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 **9**

Introduction **10**

Rapid effects of climate change on Polar systems 10

Microbes and Biogeochemical cycles 10

Arctic climate change 10

AtlantOS 10

FRAM & HAUSGARTEN 10

Make better use of the generated data **11**

Need for semantics in Environmental data 11

linked data 11

semantic web 12

linked open data 12

open science 12

FAIR 13

OPeNDAP 13

Internet of things 13

UN decade of ocean science for sustainable development 2021-2030. 13

ontology management of big data 13

Ontologies and the OBO Foundry 13

ENVO for representing environmental semantics. 14

Gene Ontology 14

Example Ontology uses 15

PANGAEA 15

role of data in 2015 - 2020 ESIP Strategic Plan 15

SDGIO 15

Policy and SDGIOs 15

UN sustainability development goals in response to climate change 16

Linking earth science data initiatives such ESIP Open knowledge network to the UN SDGIO's . . 16

knowledge outreach 16

competency questions: 17

Materials and Methods **17**

Datasets used in Datastore 17

programs used: 17

Semantic Data Annotation 18

sparql query scripting 18

Results 18
Competency Questions . . . 19
Lookup author of ontology term . . . 19
Retrieve any data which is about a subclass of sea ice . . . 20
What compounds play a role as algae metabolites? . . . 20
Hackathon competency question . . . 21
make one or cc's by breaking the OR statement in my scripts . . . 22
is it possible to move annotated data . . . 22
what level of granularity . . . 22
Can a semantic knowledge graph guide scientists toward data which is relevant to a natural process they are studying? . . . 22
Connecting GO and ENVO terms . . . 23
biogeochemical cycling . . . 23
PCO contributions & Plankton Ecology . . . 24
Tilman Satelite Data . . . 24
cryoMIxS . . . 24
ENVO releases of interest: . . . 24
post compositional data annotation model . . . 25
Example of Post Compositional Data Annotation with Ontology Terms . . . 25
Vocamp Virtual Glacial Hackathon . . . 26
AWI DBPEDIA contributions . . . 26
UNEP SDGIO . . . 26

Discussion 26
Querying Semantically Annotated Data . . . 26
Creating Classes vs post compositional annotation for data annotation . . . 27
VOCAMP Discussion section: . . . 27
consistent data structures for published data . . . 27
Semantics as AWI Public Outreach . . . 28

Conclusion 28

Outlook 28

References 30

Appendices 34
Python Scripts and Documentation . . . 34
script 1 34
script 2 36

Summary

Well this about sums it up xP

Introduction

Rapid effects of climate change on Polar systems

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

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

Microbes and Biogeochemical cycles

//maybe this can be fit in here? perhaps start with a microbial perspective to keep marmic happy?

> The prokaryotic and eukaryotic microorganisms that drive the pelagic ocean's biogeochemical cycles are currently facing an unprecedented set of comprehensive anthropogenic changes [2]

Arctic climate change

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

cite and discuss some Arctic climate change research …

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

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

AtlantOS

//maybe mention this?

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

FRAM & HAUSGARTEN

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

HAUSGARTEN intro: [3]

FRAM intro: [4]

Make better use of the generated data

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

Need for semantics in Environmental data

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

FROM [5] paraphrase and harvest useful introductory material

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

linked data

wiki

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

This is done by making use of standard web technologies.

Linked data makes use of Hypertext Transfer Protocol (HTTP) to give data objects a web address, as well as the Resource Description Framework (RDF) [6] a … to share information in a machine-readable format. This allows for

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

semantic web

wiki

> The Semantic Web is an extension of the World Wide Web through standards by the World Wide Web Consortium (W3C). [7] The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF). [8]
>
> According to the W3C, “The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries”.[2] The term was coined by Tim Berners-Lee for a web of data that can be processed by machines[3]—that is, one in which much of the meaning is machine-readable.
>
> Linked data may also be open data, in which case it is usually described as linked open data (LOD).

linked open data

read and cite [9] about linked open data arguments presented by tim_berners-lee and on the wiki page on open data https://en.wikipedia.org/wiki/Open_data

from the wiki: citing [9] > Open data is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.

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

parlalles with open science wiki

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

open science

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

FAIR

the FAIR data guiding principles (machine-focused findability, accessibility, interoperability reusability) [11]

AWI data is currently Findable and accessable at a high level for example within Pangaea files. Improvements would be to make the data findable and accessible. Improve Polar data re-usability with the cryo-MIXS extension paper in prep. Most importantly Interoperability, a formally controlled and machine accessible vocabulary, through ontologies, (ENVO, PATO, PCO, ECOCORE).

OPeNDAP

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

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

Internet of things

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

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

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

ontology management of big data

For example the HASNetO ontology [12] > has been in use to support the data management of a number of large-scale ecological monitoring activities (observations) and empirical experiments.

Ontologies and the OBO Foundry

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

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

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

//revise a bit from lab rotation: Because, no single knowledge graph can encompass the needs of interdisciplinary projects, work must be done in a coordinated fashion with other ontology researchers and developers. In order to interconnect ontologies representing scientific knowledge from different domains, the Open Biological and Biomedical Ontology (OBO) Foundry and Library was created [13]. The OBO Foundry and Library established a set of principles by which to develop and coordinate ontologies such that the scientific knowledge they represent and hence the data they link can interoperate. These ontologies share a common upper level in the hierarchy and use of the same types of logical connective operations to interlink their knowledge. Following these principles are a family of ontologies representing scientific knowledge from non-overlapping domains, which can be used in combination to describe natural phenomena in greater depth. OBO compliant ontologies make use of the Basic Formal Ontology (BFO) [14] [15] [16], to ensure they have a compatible hierarchical structure, and use logical relations from the Relations Ontology (RO) [17], to standardize the connections between their knowledge.

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

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

ENVO for representing environmental semantics.

ENVO papers: [18] [19]

The Environment Ontology (ENVO) represents expert knowledge about different types of environments[18][19]. ENVO is an OBO aligned ontology.

Environmental knowledge represented by ENVO is used to annotate data from a variety of life science disciplines including oceanography and polar research. [18][19]

Gene Ontology

go paper: [20]

GO frequently used to interpret omic data [20]. It has been used to do genomewide RNA expression profile data to compare samples based on shared biological pathways. [21]

The combination of GO and ENVO is less frequently attempted. [22]

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

Example Ontology uses

A communal catalogue reveals Earth's multiscale microbial diversity. //Uses EMPO a light-weight application ontology built on ENVO the Earth Microbiome Project Ontology [23] //good to have an example which demonstrates the utility of ENVO for an application ontology to provide utility.

//from my rotation rewrite example > Thesen et al.13. show how such a federated semantic approach can enhance handling of environmental and phenotype data, in order to ask increasingly complex questions such as "Which crop varieties are expected to do well in a particular location over the next century?". Thesen et al Emerging semantics to link phenotype and environment [24]

PANGAEA

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

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

role of data in 2015 - 2020 ESIP Strategic Plan

link to my log

SDGIO

Policy and SDGIOs

Making Marine Life Count: A New Baseline for Policy [25] Just use a little bit from this as policy intro.

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

UN sustainability development goals in response to climate change

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

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

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

14.1

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

link the nitrogen phosphorus data to the concept of those cycle being out of balance as doccumented in the Planetary Boundaries: Exploring the Safe Operating Space for Humanity paper. [26]

United Nations Environment Programme

SDGIO is an OBO compliant ontology

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

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

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

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

knowledge outreach

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

competency questions:

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

Materials and Methods

Datasets used in Datastore

1. Inorganic nutrients measured on water bottle samples at AWI HAUSGARTEN during POLARSTERN cruise MSM29. [27]
2. Physical oceanography and current meter data from mooring TD-2014-LT. [28]
3. Chlorophyll a measured on water bottle samples during POLARSTERN cruise ARK-XXIV/2. [29][30]
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. [31][32]
5. Biogenic particle flux at AWI HAUSGARTEN from mooring FEVI7. [33][34]
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. [35][36]
7. Sea ice thickness at Ice Camp 1 on 2013-09-01 (GEM2IceTh_DiveHole_IceStation1). [37][38]
8. Ice-algal chlorophyll a and physical properties of multi-year and first-year sea ice of core CASIMBO-CORE-1_10. [39][40]
9. Ice-algal chlorophyll a and physical properties of multi-year and first-year sea ice of core CASIMBO-CORE-2_11. [41][40]
10. Unpublished metagenomic data from deep sea sediments from Hausgarten POLARSTERN Polarstern cruise PS85, encompassing both functional genomic data, and 16S taxonomic data, courtesy of Josephine Z. Rapp.

programs used:

semantic technologies make use of the specifications of the World Wide Web Consortium (W3C) [7]

SPARQL 1.1 Query Language and W3C Recommendation 21 March 2013 query language for RDF [42]

Python [43] Version 2.7.12

RDF 1.1 Concepts and Abstract Syntax W3C Recommendation 25 February 2014 [6]

RDF specifications turtle [8],

Anything To Triples (any23) a library, a web service and a command line tool that extracts structured data in RDF format from a variety of Web documents [44].

owl [45]

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

Protégé [46] [47]

Semantic Data Annotation

Semantic annotation of example data was conducted in the RDF serialization turtle, drawing upon its blank node feature to facilitate scripting owl code in RDF. Annotations make use ontology terms from the OBO Foundry [13]. Ontology terms can be search for using Ontobee A linked data server hosting ontologies and their terms. [48]

sparql query scripting

scripts to perform queries were written in python verion?

using the rdf-lib module

Queryies preformed against the ontobee endpoint http://sparql.hegroup.org/sparql/ a serive provied by the He Group [48]

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

Results

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

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

Competency Questions

experiments to test knowledge model against competency questions.

Lookup author of ontology term

see my thesis here

The Internet is enabling collaborative dissemination of knowledge and data

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

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

An example question question to this effect asks:

> “What are the email addresses of all authors who contributed to ontology classes about any kind of ‘sea ice’?”

Evaluating if the semantic layer encoded into ontologies is capable of answering this question, we begin by sending the following query to the OBO ontology knowledge graph:

```
PREFIX obo: <http://purl.obolibrary.org/obo/>
SELECT DISTINCT ?term (STR(?label) as ?label) (STR(?author) as ?author)
WHERE
{
  ?term rdfs:subClassOf+ obo:ENVO_00002200;
        rdfs:label ?label;
        obo:IAO_0000117 ?author .
}
ORDER BY ?term
```

This query

Table 1: Authors information about contributors to sea ice classes.

label	author
first year ice	http://orcid.org/0000-0002-3410-4655, http://orcid.org/0000-0002-4366-3088
second year ice	http://orcid.org/0000-0002-3410-4655, http://orcid.org/0000-0002-4366-3088
multiyear ice	http://orcid.org/0000-0002-3410-4655, http://orcid.org/0000-0002-4366-3088

Retrieve any data which is about a subclass of sea ice

//easy to bang out see my thesis here

What compounds play a role as algae metabolites?

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

purl

querying the ontobee sparql endpoint

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

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

using the restriction has role.

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

Returning the following results:

Table 2: Compounds serving as algal metabolites.

purl	label
http://purl.obolibrary.org/obo/CHEBI_17992	Sucrose
http://purl.obolibrary.org/obo/CHEBI_80716	aplysiatoxin
http://purl.obolibrary.org/obo/CHEBI_90820	11(R)-HEPE(1-)
http://purl.obolibrary.org/obo/CHEBI_86386	3-mercaptopropionate
http://purl.obolibrary.org/obo/CHEBI_17754	Glycerin
http://purl.obolibrary.org/obo/CHEBI_17754	glycerol
http://purl.obolibrary.org/obo/CHEBI_16810	2-oxoglutarate(2-)
http://purl.obolibrary.org/obo/CHEBI_16914	salicylic acid
http://purl.obolibrary.org/obo/CHEBI_16914	Salicylic Acid
http://purl.obolibrary.org/obo/CHEBI_16811	Methionine

Hackathon competency question

> To what extend are polar semantics encoded into the environment ontology accepted by domain experts such as glaciologists.

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

tuned into and deliberated expert knowledge

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

table of how many and which of those terms are semantically represented in ENVO? perhaps a column for number of links to constituents? Pier says yes.

Final envo representation relative to vocamp consensus diagram.

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

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

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

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

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

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

make one or cc's by breaking the OR statement in my scripts

to fetch less of the data, taking only the simpler querying cases and compare how much data I get back make a table, could use the output from a combination of several scripts perhaps even split into several competency questions.

is it possible to move annotated data

> "Do we have any data about sea ice and if so can we use semantic annotations to retrieve such data?"

what level of granularity

> "Do we have any data about the depth of any type of sea ice?" "About the temperature of sea ice.?"
> or depth or sea ice.

Can a semantic knowledge graph guide scientists toward data which is relevant to a natural process they are studying?

for example:

> What kinds of data could be used to assess the phenomena of ice formation?

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

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

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

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

Also potentially valuable to asses such sea water for it's potential to freeze are data about nitrate, phosphate and chlorophyll, as they are indicators of the biotic activities in such seawater which are most likely linked to the extent to which the ice is freezing. This leads to the generation of more hypothesis about relationships between for example sea ice and nutrient concentrations or chlorophyll which can be harvested into the knowledge graph. Examples of which include the occurrence of blooms associated with the melting of sea ice, or the effects of sea ice melting on on water body stratification, which could potentially inhere in the nutrient data. Codifying such relationships into the semantic layer and using that layer to annotate and mobilize data provides a way illuminate the connections between otherwise disparate data sets, for example using nitrate and phosphate data in combination with other data such as temperature, to help report on the potential freezing or melting processes.

Connecting GO and ENVO terms

//I should be able to get two different competency questions out of this.

Can we use the semantic layer to get any genomic annotation data about a given set of molecular functions in any type of X environment for example:

> What 'transition metal ion binding' molecular functions are found in marine benthic biomes?

From a preliminary look it appears that in such biomes 'zinc ion binding', and 'iron ion binding' are much more prevalent than the other transition metal ion binding molecular functions.

//Either extend this example or make a new one in which I find some go terms which differentiate the abyssal and bathyl samples and feed those results into a PCA or 2 as proof of concept for an automated system pipeline which to pulls in data and does some ecological analysis on them. Go through the Go term data and try to find GO term which are different between the samples and then query for subclasses of these and queyr for data about those terms and use those results to do the PCA's. (hopefully one or 2 different ones from the different GO upper level terms.)

biogeochemical cycling

> What types of in interdisciplinary data about the Biogeochemical Silicon Cycle would we expect to access?

for example I get back * Genomic data annotated with GO terms about silicate metabolic process, silicate transmembrane transporter activity, silicate transport,

- concentration of silicate in seawater
- Data about Particulate Silicon Flux (from the Arctic phytoplankton and particle flux under climate change)

PCO contributions & Plankton Ecology

//Assuming I get any of this stuff pushed to PCO and or ENVO. I can still write about the semantics qued for addition even if I don't get to push them.

I may still be able to write about the proposed design patterns for PCO, even if I don't get to submit a pull request.

Tilman Satelite Data

Paper: Diatom Phenology in the Southern Ocean: Mean Patterns, Trends and the Role of Climate Oscillations. [49] //Associated with the plankton ecology project using Tillman satellite chlorophyll data and the plankton bloom ontology classes. **maybe move this to outlook for how this could be used in a system which draws from larger datasets (like my email to Anya explained)** Plus I could also talk about this paper as a motivation for the PCO terms, harvesting expert Domain knowledge for example from Anya's section doing the full cycle of the scientific with semantic questions.

form a hypothesis, test etc.

example of expert awi knowledge to harvest:

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

cryoMIxS

original MIxS paper: [50]

//talk about my contributions to the cryoMIxS project. Including work from my lab rotation.

talk about the annotaion for alreal chlorophyll a, plus some of the other terms used to annotate some of the data in the datastore, will work toward the semantic axiomatization and definitions of terms which will be included in the cryoMIxS paper.

ENVO releases of interest:

Ecotone, Polar express, Hot tub time machine.

post compositional data annotation model

//Maybe this could go in material and methods but I'll argue this is a result in the sense of semantic research, a model for data annotation.

In this work we present a novel semantic data annotation model. Semantics have been used to represent data … //TODO FIND REFS. In this model data annotations are composed of terms from the OBO Foundry. Data annotations are written in The RDF turtle specification, and structured as nested owl classes. Annotating the data as owl classes ensures parity to the OBO ontologies. This enables us to perform sparql queries on the annotated data in the same manor as would be done to query OBO Foundry ontologies.

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

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

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

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

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

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

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

Example of Post Compositional Data Annotation with Ontology Terms

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

Vocamp Virtual Glacial Hackathon

vocamp:

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

Virtual-Hackahon-on-Glacier-topic

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

AWI DBPEDIA contributions

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

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

[9] (citation for)

UNEP SDGIO

Despite operating within a semantically which is interoperable with the OBO Foundry the UNEP ontology is currently non queryable. Future work needs to be done to improve the way SDGIO purls are hosted via UNEP so that they can be querable. This would allow for the the incorporation of data mobililzed via semantics to the UN SDGs to help achieve their objectives.

Discussion

Querying Semantically Annotated Data

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

Creating Classes vs post compositional annotation for data annotation

VOCAMP Discussion section:

An example of the semantic clarification that took place during the "hackathon" was coming to a consensus definition of *ablation*. Ontologies take an agnostic stance when representing knowledge which has multiple definitions or which pertains to competing hypothesis. In the *ablation* example the NOAA National Weather Service Glossary 2009 [51] stipulates the restriction that only melting and evaporation processes contribute to ablation. The Cogley et al. IACS-UNESCO Glacier Mass Balance 2011 [52] definition however, refers to all processes which reduce the mass of a glacier. Specifically noting the inclusion of calving processes as significantly contributing to ablation processes.

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

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

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

consistent data structures for published data

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

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

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

Semantics as AWI Public Outreach

AWI Education & Communication

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

Conclusion

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

Outlook

add the stuff from the email to Anya.

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

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

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

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

References

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

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

3. **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.

4. **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.

5. **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.

6. **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).

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

8. **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).

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

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

11. **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.

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

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

14. **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.

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

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

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

18. **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.

19. **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.

20. **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.

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

22. **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.

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

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

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

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

27. **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.

28. **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.

29. **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.

30. **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.

31. **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.

32. **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.

33. **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.

34. **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.

35. **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.

36. **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.

37. **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.

38. **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.

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

40. **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.

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

42. **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).

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

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

45. **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).

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

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

48. **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.

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

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

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

52. **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).

Appendices

Python Scripts and Documentation

script 1 …

```python
#!/usr/bin/python

from SPARQLWrapper import SPARQLWrapper, JSON
import rdflib
import rdfextras
import rdflib.graph as g
rdfextras.registerplugins() # so we can Graph.query()
import sys
import re                   #to filter files using regex

########################################################################
# import from sys
#could later use this to accept a list of input classes to query
#in_args = sys.argv[1:]

in_arg = sys.argv[1]

########################################################################
# Put together a sparql query from pieces.

# Put together the PREFIX block:
def prefix_func():
    prefix_list = ['obo: <http://purl.obolibrary.org/obo/>', 'rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>', 'rdfs: <http://www.w3.org/2000/01/rdf-schema#>', 'owl: <http://www.w3.org/2002/07/owl#>', 'html: <http://tools.ietf.org/html/>']
    insert_prefix = ' \nPREFIX '
    return ('PREFIX ' + insert_prefix.join(prefix_list) + '\n' )

# Put together a select block
#takes the query and a bool for whether or not to add a distinct
def select_func(variables, distinct):
    insert_p = ' ?'
    if distinct == 1:
        return ('SELECT DISTINCT ?' + insert_p.join(variables) + '\n' )
    else:
        return ('SELECT ?' + insert_p.join(variables) + '\n')

```

```python
# Put together a WHERE clause which will only query for subclasses+ of a given input class
def where_subclass_query_func(input_class):
    return 'WHERE {' + '?s rdfs:subClassOf+ <' + str(input_class) + '> . } \n'

#put together a sparql query string which will query for subclasses of a
#given input class
def subclass_query_function(input_class):
    function_list = [prefix_func(), select_func(['s'], 1), where_subclass_query_func(input_class)]
    return ''.join(function_list)

#call the function to query for the subclass of the in_arg
query_for_subclasses = subclass_query_function(in_arg)

########################################################################
# wrap the ontobee SPARQL end-point
endpoint = SPARQLWrapper("http://sparql.hegroup.org/sparql/")
# set the query string
endpoint.setQuery(query_for_subclasses)
# select the return format (e.g. XML, JSON etc...)
endpoint.setReturnFormat(JSON)
# execute the query and convert into Python objects
# Note: The JSON returned by the SPARQL endpoint is converted to nested Python dictionaries, so additional parsing is not required.
results = endpoint.query().convert()

#save the results to a file with the name of the input purl
namestring = str(sys.argv[1])
#remove the http://purl.obolibrary.org/obo/ from the input file.
namestring = re.sub('http://purl.obolibrary.org/obo/', '', namestring)
namestring = re.sub('http://purl.unep.org/sdg/', '', namestring)
#make the file name to write to
outstring = 'subclasses_of_' + namestring + '.txt'

#write out to outfile: query_for_subclasses_of_out.txt
f = open(outstring, 'w')

#write each PURL fetched in query to outfile.
for res in results["results"]["bindings"] :
    f.write(res['s']['value'] + '\n')
```

script 2 …