Assessing if ontologies could serve to connect users to primary literature associated with datasets annotated with ontology term, I asked the question:

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

The results of this question were as follows. In the example datastore there are two datasets which are annotated with a terms which satisfy the condition of being part of an ENVO:*marine biome* [ENVO_00000447]: *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.* [87][88], and *Influence of snow depth and surface flooding on light transmission through Antarctic pack ice, supplementary data.* [89][90]. Both of which are about an ENVO:*marine water body* [ENVO_00001999]. 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 **Table A3**. Selected example results of publications their digital object identifiers as well as the dataset from which the publications were retrieved are shown in **Table 9**. Retrieved publications about variety of ENVO:*marine water body* [ENVO_00001999] related topics, including using chlorophyll pigments to determine phytoplankton taxonomy, plankton ecology, vertical distributions of phytoplankton communities and light transmission through pack-ice.

Table 9 Selected examples of digital object identifiers of publications obtained querying for references of datasets which are about BFO:*part of* [BFO_0000050] a ENVO:*marine biome* [ENVO_00000447].

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 [91]
	10.3402/polar.v34.23349	Summertime plankton ecology in Fram Strait-a compilation of long- and short-term observations [92]
	10.1029/2005JC003207	Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll [93]

data set	reference doi	reference title
influence snow depth	10.1002/2016JC012325	Influence of snow depth and surface flooding on light transmission through Antarctic pack ice [90]

3.5 Proposed ENVO ablation terms

The Feb 2, 2018 VoCamp Glacier Ontology Hackathon [75], was a community consultation session in which polar-related domain experts and ontology developers met to work toward the alignment of base semantics around a set of glacier-related terms. A list of participants is included in appendix AX. During the community consultation session, the semantic representation of a variety of glacier-related topics were discussed. I chose to follow up on one of the more defined outputs of the Hackathon session in which there was discussion about how best to capture the semantics of an ablation process, given that different sources of expert knowledge gave partially different definitions of of the process. The NOAA National Weather Service (2009) glossary [94] defined ablation as only including melting and evaporation processes, whereas the Cogley et al. (2011) IACS-UNESCO Glacier Mass Balance glossary [95] definition referred to all processes which reduce the mass of a glacier, including calving processes. Addressing CQ (X), with these definitions of ablation serving as a case study, I proposed for the following classes to be added to ENVO.

In order to handle cases in which an ablation processes results only from melting and evaporation, I proposed the creation of the term ENVO:*icemelt-derived ice ablation process* [URI pending] which I define as:

> An ice ablation process during which ice is lost due to an icemelt process.

This term would include the subclasses axioms: ENVO:*ice ablation process* [ENVO_01000919], BFO:*has part* [BFO_0000051] some ENVO:*icemelt* [ENVO_01000721], and oboInOwl:*database_cross_reference* [hasDbXref] http://w1.weather.gov/glossary/.

In order to handle cases in which an ablation processes results from both melting and calving processes, I proposed the creation of the term ENVO:*icemelt or calving-derived ice ablation process* [URI pending] which I define as:

> An ice ablation process during which ice is lost due to an icemelt process or ice calving process.

The subclass axioms included in this term would be: ENVO:*ice ablation process* [ENVO_01000919], BFO:*has part* [BFO_0000051] some (ENVO:*icemelt* [ENVO_01000721] or ENVO:*ice calving process* [ENVO_01000917]), and the oboInOwl:*database_cross_reference* [hasDbXref] http://unesdoc.unesco.org/images/0019/001925/1

3.6 Resilience of ontology-enabled data-mobilization to changes in underlying semantic models

As illustrated in **Figure 3** I developed a workflow which makes use of knowledge contained within the OBO ontologies, to discover new knowledge based on connections to stated input knowledge, and retrieve data about these discovered classes. Using this workflow I was able to find classes which are involved in any type of RO:*has participant* [RO_0000057] relation with any type of ENVO:*sea ice formation process* [ENVO_03000044]. Additionally I was able to retrieve data about the discovered classes. The results of this workflow are shown in **Table 10**.

Table 10 Shows the results of submitting a query for classes involved in any RO:*has participant* [RO_0000057] relation with subclasses of ENVO:*sea ice formation process* [ENVO_03000044], as well as the number of data items about the discovered classes retrieved from the datastore. The first column shows the ENVO:*sea ice formation process* [ENVO_03000044] subclasses which are the subjects in a RO:*has participant* [RO_0000057] relation with another class. The second column shows the subproperties of the RO:*has participant* [RO_0000057] relation, which related an ENVO:*sea ice formation process* [ENVO_03000044] subclass to another class. The third column shows classes discovered to be related to a subclass of an ENVO:*sea ice formation process* [ENVO_03000044] by a subproperty of RO:*has participant* [RO_0000057]. The fourth column shows the number of data items about the discovered class which were retrieved from datastore.

subject class	property	discovered class	number of data items
ENVO:*sea ice formation process* [ENVO_03000044]	ENVO:*has input* [RO_0002233]	ENVO:*sea water* [ENVO_01000321]	13
ENVO:*sea ice formation process* [ENVO_03000044]	ENVO:*has output* [RO_0002234]	ENVO:*sea ice* [ENVO_03000066]	6
ENVO:*nilas formation process* [ENVO_03000058]	ENVO:*has input* [RO_0002233]	ENVO:*new ice* [ENVO_03000063]	0
ENVO:*nilas formation process* [ENVO_03000058]	ENVO:*has output* [RO_0002234]	ENVO:*nilas* [ENVO_03000068]	0
ENVO:*young ice formation process* [ENVO_03000059]	ENVO:*has output* [RO_0002234]	ENVO:*young ice* [ENVO_03000069]	0
ENVO:*first year ice formation process* [ENVO_03000060]	ENVO:*has output* [RO_0002234]	ENVO:*first year ice* [ENVO_03000071]	8
ENVO:*second year ice formation* [ENVO_03000061]	ENVO:*has output* [RO_0002234]	ENVO:*second year ice* [ENVO_03000072]	0
ENVO:*multiyear ice formation process* [ENVO_03000062]	ENVO:*has output* [RO_0002234]	ENVO:*multiyear ice* [ENVO_03000073]	8

This workflow makes use of knowledge contained within the OBO ontologies to search for subclasses of an input class e.g. ENVO:*sea ice formation process* [ENVO_03000044] and subproperties of an input property e.g. RO:*has participant* [RO_0000057]. Referring to CQ (12), I wanted to evaluate how much of this data would

be retrievable if changes were made to the semantic models which the ontologies used this workflow themselves use. I tested the susceptibility of this workflow for retrieving data, to changes in the Relations Ontology. I did so by simulating what would happen if substantial changes were to be made to the Relations Ontology without such changes being applied to my example data. I also tested how well this workflow would perform if ontology knowledge graphs didn't employ hierarchically structured subclass relations to represent knowledge about classes and their subclasses. The results of these susceptibility simulations are as follows. If ontologies were to not make use of hierarchically structured subclass relations, only 54.3% of the data items would be retrievable. Faced with major changes to the Relations Ontology 0% of the relevant data would be retrievable.

3.7 Analysis of polar knowledge graph

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. I 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 **Figures A3**, and **A4**. 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. Additional network parameters are summarized in **Table 11**.

Table 11 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 9**, shows a positive skew with the median shortest path length being 2.0, a value less than that of the mean.

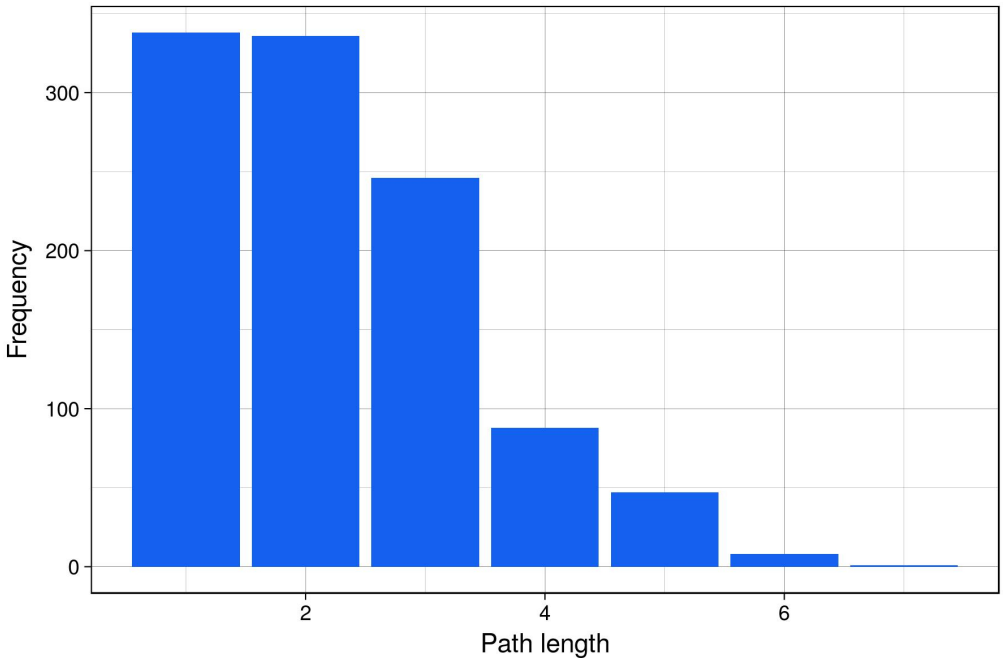

Figure 9 Distribution of shortest path lengths of the envoPolar subset analyzed as a network.

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 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 10**, I 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 I analyzed the betweenness centrality of nodes in the envoPolar network, see **Figure A5**. 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 a function of number of neighboring nodes I 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.

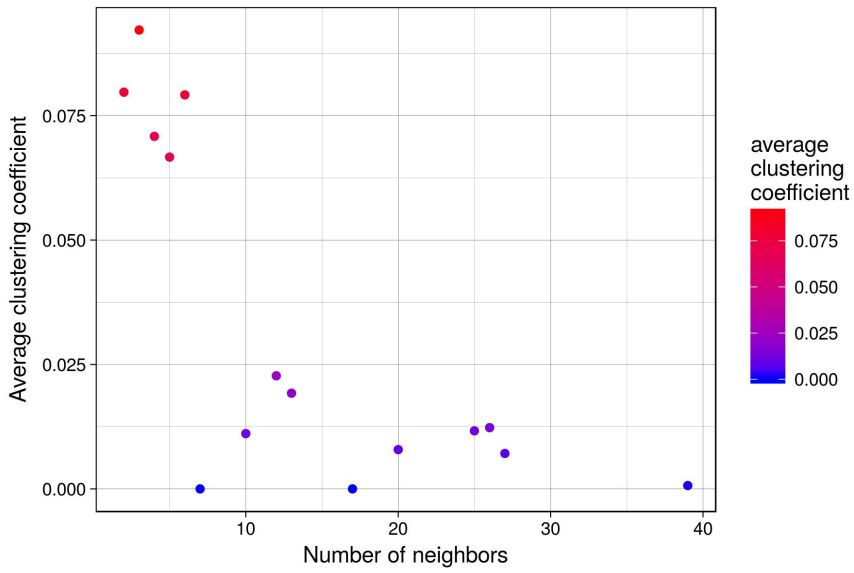

Figure 10 Average clustering coefficient as a function of number of neighboring nodes in the envoPolar subset analyzed as a network.

Table 12 Top ten largest in-degree value terms for polar related terminology from the envoPolar subset.

term	in degree
ENVO:*water ice* [ENVO_01000277]	24
ENVO:*glacier* [ENVO_00000133]	19
ENVO:*snow* [ENVO_01000406]	10
ENVO:*ice mass* [ENVO_01000293]	10
ENVO:*powdery snow* [ENVO_03000027]	6
ENVO:*glacial erosion process* [ENVO_03000006]	4
ENVO:*permafrost* [ENVO_00000134]	4
ENVO:*frazil ice* [ENVO_03000046]	4
ENVO:*glacial ice* [ENVO_03000004]	3
ENVO:*permafrost thawing process* [ENVO_03000086]	3

From the results presented in **Table 12** and **13** showing the nodes with the highest and lowest in-degree values, I observed that there are very few nodes of substantial in-degree values. The distribution of in-degree drops off very rapidly with only a few nodes such as ENVO:*water ice* [ENVO_01000277], ENVO:*glacier* [ENVO_00000133], and ENVO:*snow* [ENVO_01000406] having high in-connectivity values. Even within the top ten in-degree nodes the value decrease substantially from ENVO:*water ice* with 24 in connections, to ENVO:*permafrost thawing process* [ENVO_03000086] with only 3 interconnections. From **Table 13** we see that many of the nodes have in-degrees of 0.

Table 13 Bottom ten smallest in-degree value terms for polar related terminology from the envoPolar subset.

term	in degree
ENVO:*ice field* [ENVO_00000299]	0
ENVO:*shuga formation process* [ENVO_03000079]	0
ENVO:*erosion through nivation* [ENVO_03000121]	0
ENVO:*pingo* [ENVO_00000413]	0
ENVO:*ice cap dome* [ENVO_00000342]	0
ENVO:*ice tongue* [ENVO_00000392]	0
ENVO:*sea ice floe* [ENVO_03000066]	0
ENVO:*snowpack* [ENVO_03000116]	0
ENVO:*ice cap ridge* [ENVO_00000528]	0
ENVO:*perennial snow patch* [ENVO_03000115]	0

3.8 Feasible semantic data annotations

Assessing the practicality of retrieving ontology term annotated data from the example data store, I evaluated the relative amount of data which would be retrieved by estimated user stories programmed to estimate users of various levels of querying expertise, estimating the retrieval rates of data for users of the system with basic, intermediate and advanced querying expertise.

Two estimated user story querying proficiency experiments 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 experiment queried for data which is about a part of an input term, for example, *part of a glacier*. The results of the first experiment, displaying the percentages of annotation terms and data matrix columns retrieved about the intersection of two terms, are presented in **Figure 11**.

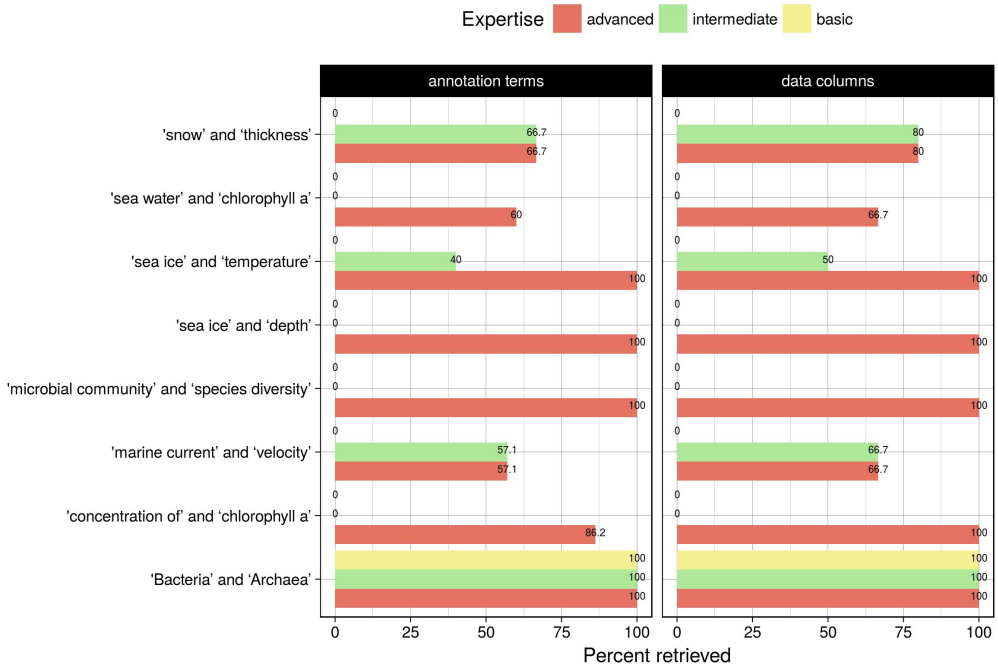

Figure 11 Analysis of querying expertise required to obtain data matrix columns and annotations when querying for data about subclasses of a term AND another term.

User stories estimating basic querying expertise were only able to retrieve data and annotations from the *Bacteria and Archaea* case. User stories estimating intermediate querying expertise were only able to retrieve data from 4 out of the 8 cases tested. Excluding the *Bacteria and Archaea* cases which were covered with 100% success by all three expertise classes, the percentage of annotations retrieved by intermediate user stories ranged from 40-66.7%, whereas the percentage of data columns retrieved ranged from 50-80%. Advanced user stories 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 experiment, displaying the percentages of data matrices and data points retrieved from

BFO:*part of* [BFO_0000050] input terms are presented in **Figure 12**. Basic user stories 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 story expertise case retrieved 25%, the intermediate case retrieved 75% and advanced case retrieved 100%. Intermediate expertise user stories were only able to retrieve data points from 4 out of the 8 *parts of cases*. Although advanced user stories 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*.

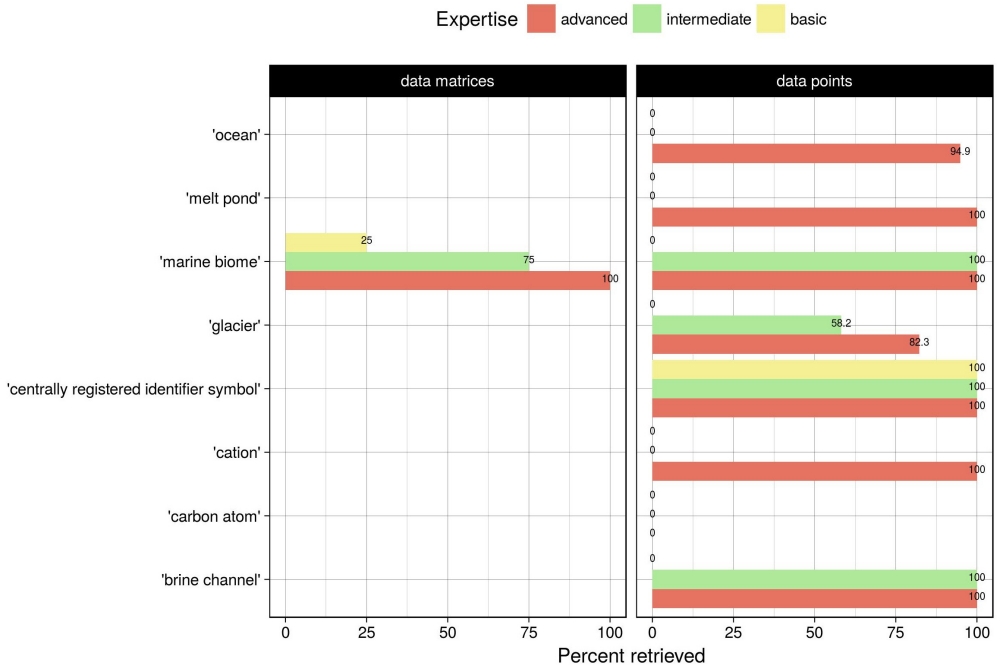

Figure 12 Analysis of querying expertise required to obtain data matrices and data points when querying for data about parts associated with an ontology term.

3.9 Identifying phenomenal interconnections

Attempting to build upon the semantic framework proposed as necessary by Stec et al. (2017), for future modeling of plankton ecosystems [40]. Asking the question:

> "Does the inclusion of novel expert knowledge about phenomena relating to plankton ecology into the ENVO knowledge graph aid to better understand the interconnections of such phenomena?"

To answer this question I proposed the following plankton ecology related ontology term to the ENVO [19][26] PATO [78] ECOCORE [79] and PCO [80] ontologies. A complete description of these potential ontology terms, are available from the following URL: https://github.com/kaiiam/kblumberg_masters_thesis/wiki/plankton-ecology

An example of a term prepared for the Population and Community Ontology is PCO:*phytoplankton bloom process* which I have 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.

I proposed for this PCO:*phytoplankton bloom process* term to have a variety of subclass axioms, which include other proposed PCO and ECOCORE terms. The subclass axioms I've proposed include: PCO:*plankton bloom process* [URI pending], RO:*has participant* [RO_0000057] some PCO:*phytoplankton community* [URI pending], BFO:*part of* [BFO_0000050] some ECOCORE:*surface photoautotrophic biomass formation* [URI pending], BFO:*has part* [BFO_0000051] some PCO:*population bloom* [URI pending], BFO:*occurs in* [BFO_0000066] some ENVO:*water body* [ENVO_00000063], and finally an oboInOwl:*database_cross_reference* [hasDbXref] https://en.wikipedia.org/wiki/Phytoplankton.

An example of a term prepared for the Phenotypic Quality Ontology is PATO:*planktonic*, which I have 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.

This definition encodes the classic oceanography definition of plankton into the ontology knowledge graph, characterizing planktonic organisms as drifting organisms unable to swim against a current [96]. I proposed this planktonic class to be a subclass of PATO:*organismal quality* [PATO_0001995], and for it to include the oboInOwl:*database_cross_reference* [hasDbXref] https://en.wikipedia.org/wiki/Phytoplankton.

Another term I propose to be added to PATO is PATO:*ice cover of a planetary surface*, which I define 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.

This class has the following proposed subclasses: PATO:*physical quality* [PATO_0001018], RO:*inheres in* [RO_0000052] some ENVO:*planetary surface* [ENVO_01000324] or ENVO:*water body* [ENVO_00000063], BFO:*has part* [BFO_0000051] some ENVO:*surface layer* [ENVO_00010504] and (RO:*adjacent to* [RO_0002220] some ENVO:*atmosphere [ENVO_01000267] or ENVO:*outer space [ENVO_01000637]), RO:*adjacent to* [RO_0002220] some ENVO:*water ice* [ENVO_01000846], finally having oboInOwl:*exact synonym* [hasExactSynonym] *ice cover* and *ice coverage*.

I also prepared many terms for ENVO. For example the term ENVO:*marginal ice zone*, which I defined as:

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

This term includes the subclass axioms: ENVO:*environmental zone* [ENVO_01000408], RO:*has quality* [RO_0000086] some PATO:*ice cover of a planetary surface* [URI pending], RO:*overlaps* [RO_0002131] some (ENVO:*sea ice* [ENVO_03000066] and ENVO:*marine water body* [ENVO_00001999]), RO:*causally upstream of, positive effect* [RO_0002304] some PCO:*phytoplankton bloom process* [URI pending], RO:*causally upstream of, positive effect* [RO_0002304] some ECOCORE:*photoautotrophic biomass formation* [URI pending], oboInOwl:*has related synonym* [hasRelatedSynonym] *sea ice edge*, with oboInOwl:*database_cross_reference*s [hasDbXref] http://www.npolar.no/en/facts/the-marginal-ice-zone.html and https://doi.org/10.1016/j.jmarsys.2013.11.008 [76].

For ENVO I also propose the term ENVO:*marine environment determined by a phytoplankton community bloom*, which I define 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.

This term would include the subclass axioms: ENVO:*marine environment determined by a community bloom* [URI pending], ENVO:*determined by* [ENVO_2100001] some PCO:*phytoplankton bloom process* [URI pending], RO:*causally downstream of, negative effect* [RO_0002305] some ENVO:*marine water body stratification* [URI pending]

I also prosed to encode the ENVO:*marine water body stratification* term in ENVO with the definition:

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

This term includes the subclass axioms: ENVO:*water body stratification* [URI pending], BFO:*occurs in* [BFO_0000066] some ENVO:*marine water body* [ENVO_00001999], RO:*results in formation of* [RO_0002297] some ENVO:*stratified marine water body* [URI pending], and finally RO:*results in formation of* [RO_0002297] min 2 ENVO:*marine layer* [ENVO_01000295].

I have also prosed to include the term ENVO:*sea ice melting* in ENVO, defining it 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.

Describing it with the subclass axioms: ENVO:*ice melt* [ENVO_01000721], RO:*has input* [RO_0002233] some (ENVO:*sea ice* [ENVO_00002200] and (RO:*adjacent to* [RO_0002220] some ENVO:*troposphere* [ENVO_01000540]) and (RO:*adjacent to* [RO_0002220] some ENVO:*marine water body* [ENVO_00001999])) and PATO:*decreased degree of illumination* [PATO_0015015], RO:*has output* [RO_0002234] some (ENVO:*meltwater* [ENVO_01000722] and (RO:*adjacent to* [RO_0002220] some ENVO:*troposphere [ENVO_01000540]) and (RO:*adjacent to* [RO_0002220] some some ENVO:*marine water body* [ENVO_00001999])) and PATO:*increased degree of illumination* [PATO_0015014], RO:*causally upstream of* [RO_0002411] some ENVO:*marine water body stratification* [URI pending], finally with oboInOwl:*database_cross_references* [hasDbXref] https://doi.org/10.1016/j.jmarsys.2013.11.008 [76], https://en.wikipedia.org/wiki/Attenuation and https://en.wikipedia.org/wiki/Optical_properties_of_water_and_ice.

I bring these terms together in the following **Figure 13** as to visualize the interconnecting relationships between them.

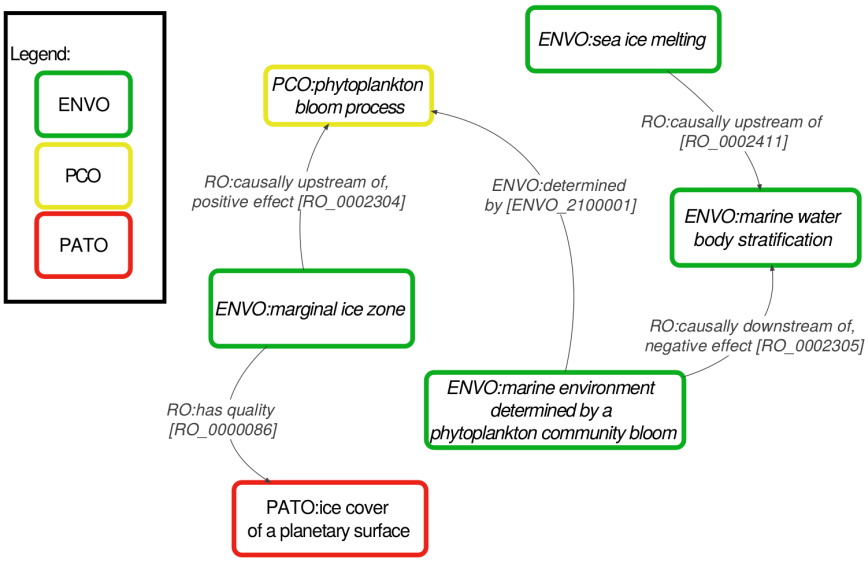

Figure 13 Diagram indicating relationships interconnecting highlighted subset of ontology term contributions made during the course of this work to encode expert knowledge about plankton ecology. Nodes are proposed ontology classes for the ENVO, PCO and PATO ontologies. Edges are RO relationships.

I assessed if ontologies can help us to contextualize the interconnections between encoded phenomena about which we have expert knowledge. **Figure 13** illustrating the relationships interconnecting the various phenomena

which effect phytoplankton blooms shows how ontologies help us to understand the interconnections between pieces of expert knowledge. Starting by examining physical processes which have effects on phytoplankton blooms, I assembled and encoded the following expert knowledge. As the onset of phytoplankton blooms have been shown to be dependent on the timing of the retreat of melting sea ice [77], I encoded the knowledge of such a relationship into the potential term sea ice melting, by specifying that it is RO:*causally upstream of* [RO_0002411] some ENVO:*marine water body stratification* [URI pending].

I encoded the concept of the ENVO:*marginal ice zone*, which is described as the transition zone between the open ocean and sea ice [78]. In the ENVO:*marginal ice zone*, melting sea-ice has been shown to promote phytoplankton growth by stratifying the water column [76]. To encode the knowledge of the process of stratification I created the potential class ENVO:*marine water body stratification*, which RO:*results in formation of* [RO_0002297] at least two ENVO:*marine layer*s [ENVO_01000295]. In the ENVO:*marginal ice zone*, stratification of the water body resulting from melting sea ice has been shown to be the location of maximum chlorophyll [76], controlling the onset of the seasonal phytoplankton blooms [77]. To encode knowledge about phytoplankton blooms, I 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. I have encoded into the ENVO:*marginal ice zone* term an axiom stating that marginal ice zones are RO:*causally upstream of, positive effect* [RO_0002304] some PCO:*phytoplankton bloom process* [URI pending]. As phytoplankton bloom processes have a profound impact on their surrounding environment, I have also created the term ENVO:*marine environment determined by a phytoplankton community bloom*, which I have specified to be related to PCO:*phytoplankton bloom processes* with the axiom ENVO:*determined by* [ENVO_2100001] some PCO: *phytoplankton bloom process* [URI pending]. I have also encoded the connection between an ENVO:*marine environment determined by a phytoplankton community bloom* and ENVO:*marine water body stratification* with the RO:*causally downstream of, negative effect* [RO_0002305] relationship.

4. Discussion

4.1 Comparative environmental genomics

Making use of data annotated with interoperable Gene and Environment Ontology terms, I mobilized data to answer the question:

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

I examined the results of **Table 1**, showing the relative genomic and transcriptomic abundances of GO:*oxidation-reduction process* [GO_0055114] genes, in various ENVO:*marine biome*s [ENVO_00000447], to address 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 ENVO:*marine neritic benthic zone biome* [ENVO_01000025] samples are relatively enriched in GO:*photosynthetic electron transport in photosystem II* [GO_0009772], GO:*aerobic respiration* [GO_0009060] and GO:*respiratory electron transport chain* [GO_0022904] related genes. Relative to the ENVO:*marine abyssal zone biome* [ENVO_01000027] and ENVO:*marine bathyal zone biome* [ENVO_01000026] samples enriched in undifferentiated GO:*oxidation-reduction process*es [GO_0055114] and GO:*methanogenesis* [GO_0015948] related genes. This preliminary comparison 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, I asked another question:

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

Table 2, showing the relative abundance of GO:*vitamin biosynthetic process* [GO_0009110] genes in various types of ENVO:*marine benthic biomes* [ENVO_01000024], further addresses the use of ontologies to interconnect genomic data. Analyzing the result of finding elevated GO:*vitamin biosynthetic process* [GO_0009110] genomic capacity in deep samples relative to the lower transcriptomic capacity in ENVO:*marine neritic benthic zone biome* [ENVO_01000025] samples. I made use of the ontology knowledge graph to search for the relative abundances of other GO:*biological process* [GO_0008150] and GO:*molecular function* [GO_0003674] which may help to explain the differences in GO:*vitamin biosynthetic process*es [GO_0009110]. As flavin compounds 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 [84][85], I investigated GO:*transition metal ion binding* [GO_0046914] and GO:*transition metal ion transport* [GO_0000041] subclasses, with the aid of the Gene Ontology knowledge hierarchy. The results of elevated GO:*ferrous iron binding* [GO_0008198] and GO:*ferrous iron transport* [GO_0015684] gene abundance in ENVO:*marine abyssal zone biome*s [ENVO_01000027] and ENVO:*marine bathyal zone biome*s [ENVO_01000026] relative to ENVO:*marine neritic benthic zone biome*s [ENVO_01000025] samples, combined with the elevated GO:*riboflavin biosynthetic process* [GO_0009231]

suggest a potential ecophysiological connection. Allowing us to hypothesize that riboflavin mediated iron reduction differentiates ENVO:*marine abyssal zone biome* [ENVO_01000027] and ENVO:*marine bathyal zone biome* [ENVO_01000026] sediments from ENVO:*marine neritic benthic zone biome* [ENVO_01000025] 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 [34]. Multiple competing hypotheses 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. The continued development of cyberinfrastructure by which to conduct these types of comparative genomic analysis could be scaled up to a large machine-actionable system the analysis of microbial genomics. Thematically this could build upon previous efforts such as the Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA), a semantically-annotated environmental genomic data base supporting semantic queries [97].

To further investigate potential knowledge which can be derived from the interconnection of data annotated with GO and ENVO terms, I asked the following question of the example datastore:

> "What biological processes differentiate various types of marine benthic biomes?"

By drilling down into finer levels of granularity within the GO:*biological process* [GO_0008150] hierarchy, see **Figures 4** and **A2**, I was able to pinpoint processes differentiating deep ENVO:*marine abyssal zone biome* [ENVO_01000027] and ENVO:*marine bathyal zone biome* [ENVO_01000026] samples from ENVO:*marine neritic benthic zone biome* [ENVO_01000025] samples. **Figure 5** An examination of GO:*cellular amino acid biosynthetic processes* [GO_0008652], a subclass of GO:*biological process* [GO_0008150], shows a much clearer differentiation of samples, than does the higher level class in **Figures 4**. From these results I was able to pinpoint even more specific subclasses to investigate differences between types of ENVO:*marine benthic biomes* [ENVO_01000024].

Figures 5 showing GO:*serine family amino acid biosynthetic processes* [GO_0009070] illustrates a very clear and potentially biologically interesting difference. GO:*glycine biosynthetic process* [GO_0006545], which has as subclass GO:*glycine biosynthetic process from serine* [GO_0019264], is more abundant in the deep ENVO:*marine benthic biomes* [ENVO_01000024] samples, whereas GO:*cysteine biosynthetic process from serine* [GO_0006535] is more abundant in the ENVO:*marine neritic benthic zone biome* [ENVO_01000025] 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 ENVO:*marine neritic benthic zone biome* [ENVO_01000025] tend to pro-

duce cysteine from serine, whereas organisms from deep ENVO:*marine abyssal zone biome* [ENVO_01000027] and ENVO:*marine bathyal zone biome* [ENVO_01000026] 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 [98]. This may help cells to cope with high pressure deep environments.

What is most notable from this finding is that I was able to discover this potential difference based solely on information contained within the gene ontology. Having no preliminary ideas about what GO:*cellular amino acid biosynthetic processes* [GO_0008652] which could differentiate deep ENVO:*marine abyssal zone biome* [ENVO_01000027] and ENVO:*marine bathyal zone biome* [ENVO_01000026] from ENVO:*marine neritic benthic zone biome* [ENVO_01000025] samples, nor knowledge of what amino acids serine is a precursor to. This example shows how ontologies are fit for purpose to interconnect disparate omic datasets and generate working hypotheses therefrom.

4.3 Ontology guided data and knowledge discovery: are ontologies fit for purpose?

//

(CQ Y > "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?")

Being able to navigate machine-accessible information to discover data and knowledge are key competencies for ontologies to be able to perform CQ (7,8). To explore the question of ontological fitness for purpose for the discovery of knowledge and data, I have analyzed how ontology annotation terms can be used to facilitate the assembly of relevant knowledge and data about a phenomenon of interest. I also evaluated the network parameters of an example ontology knowledge graph to asses if it is suitable to connecting knowledge explicitly stated by a researcher to new, unstated but related knowledge.

Evaluating CQ (X), I analyzed how ontology terms can be used to facilitate the assembly of relevant knowledge and data about a phenomenon of interest, for example environmental factors which may influence a sea-ice associated phytoplankton community. Doing so, I encountered both expected and unexpected data which may potentially be of use to perform an ecological analysis on the data. For example when I retrieved data about environmental factors which may influence sea-ice associated phytoplankton communities, I retrieved some variables I expected to find such as sea ice thickness, multi-year sea ice temperature, degree of illumination of sea ice. Additionally, I also retrieved some unexpected but potentially relevant data such as sea water salinity, oxygen, phosphate and nitrate concentrations. In an ideal case the ontology annotations would help us to assemble data which researchers would not have thought to include in the analysis. Sea water salinity data for example, may give an indication of the existence of meltwater released from sea ice melting. Such data coinciding with available nutrient concentrations could indicate the beginning of a phytoplankton bloom.

Despite being able to harness the rich wealth of knowledge encoded within ontologies, a lack of machine-accessible data hinders our ability to perform ecological analysis on ontology-discovered data. ence I would

describe the current situation as rich in knowledge but poor in data. In order for future machine-guided routines to make use of ontologies to discover data relevant to a phenomena of interest and perform ecological analysis thereon, more machine-accessible data is required. This is not necessarily an infeasible goal for the scientific community to strive toward. Simple efforts to toward standardizing data outputs generated by projects such as environmental observatories could go a long way toward providing machine-accessible data to analyze. Striving toward the long term goal of improved data standardization, could help us to enable data to be used to its full potential. This would allow for future machine-guided meta-analyses to be conducted on large machine-accessible data sets. Using ontologies and machine-accessible data as the source material for future artificial intelligence knowledge representation endeavors may allow us to break through traditional analysis barriers. Enabling deeper, expert-knowledge and large-data-informed machine-guided ecological meta-analyses to be performed.

//CQ(8) > "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?"

In addition to the discovery of data, I assessed if ontological knowledge graphs are fit for the purpose of discovering new knowledge from ontology knowledge graphs which is related to the stated input knowledge. For this I made use of the envoPolar ENVO subset as an example ontology knowledge graph. Answering CQ(X), I assessed the network parameters of the envoPolar graph, assuming that connectivity within the graph is analogous to the facility of researchers searching the network to discover knowledge associated with their stated input knowledge.

My 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. I additionally examined the diameter of the network, the longest possible path between connected nodes in a network to gain insight into how well integrated the network is. Longer maximum path lengths equate to less well integrated networks [99]. In the envoPolar network, the maximum path length is only 7. Examining the distribution of path lengths, see **Figure 9** I observe that the majority of nodes have a path length of 2. This means that the average node in the network is only 2 steps away from most other nodes. Hence the overall network is well connected. Examining the in-degree distributions of nodes in the network, see **Figure A4** in the appendix, I remarked 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 10** the graph of average clustering coefficient as a function of number of neighboring nodes, where I observed 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. **Figure A6** 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 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. ENVO:*geographic feature* [ENVO_00000000] 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 [100], which describe an exponential relationship between the degree of connectivity a node has and the frequency of its occurrence. Examining the topology of the envoPolar network I 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 [101]. Allowing for a very short average path length within a very large network, which is what I observed 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 were no longer used and suddenly a very central node such as ENVO:*geographic feature* [ENVO_00000000] were to be removed without replacement, the network would shatter into many unconnected components, rendering it unable to interconnect information. In the current organizational structure the majority of nodes are only two steps away from highly centralized and well-connected *hub* nodes. Through these highly centralized nodes the network is very highly interconnected. This is due to the hierarchical organizational structure of the ontology.

The majority of nodes, however, are not very well connected to. The results of **Table 12** show that only a handful of nodes with high in-degree values are polar related semantic terms sourced from lower down in the hierarchy. Furthermore the low average in connectivity of nodes in the network implies that most nodes are not well connected to by other nodes. Hence it is my assessment that more work is needed to encode the potential relationships which exist between classes. Building upon the relational connectivity of the envoPolar knowledge graph will be necessary in order for nodes within the graph to be of sufficient in-connectivity to facilitate the discovery of new knowledge based on relationships to stated input knowledge.

4.4 Tracking information provenance

Jackson's (2012) bestiary of ignorance proposes four categories in an overview of knowledge or lack of knowledge about a subject [102]. 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 [5]. Hortal et al. (2015) propose that informatics methods could be employed to facilitate community access to non-easily search-able knowledge collections [5]. Considering informatics strategies by which to improve community access to unknown known knowledge, I examined various types of information which ontologies could be used to track the provenance of. I discuss ways in which ontology knowledge graphs can help to identify the provenance of primary literature associated with annotated datasets, specimens from a museum or collection, and authors who contribute expert knowledge to ontologies. Mobilizing known unknown information

data and knowledge into a greater ontology knowledge graph, is a first step toward overcoming the limitation of known unknown knowledge.

Evaluating the fitness for purpose of ontologies to connect users to primary literature about datasets annotated with ontology terms, I posed the following question of the 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. For example, by searching for publications about BFO:*part of* [BFO_0000050] a ENVO:*marine biome* [ENVO_00000447] the ontology knowledge graph lead us to papers about a ENVO:*marine water body* [ENVO_00001999]. 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* [90], 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 an OBCS:*data matrix* [OBCS_0000120] about some PATO:*physical* quality* [PATO_0001018] and (RO:*inheres in* [RO_0000052] some (ENVO:*marine water body* [ENVO_00001999]) and (RO:*adjacent to* [RO_0002220] some ENVO:*sea ice* [ENVO_00002200]))

In other cases publication less directly related to a dataset about a part of an ENVO:*marine biome* [ENVO_00000447], were retrieved. Such as for example the publication *An evaluation of the application of CHEMTAX to Antarctic coastal pigment data* [91], 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 an OBCS:*data matrix* [OBCS_0000120] about some CHEBI:*chlorophyll a [CHEBI_18230] and (RO:*part of* [BFO_0000050] some ENVO:*marine water body* [ENVO_00001999])

Although these are relatively uninteresting 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 some: PCO:*phytoplankton community bloom* [URI pending] and (RO:*composed primarily of* [RO_0002473] some *prokaryotic organisms* and (RO:*overlaps* [RO_0002131] some RO:*output of* [RO_0002353] some ENVO:*iceberg calving process* [ENVO_03000031] and (RO:*located in* [RO_0001025] some GAZ:*Weddell Sea* [GAZ_00004045]))).

A semantic annotation such as this could be realized by using the open source gazetteer GAZ [103], 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 GAZ:*Weddell Sea* [GAZ_00004045]. Terms such as *phyto-*

plankton 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 questions 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 NCBITaxon:*Alveolata* [NCBITaxon_33630] (dinoflagellates), were to encode information about the morphologies of such specimens in queryable formats with ontology term annotations. Providing an specimen annotation such as some: owl:*NamedIndividual* [NamedIndividual] and (rdf:*subClassOf* [subClassOf] some NCBITaxon:*Alveolata* [NCBITaxon_33630] and (*has role* [RO_0000087] some OBI:*specimen role* [OBI_0000112] and (RO:*has quality* [RO_0000086] some PATO:*morphology* [PATO_0000051]))).

Users would be enabled to ask questions of the query-able natural history specimen collections such as:

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

This would go a long 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.

Ontologies can also be used to track the provenance of term authors who have contributed expert knowledge to an ontology knowledge graph. There is a need to track the provenance of expert knowledge authorship, as scientific discoveries are increasingly being enabled through Internet based collaboration [104]. Ontologies are semantic representations of expert knowledge, and thus have the potential to facilitate on-line networking among scientists, allowing users to connect to the authors who have contributed their expert 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 unambiguous personal identifiers. These identifiers would need to be connected to a living system which is query-able. Allowing for users to query the ontology knowledge for any authors who contributed knowledge related to specific input terminology of interest. Enabling 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 terms with a link to an Open Researcher and Contributor ID (ORCID) [86]. ORCIDs are persistent digital identifier serving as primary keys to distinguish researchers. ORCID satisfies the requirement of being a permanently maintained persistent living system by which to track author provenance. Being a web service, ORCID also 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 serve to interconnect people who contribute knowledge to the knowledge they have contributed, I 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 a given term. Although ontologies such as ENVO 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 without a priori knowledge of its existence. The practice of using author ID ranges works for ontologies with smaller numbers of contributing authors, but constituents a cumbersome solution for the micro-crediting of many authors who may only ever contribute to a single term. Directly annotating terms with links to contributing author ORCIDs provides a more easily scalable solution for future influxes of contributing authors. Directly annotating ontology terms with links to the ORCIDs of contributing authors would serve to identify term author provenance. As well as facilitate future networking amongst scientists by connecting ontology term authors to OCRIDs, from which current contact information can be pulled.

4.5 Practical and resilient systems for knowledge-representation and data-mobilization

In principal, ontologies can be used to provide machine-readable representations of expert knowledge as well as the semantic infrastructure by which to interconnect and mobilize machine-accessible data. In this section I discuss the practicality of using ontologies to perform such functions, as well as their resilience to the incorporate of new knowledge, or changes to their underlying semantic models. First in reference to CQ (11) I discuss strategies for ontology development to make them resilient to the incorporation of new and possibly contradictory expert knowledge. Next, referencing CQ (12), I discuss the resilience of ontology-enabled data-mobilization workflows, if subject to changes in their underlying semantic models. Finally, referencing CQ (13), I evaluated how practical it is for predicted users to retrieve ontology-annotated machine-accessible data.

Addresssing CQ (11), an example of the requirement of ontologies to be resilient to the incorporation of expert knowledge which may have conflicting or partially non-overlapping definitions came up during the VoCamp Glacier Ontology Hackathon community glacial-consultation session [75], in which different expert knowledge sources presented different definitions of ablation. As ontologies take an agnostic stance when representing knowledge which has multiple definitions or which pertains to competing hypotheses [34], a variety of approaches can be taken in parallel to incorporate such definitional discrepancies into the ontology knowledge graph. This involves determining the differences between competing definitions, finding or creating other semantics to represent these

differences and finally creating multiple versions of the term of interest which have subclass axioms referencing their differences. In the ablation case the difference concerned whether or not calving processes contribute to ice ablation processes. Thus I suggested the creation of two different ablation classes, both of which are to be subclasses of the general ENVO:*ice ablation process* [ENVO_01000919]. The first class ENVO:*icemelt-derived ice ablation process* [URI pending] would have the axiom: BFO:*has part* [BFO_0000051] some ENVO:*icemelt* [ENVO_01000721], telling the semantic knowledge layer that this class only refers to ablation which is due to an icemelt process. The second class ENVO:*icemelt or calving-derived ice ablation process* [URI pending], would have the axiom: BFO:*has part* [BFO_0000051] some (ENVO:*icemelt* [ENVO_01000721] or ENVO:*ice calving process* [ENVO_01000917]), telling the semantic knowledge layer that this class represents ablation processes which are due to icemelt and or calving processes. If in the future, more polar data were to be available for analysis, these terms could be used to evaluate the overlap of data retrieved when performing queries for these different definitions of ablation. This could possibly help to ask and answer a question such as:

> "To what extent are calving processes contributing to ice ablation process relative to icemelt processes?"

Addressing CQ (12), I evaluated the extent to which ontology-enabled data-mobilization workflows such as those I developed to mobilize data about participants in any type of ENVO:*sea ice formation process* [ENVO_03000044], are susceptible to potential changes made to their underlying semantic models. I simulated the effects of not using hierarchically structured subclass relations or structured RO relations to retrieve data about participants in ENVO:*sea ice formation process*es [ENVO_03000044] from my datastore. The results of these simulations indicate that my ontology guided data-mobilization workflow would be quite susceptible to changes made to the underlying semantic models employed by ontologies.

The results show that if ontologies didn't make use of hierarchically structured subclass relationships to represent the taxonomy of knowledge, for example representing the knowledge that a first year ice formation process is a type of sea ice formation process, a data-mobilization workflow such as the one I devised would retrieve substantially less data. Additionally, if a non-standardized set of relations were to be used in place of the structured Relations Ontology relations, it would not be possible to use a data-mobilization workflow to retrieve data.

These results illustrate that in order for ontologies to be used in data-mobilization workflows, it is important that the foundational semantic models they employ are not radically changed. This underscores the need for a well-established and well-structured set of semantics to be used to annotate and mobilize machine-accessible data. Although I am not specifically advocating for the use of the OBO Foundry and Library ontologies to serve as the semantic infrastructure for the annotation and mobilization of machine-accessible data, OBO ontologies would be a reasonable choice to do so. OBO ontologies provide a pre-existing, interoperable semantic infrastructure, including the arguably most successful and widely used biomedical ontology, the Gene Ontology [25]. 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 [51]. The SWEET ontologies are the *de facto* standards for the formal representation of earth and environmental science domain knowledge [105]. Aligned OBO and SWEET semantics hold great potential to aid in the future interconnection of environmental

and genomic data.

Addressing CQ (13) I evaluated how practical it would be for potential users of various levels of querying expertise to retrieve ontology-annotated machine-accessible data from my example 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, I evaluated this question by testing the performance of predicted user stores estimating users of various querying proficiencies to retrieve example data. The results of these predicted user stories 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.

Clearly these outcomes are limited being only user story simulation conducted in place of proper user group analyses. Semantic science advances could be made by creating an open linked data repository with various styles of data annotations of which a user focus group would be tasked with querying. This would serve to inform us as to what axiomatic annotations are readily querable lending themselves to successful data mobilization, and which are not. Additional efforts could make use of Neuro-Linguistic Programming (NLP) knowledge to create data annotations which are intuitive to average users. Additionally, efforts could be undertaken to find better ways of connecting and storing linked data which make use of more natural linguistic patterns, to attempt to make linked data accessible by methods other than cumbersome and technical SPARQL queries.

Next, I discuss the axiomatic data annotation patterns which the predicted user stories were and were not able to query. I first examine an example axiomatic pattern used to annotate data about *snow thickness*. This annotation axiom was successfully queried by user stories of intermediate expertise and makes use of the following axiom: PATO:*thickness* [PATO_0000915] and (RO:*inheres in* [RO_0000052] some ENVO:*snow*[ENVO_01000406])). 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:

```
1 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 vs. a more pragmatic linked data approach intended to make data easily mobilizable. The example of data about *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 phenomenon, questions arise about how to address such a trade-off. For example data about *mean snow thickness*, which was not retrieve by the advanced user stories, is annotated with the axiom: OBCS:*expected value* [OBCS_0000083] and IAO:*is about* [IAO_0000136] min 2 (IAO:*data item* [IAO_0000027] and IAO:*is about* [IAO_0000136] some (PATO:*thickness* [PATO_0000915] and [RO:*inheres in* [RO_0000052] some ENVO:*snow*[ENVO_01000406])). 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 different owl syntax for handing restrictions about *some* vs. *min* or *max*. Performing SPARQL queries 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 access 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 phenomenon of interest, satisfying both the ontology philosopher and the pragmatic mobilizer of data.

5. Conclusion

In conclusion this work has demonstrated that ontologies are fit for purpose to perform a variety of tasks related to the interconnection and mobilization of interdisciplinary data. Ontologies can be used to represent expert knowledge about environmental phenomena such as plankton ecology with machine-actionable semantics. Ontology terms could additionally serve to track the provenance of information, datasets, specimens and scientific publications, helping to the scientific community to overcome the limitation of known unknown knowledge.

Ontologies can aid to mobilize interdisciplinary datasets, by annotating and querying datasets using ontology terms. Ontology terms through their axiomatic relationships could in principal be used to facilitate the assembly previously unconnected but relevant data to perform ecological analyses. More open linked data is required in order to make this possible.

When using ontology terms to annotate data, uniform and repetitive annotation patterns would aid to mobilize annotated data while providing sufficient depth to adequately represent phenomena of interest. Additionally, it is of crucial importance when utilizing ontology terms to mobilize data, to maintain consistent usage of upper level semantic models such as the Basic Formal Ontology and the Relations Ontology, deviation from which could have serious consequences on the effectiveness of data mobilization.

In order for ontologies to help facilitate future networking amongst creators and users of knowledge captured in ontologies, term author annotation could be implemented at the term level. With direct annotations to pull-able and living author identifying systems such as ORCID, helping to track the provenance of term authors.

Ontological knowledge graphs are highly centralized due to the hierarchical structure provided by the BFO upper level semantic model. Such graphs exert scale free properties as an exponential number of nodes can be reached from a linear increase in path length traversed. The majority of nodes, however, have few connections. Hence more axiomatic relationships are required in order for the networks to facilitate the discovery of new information based on connections to stated information. Due to their centralized nature, ontologies knowledge graphs are robust to changes made to lower level terms, but vulnerable to substantial changes made to higher level terms.

Finally this work has demonstrated that interoperable environmental and genomic semantics provided by the Environment and Gene ontologies can be leveraged to generate bioinformatic hypotheses from the comparison of environmentally annotate omics datasets.

6. Outlook

As a first outlook form this work, polar related semantics contributed to the envoPolar subset could be aligned with the dpbedia ontology [90] glacial semantics. Disseminating knowledge output from the Alfred Wegener Institute (AWI) to the open source encyclopedia Wikipedia would serve to simultaneously to improve Wikipedia and improve AWI public outreach. Helping to communicate AWI research outputs, as well as educate the public on polar knowledge.

A second outlook to this work which is currently underway is the creation of polar terminology to be used for the annotation of polar genomic sequence data. Such terminology would be detailed in cryosphere extension paper to the Minimum Information about any (x) Sequence (MIxS) genomic sequence submission standards [91].

A third outlook from this work would be to research strategies for effective data annotation and mobilization. The knowledge gained therefrom could be used to develop tools to automate and standardized the annotation of data, datasets, samples and or specimens with ontology terms, aiding to make published data interoperable, mobilizable, and precisely annotated.

A final outlook from this work could involve the creation of an Polar Environmental Observatory Intelligence System. Such a system would be built using knowledge representation, a branch of artificial intelligence. This would involve using the relational knowledge represented by axioms within the Environment and Gene ontologies, in propositional or first order logic models. These ontologically derived logic models would additionally be given information in the form of ontology term annotated data, allowing for reasoning to be conducted over the knowledge model. Input data from which to seed the model would be sourced from data generated by the AWI HAUSGARTEN and FRAM observatories. Input data could include: AWI autonomous buoy sea ice thickness data, physical and chemical oceanography data, FRAM microbial observatory program genomic data, and NASA earth observatory and National Snow and Ice Data Center satellite chlorophyll and sea ice cover data. Additionally, the Gene, Environment and other related ontologies would be extended to create a rich first order and or propositional logic model. Such a system would be programed to dynamically incorporate new data as it is published, and would be developed to systematically address interdisciplinary questions such as:

> “Does microbial taxon diversity in deep-sea sediments show resilient patterns over seasonal cycles?”
>
> “What metabolic processes are enriched in the microbial communities of multi-year sea ice?”
>
> “Which ranges of nutrients result in high taxon turnover in the epipelagic zone?”
>
> “Which cellular components are enriched during a bloom induced by sea ice retreat?”

References

1. **Rockström J, Steffen W, Noone K, Persson Å, Chapin F *et al.*** Planetary boundaries: Exploring the safe operating space for humanity. *Ecology and Society*;14.

2. **Lenhard J, Lücking H, Schwechheimer H**. Expert knowledge, mode-2 and scientific disciplines: Two contrasting views. *Science and Public Policy* 2006;33:341–350.

3. **Bjurström A, Polk M**. Climate change and interdisciplinarity: A co-citation analysis of IPCC third assessment report. *Scientometrics* 2011;87:525–550.

4. **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.

5. **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.

6. **Cox S, Jones R, Lawrence B, Milic-Frayling N, Moreau L**. Interoperability issues in scientific data management. https://eprints.soton.ac.uk/409015/1/CoxEA06.pdf (2006).

7. **Horsburgh JS, Tarboton DG, Piasecki M, Maidment DR, Zaslavsky I *et al.*** An integrated system for publishing environmental observations data. *Environmental Modelling & Software* 2009;24:879–888.

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

9. **Andreas Rauber, Asmi A, Uytvanck D van, Pröl S**. Data citation of evolving data. https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_150609.pdf (2015).

10. **Brachman R**. *Readings in knowledge representation*. Los Altos, Calif: M. Kaufmann Publishers; 1985.

11. **Takahashi R, Kajikawa Y**. Computer-aided diagnosis: A survey with bibliometric analysis. *International Journal of Medical Informatics* 2017;101:58–67.

12. **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.

13. **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.

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

15. **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.

16. **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.

17. **Lee S**. A theory for polar amplification from a general circulation perspective. *Asia-Pacific Journal of Atmospheric Sciences* 2014;50:31–43.

18. **Calvo F**. *Analysis of breast cancer cell invasion using an organotypic culture system*. s.l.s.l: Springer-Springer; 2017.

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

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

21. **Favé M-J, Lamaze FC, Soave D, Hodgkinson A, Gauvin H *et al.*** Gene-by-environment interactions in urban populations modulate risk phenotypes. *Nature Communications*;9. Epub ahead of print March 2018. DOI: 10.1038/s41467-018-03202-2.

22. **Ainsworth C**. Systems ecology: Biology on the high seas. *Nature* 2013;501:20–23.

23. **Ledford H**. Microbes reveal extent of biodiversity. *Nature* 2007;446:240–240.

24. **Karl DM, Lukas R**. The hawaii ocean time-series (hot) program: Background, rationale and field implementation. *Deep Sea Research Part II: Topical Studies in Oceanography* 1996;43:129–156.

25. **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.

26. **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.

27. **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.

28. **Johansen R**. *Get there early : Sensing the future to compete in the present*. San Francisco, Calif: Berrett-Koehler Publishers; 2007.

29. 1st AtlantOS Briefing Paper AtlantOS. https://www.atlantos-h2020.eu/2017/02/10/1st-atlantos-briefing-paper/ (accessed 28 February 2018).

30. FixO3 Fixed-point Open Ocean Observatories. http://www.fixo3.eu/ (accessed 3 May 2017).

31. **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.

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

33. Introduction to BCO-DMO BCO-DMO. https://www.bco-dmo.org/ (accessed 28 February 2018).

34. **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.

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

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

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

38. **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).

39. **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).

40. **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.

41. Hurwitz Lab. http://www.hurwitzlab.org/ (accessed 28 February 2018).

42. Ocean Cloud Commons Hurwitz Lab. http://www.hurwitzlab.org/projects/ocean-cloud-commons/ (accessed 28 February 2018).

43. Project Planet Microbe. http://www.planetmicrobe.org/project/ (accessed 28 February 2018).

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

45. United Nations Decade of Ocean Science for Sustainable Development (2021-2030). *UNESCO*. https://en.unesco.org/ocean-decade (2017, accessed 28 February 2018).

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

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

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

49. Sensor Observation Service OGC. http://www.opengeospatial.org/standards/sos (accessed 28 February 2018).

50. **Compton M, Barnaghi P, Bermudez L, García-Castro R, Corcho O *et al.*** The SSN ontology of the w3c semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web* 2012;17:25–32.

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

52. **Figueiredo AS**. Data sharing: Convert challenges into opportunities. *Frontiers in Public Health*;5. Epub ahead of print December 2017. DOI: 10.3389/fpubh.2017.00327.

53. **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.

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

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

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

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

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

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

60. **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).

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

62. **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).

63. **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.

64. **Buttigieg PL, Mungall C, Blumberg K, renzo, uberon**. EnvironmentOntology/envo: Ecotone. Epub ahead of print May 2017. DOI: 10.5281/zenodo.573849.

65. **Mungall C, Buttigieg PL, Blumberg K, renzo, uberon**. EnvironmentOntology/envo: Polar express. Epub ahead of print April 2017. DOI: 10.5281/zenodo.546433.

66. **Mungall C, Buttigieg PL, renzo, Blumberg K, uberon**. EnvironmentOntology/envo: Hot tub time machine. Epub ahead of print March 2017. DOI: 10.5281/zenodo.438339.

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

68. **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.

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

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

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

72. **Buttigieg PL, Mungall C, Blumberg K, Laporte M-A, renzo *et al.*** EnvironmentOntology/envo: Planetary ecology. Epub ahead of print August 2017. DOI: 10.5281/zenodo.846451.

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

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

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

76. **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.

77. **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.

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

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

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

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

82. 9.7. Statistics — Mathematical statistics functions — Python 3.6.4 documentation. https://docs.python.org/3/library/statistics.html (accessed 23 February 2018).

83. User Story – a journey into software best practices with maria grazia merlo. https://mariagraziamerlo.com/tag/user-story/ (accessed 5 March 2018).

84. **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.

85. **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.

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

87. **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.

88. **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.

89. **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.

90. **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.

91. **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.

92. **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.

93. **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.

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

95. **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).

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

97. **Sun S, Chen J, Li W, Altintas I, Lin A *et al.*** Community cyberinfrastructure for advanced microbial ecology research and analysis: The CAMERA resource. *Nucleic Acids Research* 2010;39:D546–D551.

98. **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.

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

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

101. **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.

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

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

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

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

106. **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.

107. **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.

108. **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.

109. **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.

110. **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.

111. **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.

112. **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.

113. **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.

114. **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.

115. **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.

116. **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.

117. **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.

118. **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.

119. **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.

120. **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.

121. **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.

122. **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.

123. **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.

124. **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.

125. **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

A.1 Semantic science adaption of scientific method

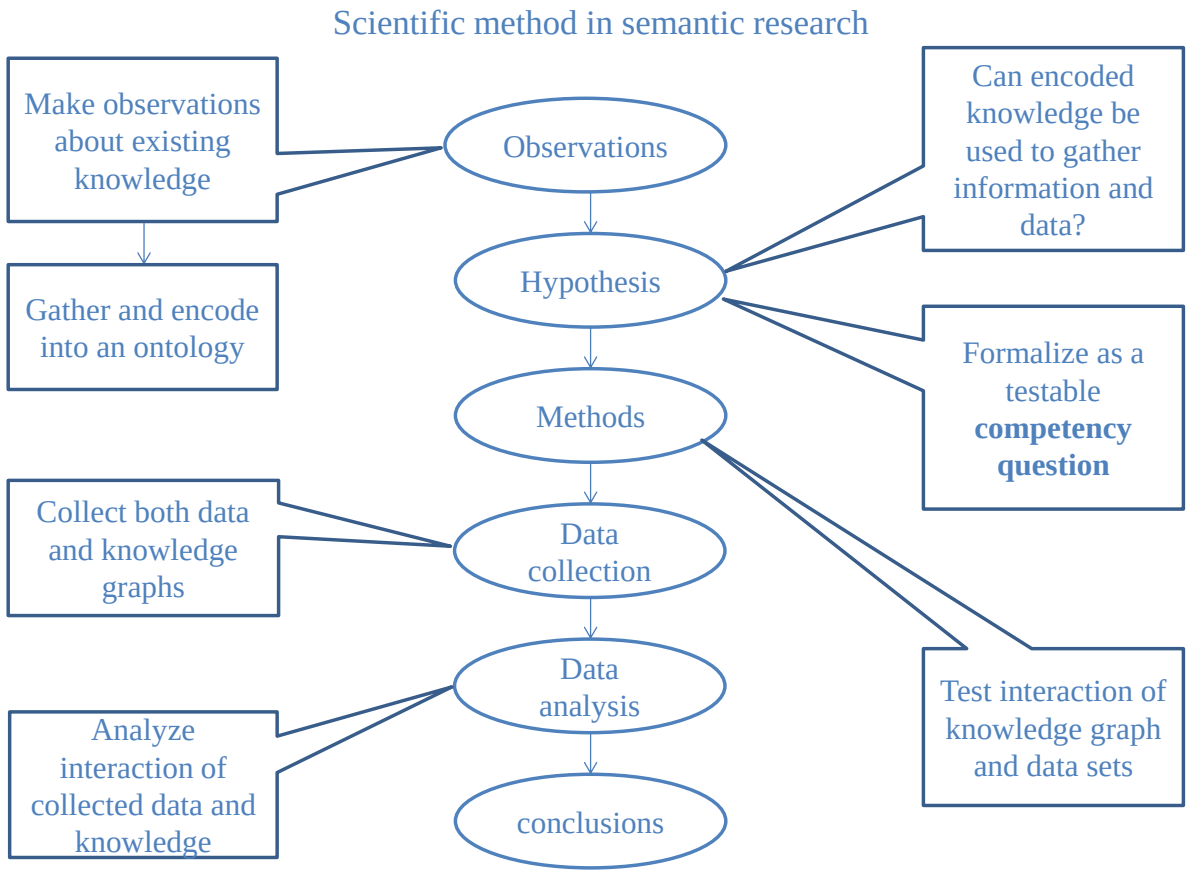

Figure A1 Shows the adaptation of semantic science to the scientific method, facilitate the reuse of published data. Item **A** shows the standard scientific method where problems are defined, hypotheses are generated, experiments are conducted to test hypotheses either generating publishable data or leading to the definition of new problems. Item **B** shows the workflow by which to convert published data to machine-actionable formats, annotate with ontology terms as meta-data. This facilitates the process of asking questions of open linked data and gather data to test hypotheses.

A.2 Model polar datastore creation

A Model data store of environmental and genomic data was created during this work. Detailed description of axioms making use of ontology terms used to post-compositionally annotate the example polar datastore are available from: https://github.com/kaiiam/kblumberg_masters_thesis/wiki/complete_datastore

The following datasets were included in the example datastore:

1. Inorganic nutrients measured on water bottle samples at AWI HAUSGARTEN during POLARSTERN cruise MSM29. [106]

2. Physical oceanography and current meter data from mooring TD-2014-LT. [107]

3. Chlorophyll a measured on water bottle samples during POLARSTERN cruise ARK-XXIV/2. [108][92]

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. [87][88]

5. Biogenic particle flux at AWI HAUSGARTEN from mooring FEVI7. [109][110]

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. [111][112]

7. Sea ice thickness at Ice Camp 1 on 2013-09-01 (GEM2IceTh_DiveHole_IceStation1). [113][90]

8. Influence of snow depth and surface flooding on light transmission through Antarctic pack ice, supplementary data. [89][90]

9. Ice-algal chlorophyll a and physical properties of multi-year and first-year sea ice of core CASIMBO-CORE-2_11. [114][115]

10. Unpublished metagenomic data from deep sea sediments from Hausgarten POLARSTERN Polarstern cruise PS85, encompassing both functional genomic data in the form of preprocessed pfam2go tables as well as 16S taxonomic data, courtesy of Josephine Z. Rapp.

11. Unpublished transcriptomic data from shallow Helgoland Marine Sediments during a spring phytoplankton bloom, encompassing preprocessed pfam2go tables, courtesy of David Probandt, and Matthew Schechter.

A.3 Metagenomic and metatranscriptomic data

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 ENVO:*marine bathyal zone biome* [ENVO_01000026] Samples 3 and 4 were collected from depths of 3531m and 5525m which best correspond to ENVO:*marine abyssal zone biome* [ENVO_01000027]

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 ENVO:*sandy sediment* [ENVO_01000118] from an ENVO:*marine neritic benthic zone biome* [ENVO_01000025]. The first 4 samples: labeled X1-X4 were used.

Table A1 Results of the relative genomic and transcriptomic proportions of GO:*transition metal ion transport* [GO_0000041] process in various types of ENVO:*marine benthic biome*s [ENVO_01000024], queried for in the example datastore.

term	ENVO:*marine abyssal zone biome* [ENVO_01000027]	ENVO:*marine bathyal zone biome* [ENVO_01000026]	ENVO:*marine neritic benthic zone biome* [ENVO_01000025]
GO:*ferrous iron transport* [GO_0015684]	0.04	0.04	0.02
GO:*mercury ion transport* [GO_0015694]	0.01	0.00	0.00
GO:*nickel cation transmembrane transport* [GO_0035444]	0.00	0.00	0.00
GO:*transition metal ion transport* [GO_0000041]	0.00	0.00	0.00
GO:*iron ion transmembrane transport* [GO_0034755]	0.00	0.00	0.00
GO:*iron ion transport* [GO_0006826]	0.00	0.00	0.00
GO:*copper ion transmembrane transport* [GO_0035434]	0.00	0.00	0.00
GO:*cobalt ion transport* [GO_0006824]	0.00	0.00	0.00
GO:*copper ion transport* [GO_0006825]	0.00	0.00	0.00

Table A2 Results of the relative genomic and transcriptomic proportions of GO:*transition metal ion binding* [GO_0046914] molecular functions in various types of ENVO:*marine benthic biome*s [ENVO_01000024], queried for in the example datastore.

term	ENVO:*marine abyssal zone biome* [ENVO_01000027]	ENVO:*marine bathyal zone biome* [ENVO_01000026]	ENVO:*marine neritic benthic zone biome* [ENVO_01000025]
GO:*transition metal ion binding* [GO_0046914]	0.70	0.67	0.52

term	ENVO:*marine abyssal zone biome* [ENVO_01000027]	ENVO:*marine bathyal zone biome* [ENVO_01000026]	ENVO:*marine neritic benthic zone biome* [ENVO_01000025]
GO:*cobalt ion binding* [GO_0050897]	0.69	0.73	1.07
GO:*ferric iron binding* [GO_0008199]	0.22	0.21	0.62
GO:*nickel cation binding* [GO_0016151]	0.12	0.13	0.12
GO:*molybdenum ion binding* [GO_0030151]	0.05	0.05	0.10
GO:*manganese ion binding* [GO_0030145]	0.04	0.04	0.02
GO:*iron ion binding* [GO_0005506]	0.04	0.04	0.03
GO:*zinc ion binding* [GO_0008270]	0.03	0.03	0.11
GO:*ferrous iron binding* [GO_0008198]	0.02	0.03	0.00
GO:*copper ion binding* [GO_0005507]	0.01	0.01	0.02
GO:*copper chaperone activity* [GO_0016531]	0.00	0.00	0.00

Thing
→ GO:***biological_process*** [GO_0008150]
→ GO:*cellular process* [GO_0009987]
→ GO:*cellular metabolic process* [GO_0044237]
→ GO:*organic acid metabolic process* [GO_0006082]
→ GO:*oxoacid metabolic process* [GO_0043436]
→ GO:*carboxylic acid metabolic process* [GO_0019752]
→ *GO:cellular amino acid metabolic process* [GO_0006520]
→ GO:***cellular amino acid biosynthetic process*** [GO_0008652]
→ GO:*alpha-amino acid biosynthetic process* [GO_1901607]
→ GO:***serine family amino acid biosynthetic process*** [GO_0009070]

→ GO:***glycine biosynthetic process*** [GO_0006545]
→ GO:*glycine biosynthetic process from serine* [GO_0019264]

→ GO:*cysteine biosynthetic process* [GO_0019344]
→ GO:***cysteine biosynthetic process from serine*** [GO_0006535]

Figure A2 GO:*biological process* [GO_0008150] hierarchy differentiating ENVO:*marine abyssal zone biome* [ENVO_01000027] and ENVO:*marine bathyal zone biome* [ENVO_01000026] from ENVO:*marine neritic benthic zone biome* [ENVO_01000025] ENVO:*marine sediments* [ENVO_03000033].

Subclasses of boldfaced terms GO:*biological process* [GO_0008150], GO:*cellular amino acid biosynthetic process* [GO_0008652], and GO*serine family amino acid biosynthetic process* [GO_0009070] are the subjects of the Principal coordinate analyses plots in **Figures 4**, **5** and **6** respectably. GO*serine family amino acid biosynthetic process* [GO_0009070] terms differentiating ENVO:*marine abyssal zone biome* [ENVO_01000027] and ENVO:*marine bathyal zone biome* [ENVO_01000026] from ENVO:*marine neritic benthic zone biome* [ENVO_01000025] ENVO:*marine sediments* [ENVO_03000033], shown in **Figure 6**, are GO:*glycine biosynthetic process* [GO_0006545], and GO:*cysteine biosynthetic process from serine* [GO_0006535].

A.4 Ecological analysis of ontology-collected environmental data

As a demonstration of how an ecological analysis could be conducted on machine-actionable data collected via an ontology semantics, I performed a principal component analysis. The data was collected by searching the datastore for data annotated with terms which are present in the axioms of the hypothetical ontology term ENVO:*environmental system determined by a community* [URI pending]

This mock analysis used the collected data to investigate which environmental variables have the greatest loading on the analysis. **Figure A3** shows a hypothetical principal component analysis showing the effects of the various environmental variables, assembled due to their inclusion in axioms of the term ENVO:*environment determined by a phytoplankton community associated with sea-ice*. The first two PCA axes explain 53.6% of the variance in this analysis with PCA axis 1 explaining 34.0% of variance and PCA axis 2 explaining 19.6% of variance. Included in **Figure A3** are the Environment Ontology terms which were referenced as axioms of the hypothetical ENVO:*environment determined by a phytoplankton community associated with sea-ice* term, and from which annotated data was retrieved. For example *SignalStrength_ENVO_00002200* represents a column which is labeled *Signal Strength* which is about a PATO:*degree of illumination* [PATO_0015013] which RO:*inheres in* [RO_0000052] some ENVO:*sea ice* [ENVO_00002200].

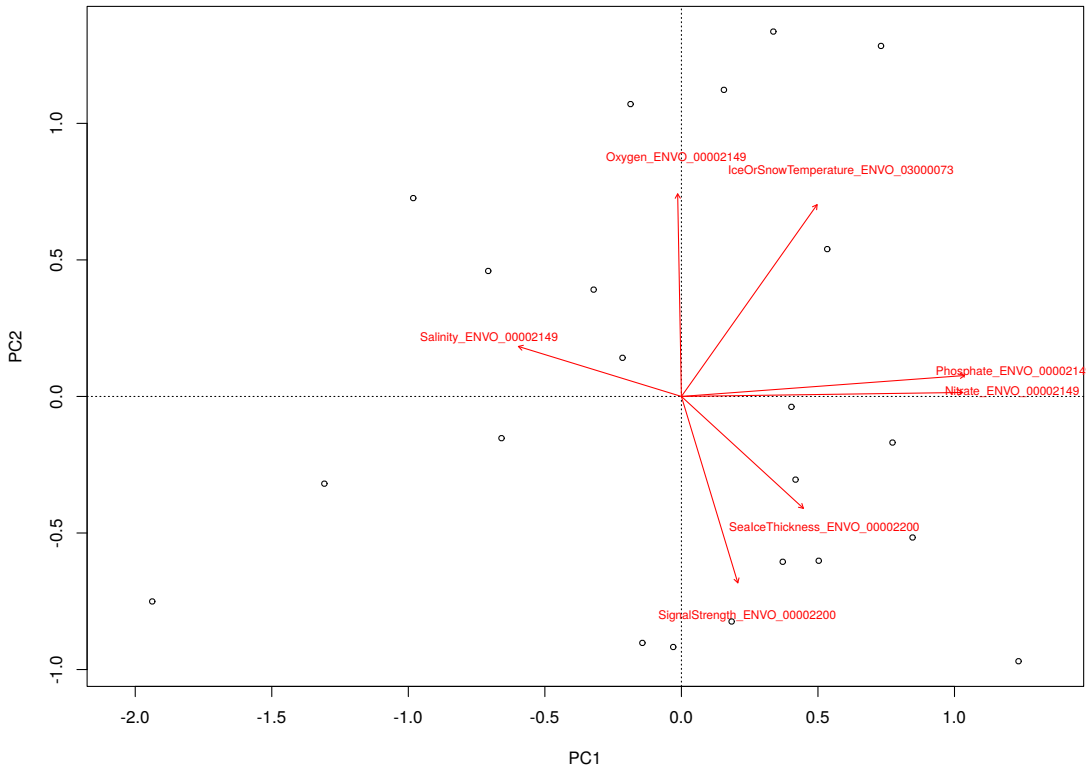

Figure A3 PCA on assembly of data about terms included as axioms of a hypothetical ENVO:*environment determined by a phytoplankton community associated with sea-ice* term.

Table A3 Shows the loadings from principal components 1 and 2 from the PCA conducted on data assembled due it being annotated with ontology term included in the subclass axioms of a hypothetical ENVO:*environment determined by a phytoplankton community associated with sea ice* ontology term. Terms are ordered in descending order based on PC 1 then PC 2. The first row of the table is a data column about the PATO:*concentration of* [PATO_0000033] CHEBI:*phosphate* [CHEBI_26020] in ENVO:*sea water* [ENVO_00002149].

data columns	annotation term	PC1 loading	PC2 loading
phosphate	ENVO:*sea water* [ENVO_00002149]	1.22009153	0.09015633
nitrate	ENVO:*sea water* [ENVO_00002149]	1.21200068	0.01720033
ice or snow temperature	ENVO:*multiyear ice* [ENVO_03000073]	0.58457003	0.82623573
sea ice thickness	ENVO:*sea ice* [ENVO_00002200]	0.52555148	-0.48198612
signal strength	ENVO:*sea ice* [ENVO_00002200]	0.24304142	-0.80299701
oxygen	ENVO:*sea water* [ENVO_00002149]	-0.01611319	0.87287456
salinity	ENVO:*sea water* [ENVO_00002149]	-0.70229106	0.21544285

Example results of this hypothetical analysis are that ENVO:*sea water* [ENVO_00002149] CHEBI:*phosphate* [CHEBI_26020] and CHEBI:*nitrate* [CHEBI_17632] concentrations have positive PC1 loading values. ENVO:*sea water* [ENVO_00002149] CHEBI:*dioxygen* [CHEBI_15379] and CHEBI:*salinity* [URI pending] have negative PC1 loading values. ENVO:*sea ice* [ENVO_00002200] PATO:*thickness* [PATO_0000915], and signal strength PATO:*degree of illumination* [PATO_0015013] which RO:*inheres in* [RO_0000052] some ENVO:*sea ice* [ENVO_00002200] have negative PC2 loading values.

A.5 Marine biome associated DOIs

Table A4 Complete list of digital object identifiers of publications obtained querying for references of datasets which are about BFO:*part of* [BFO_0000050] an ENVO:*marine biome* [ENVO_00000447].

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 [91]
	10.1016/j.pocean.2009.07.011	Dimethylsulphide, DMSP-lyase activity and microplankton community structure inside and outside of the Mauritanian upwelling [116]

data annotation	reference doi	reference title
	10.5194/bg-9-1041-2012	Environmental control on the variability of DMS and DMSP in the Mauritanian upwelling region [117]
	10.3390/rs61010089	Global Retrieval of Diatom Abundance Based on Phytoplankton Pigments and Satellite Data [118]
	10.5194/bgd-10-12115-2013	Photophysiological state of natural phytoplankton communities in the South China Sea and Sulu Sea [119]
	10.1016/j.dsr.2014.12.010	Physiological characteristics of open ocean and coastal phytoplankton communities of Western Antarctic Peninsula and Drake Passage waters [120]
	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 [121]
	10.5194/bg-10-3297-2013	Sulphur compounds,methane, and phytoplankton: Interactions along a north-south transit in the western Pacific Ocean [122]

data annotation	reference doi	reference title
	10.3402/polar.v34.23349	Summertime plankton ecology in Fram Strait-a compilation of long- and short-term observations [92]
	10.5194/essd-5-109-2013	The MAREDAT global database of high performance liquid chromatography marine pigment measurements [123]
	10.1029/2003EO380001	Unique data repository facilitates ocean color satellite validation [124]
	10.5194/os-11-139-2015	Using empirical orthogonal functions derived from remote-sensing reflectance for the prediction of phytoplankton pigment concentrations [125]
	10.1029/2005JC003207	Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll [93]
influence snow depth	10.1002/2016JC012325	Influence of snow depth and surface flooding on light transmission through Antarctic pack ice [90]

A.6 Glacial community consultation working group participants

Participants in the Feb 2, 2018 VoCamp Glacier Ontology Hackathon are listed in the following **Table A2**.

Table A5 Participants in the Feb 2, 2018 February VoCamp Glacier Ontology Hackathon and community semantic consultation session.

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

A.7 Network analysis supplemental figures

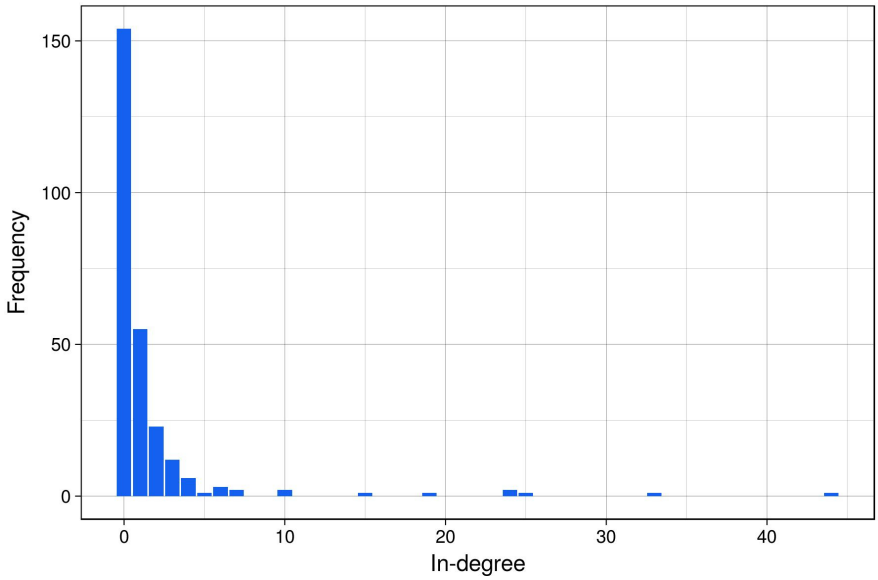

Figure A4 In degree distribution of the envoPolar subset analyzed as a network.

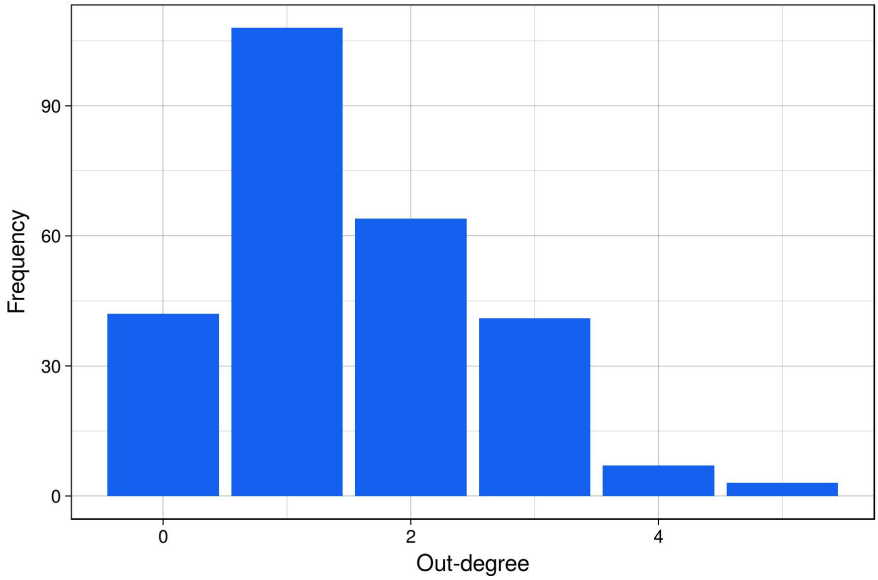

Figure A5 Out degree distribution of the envoPolar subset analyzed as a network.

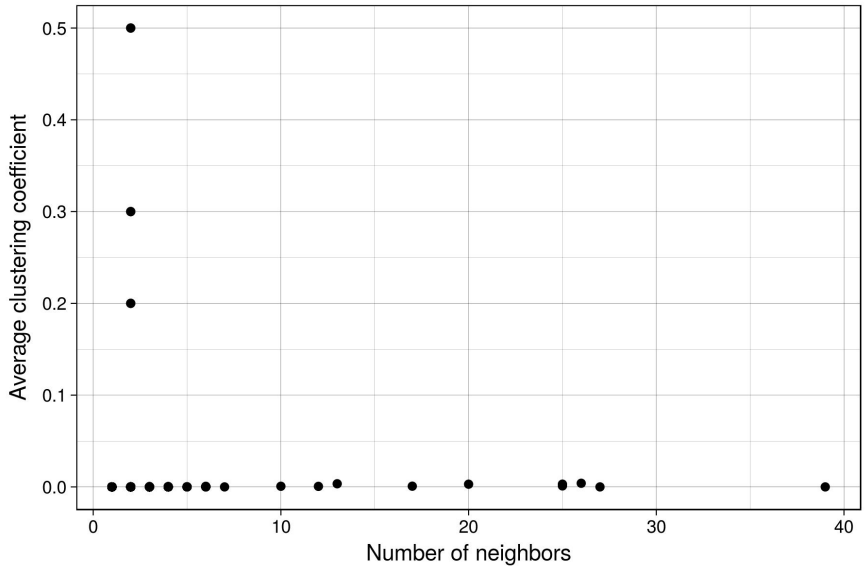

Figure A6 Plot of betweenness centrality as a function of number of neighbors of nodes from the envoPolar subset analyzed as a network.

A.8 Estimated user story simulated querying expertise

Estimated user stories were developed to evaluate the proficiencies of users of basic, intermediate or advanced querying expertise for performing semantic queryies of the model datastore.

The predicted users were tested with two mobilization tasks. 1) mobilizing annotated with an exclusive *and* intersection of two ontology terms, 2) mobilizing data annotated as being about parts associated with an ontology term.

Exclusive *and* querying cases used in the first task involved the following pairs of ontology terms: NCBITaxon:*Bacteria* [NCBITaxon_2] and NCBITaxon*Archaea* [NCBITaxon_2157], PATO:*concentration of* [PATO_0000033] and CHEBI*chlorophyll a* [CHEBI_18230], ENVO:*marine current* [ENVO_01000067] and PATO:*velocity* [PATO_0002242], PCO:*microbial community* [PCO_1000004] and PCO:*species diversity* [PCO_0000019], ENVO:sea ice* [ENVO_00002200] and PATO:depth* [PATO_0001595], ENVO:sea ice* [ENVO_00002200] and PATO:*temperature* [PATO_0000146], ENVO:*sea water* [ENVO_00002149] and CHEBI*chlorophyll a* [CHEBI_18230], ENVO:*snow* [ENVO_01000406] and PATO:*thickness* [PATO_0000915].

Data about parts associated with an ontology term querying cases used in the second task involved the following term annotations by which to query for data: ENVO:*brine channel* [ENVO_03000041], CHEBI:*carbon atom* [CHEBI_27594], CHEBI:*cation* [CHEBI_36916], IAO:*centrally registered identifier symbol* [IAO_0000577], ENVO:*glacier* [ENVO_00000133], ENVO:*marine biome* [ENVO_00000447], ENVO:*melt pond* [ENVO_03000040], and finally ENVO:*ocean* [ENVO_00000015].

The predicted user stories were programed to have knowledge of a subset of the turtle data annotation querying case property paths used in the example datastore. The querying cases the various user stories were programed

to include are shown in **Table A6**.

Table A6 Shows the SPARQL querying cases which were included for each of the predicted user stories of various querying expertise levels.

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

The following code block shows sudo-code representations of the querying cases which were included in the user story estimating basic querying expertise.

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

The following code block shows sudo-code representations of the querying cases which were included in the user story estimating intermediate querying 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))
```

The following code block shows sudo-code representations of the querying cases which were included in the user story estimating advanced querying 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))))
```

The following code block shows sudo-code representations of all the querying cases which were included to query for all annotated data included in the model datastore.

```
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)))
```

A.8 Github repository

All material used and generated in the course of this work is available from the github `kblumberg_masters_thesis` repository available from: https://github.com/kaiiam/kblumberg_masters_thesis/

The Wiki page associated with the github repository is available from: https://github.com/kaiiam/kblumberg_masters_thesis/wiki