CYVERSE™

Transforming Science Through Data-driven Discovery

MEOWL: Microbial Environments described using OWL

Sixth Annual Workshop of the Clinical and Translational Science Ontology Group

Oct. 25, 2017, Ann Arbor, MI (presented remotely)

Ramona Walls

University of Arizona

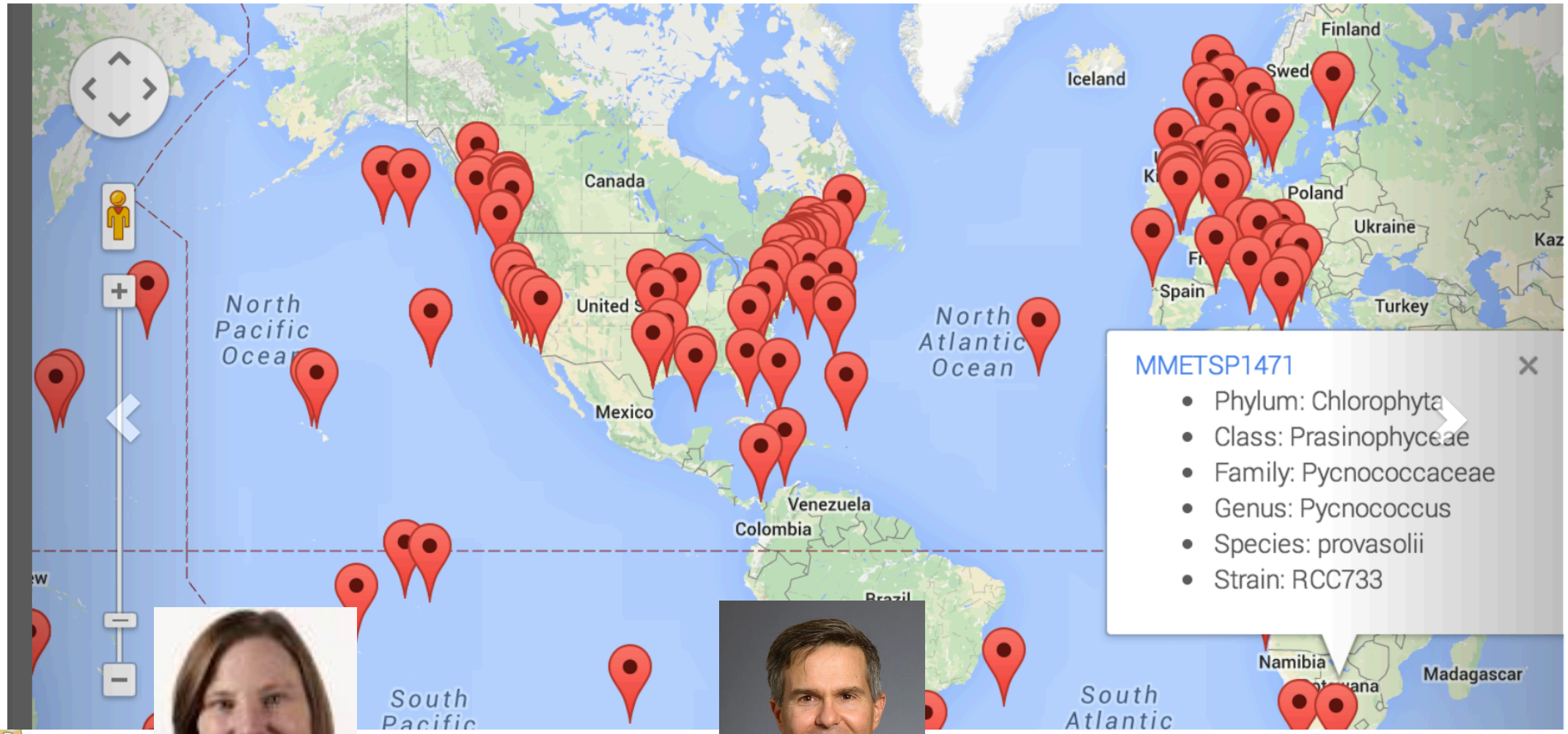
rwalls@cyverse.org [@RamonaWalls]

Outline

- iMicrobe: Search and discovery of world-wide microbial datasets
- MEOWL use cases and data mapping
- Ingesting data on microbial environments

https://imicrobe.us/

http://ivirus.us/



Bonnie Hurwitz

Ken Youens Clark

iMicrobe

- Aggregates global microbial datasets (metagenomic and genomic)
- Provides HPC tools for analyzing data via CyVerse
- Offers online data discovery portal

- Original datasets:
  - Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP)
  - CAMERA (was a repository for metagenomic data)

  - These datasets had environmental metadata (habitat, chemical parameters) but they were not standardized.

MEOWL

Goals:

- Build on existing ontologies:
  - Environment Ontology (ENVO)
  - Ontology for Biomedical Investigations (OBI)
  - Chemical Entities of Biological Interest (CheBI)
  - Biological Collections Ontology (BCO)
- Map to existing standards and vocabularies
  - MIxS
  - BCO-DMO
  - ENVO

https://github.com/hurwitzlab/imicrobe-lib

Cross-vocabulary mappings

	A	B	C	D	E	F	G	H	I	J
1	rdfs:label	MEOWL short label	BCO-DMO unit	other unit	From CAMERA data.xls	category	BCO-DMO short name	MIxS package	maps to MIxS term	other synonyms
48	carbon dioxide concentration	co2_conc		umol/kg	Carbon dioxide(CO2) - (umol/kg)	chemical		multi	dissolved carbon dioxide	
49	colored dissoloved organic ma	cdom	mg/m^3	RFU	CDOM - (RFU)	chemical	CDOM			
50	chlorophyll a concentration	chl_a_conc	mg/m^3|ug/	mg/l	Chla - (mg/1---L)	chemical	chl_a	multi	chlorophyll	
51	chlorinity	chlorinity		mM	Chlorinity(Cl) - (mM)	chemical				
52	chlorinity	chlorinity		uM	Chlorinity(Cl) - (uM)	chemical				
53	not useful without more information				-	chemical	x			
54	annual chlorophyll density	chl_dens_annual		ug/kg	chlorophyll density/annual - (ug/l	chemical	x	multi	chlorophyll	
55	annual chlorophyll density	chl_dens_annual		ug/l	chlorophyll density/annual - (ug/l	chemical	x	multi	chlorophyll	
56	chlorophyll density	chl_dens		psu	chlorophyll density - (psu)	chemical		multi	chlorophyll	
57	chlorophyll density per sample	chl_dens_samp_month		ug/kg	chlorophyll density/sample mont	chemical		multi	chlorophyll	
58	chlorophyll density	chl_dens		ug/kg	chlorophyll density - (ug/kg)	chemical		multi	chlorophyll	
59	unit missing				Chloropigment	chemical		mulit	chlorophyll	
60	chloropigment concentration	chlpig_tot_lt53		ug/l	Chloropigment	chemical	chlpig_tot_lt53			
61	chloropigment concentration	chloropig_conc		ng/g dry wt	Chloropigment	chemical	chloropig_conc			
62	chloropigment flux	chlpig_f		ug/m^2/day	Chloropigment	chemical	chlpig_f			
63	dissolved inorganic carbon co	dic_conc		mM	Dissolved Inorg C(DIC) - (mM)	chemical		multi	dissolved inorganic carbon	
64	dissolved inorganic carbon co	dic_conc		uM	Dissolved Inorg C(DIC) - (uM)	chemical		multi	dissolved inorganic carbon	
65	dissolved inorganic carbon co	dic_conc		umol/kg	dissolved inorganic carbon - (umc	chemical		multi	dissolved organic carbon	
66	dissolved inorganic nitrogen c	din_conc		umol/l	dissolved inorganic nitrogen - (v∞	chemical		water	total inorganic nitrogen	
67	dissolved inorganic phosphate	dop_conc		nmol/kg	dissolved inorganic phosphate - (r	chemical		water	dissolved inorganic phospho	
68	dissolved organic carbon conc	doc_conc		uM	dissolved organic carbon - (uM)	chemical		multi	dissolved organic carbon	
69	dissolved organic carbon conc	doc_conc		umol/kg	dissolved organic carbon - (umol/	chemical		multi	dissolved organic carbon	
70	dissolved organic nitrogen cor	don_conc		umol/kg	Dissolved Organic Nitrogen - (v∞	chemical		multi	dissolved organic nitrogen |	
71	dissolved organic nitrogen cor	don_conc		umol/kg	Dissolved Organic Nitrogen - (v∞	chemical		multi	dissolved organic nitrogen |	

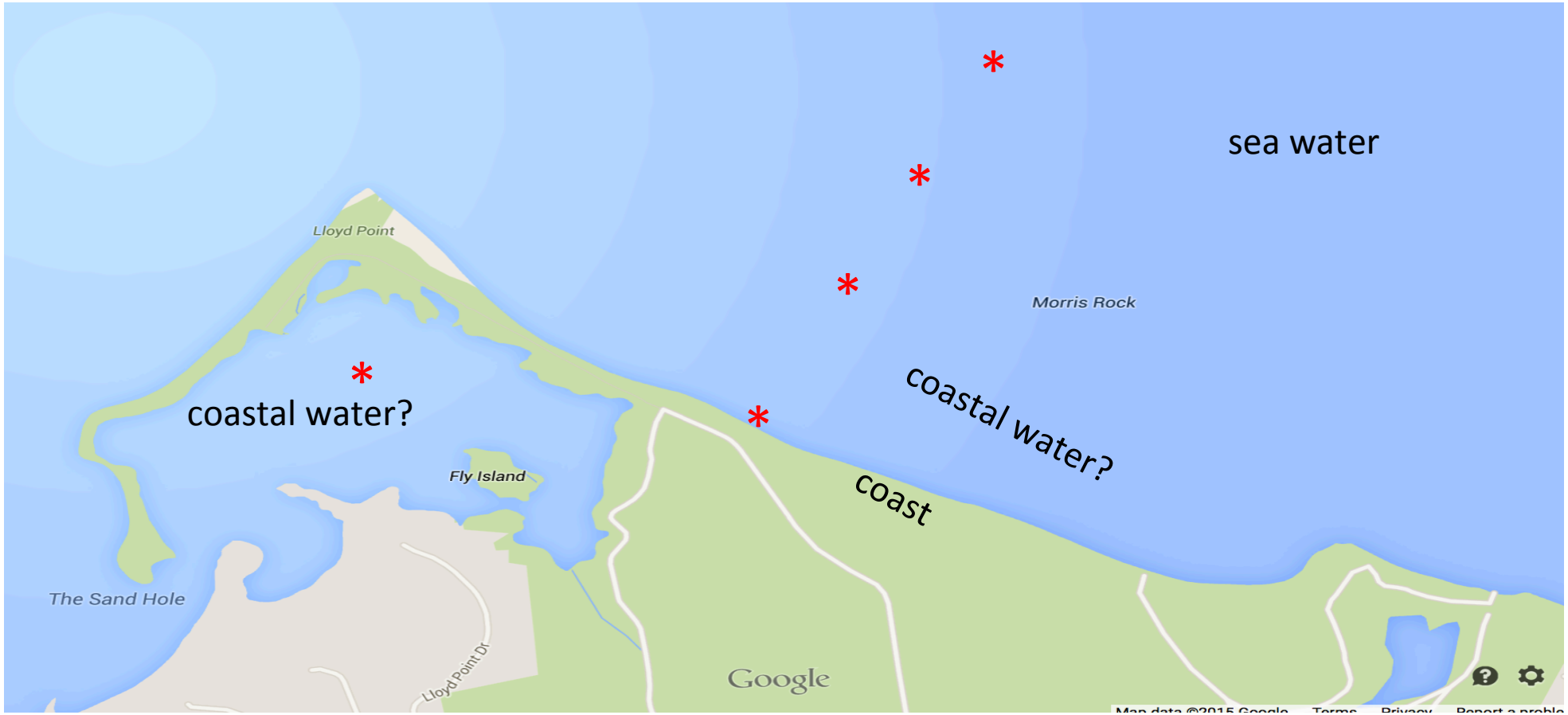
https://github.com/hurwitzlab/imicrobe-lib/blob/master/docs/mapping_files/iMicrobe_sample_metadata_master_list.txt

Data annotation with ENVO terms

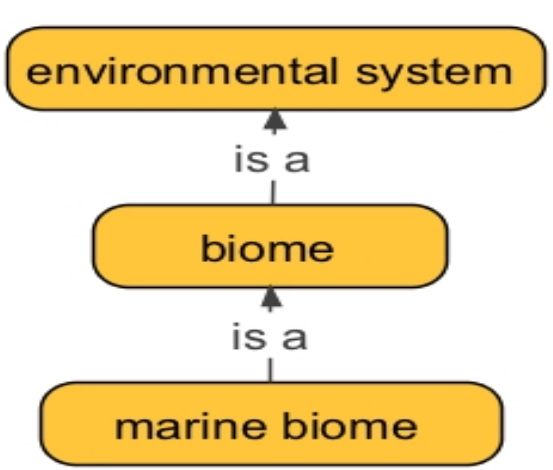
A	B	C	D	E	F	G	H	I
SAMPLE_ACC	SAMPLE_DESCRIPTION	DESCRIPTION	SITE_DESCRIPTION	REGION	HABITAT_NAME	biome_label	biome_id	environmental_mate
CAM_SMPL_000800		Freshwater	Surface waters of an Eutrophi	Lake 227, Experimenta	freshwater habitat	freshwater lake bi	ENVO:01000252	fresh water|surface v
CAM_SMPL_000800		Freshwater	Surface waters of an Eutrophi	Lake 227, Experimenta	freshwater habitat	freshwater lake bi	ENVO:01000252	fresh water|surface v
CAM_SMPL_000801		Freshwater	Surface waters of an Eutrophi	Lake 227, Experimenta	freshwater habitat	freshwater lake bi	ENVO:01000252	fresh water|surface v
CAM_SMPL_000803		Freshwater	Surface waters of an Eutrophi	Lake 227, Experimenta	freshwater habitat	freshwater lake bi	ENVO:01000252	fresh water|surface v
CAM_SMPL_000805		Freshwater	Surface waters of an Eutrophi	Lake 227, Experimenta	freshwater habitat	freshwater lake bi	ENVO:01000252	fresh water|surface v
CAM_SMPL_000807		Freshwater	Surface waters of an Eutrophi	Lake 227, Experimenta	freshwater habitat	freshwater lake bi	ENVO:01000252	fresh water|surface v
CAM_SMPL_000809		Freshwater	Surface waters of an Eutrophi	Lake 227, Experimenta	freshwater habitat	freshwater lake bi	ENVO:01000252	fresh water|surface v
CAM_SMPL_000811		Freshwater	Surface waters of an Eutrophi	Lake 227, Experimenta	freshwater habitat	freshwater lake bi	ENVO:01000252	fresh water|surface v
CAM_SMPL_000813	Station ALOHA; North P	seawater	oligotrophic ocean	Pacific Ocean	marine habitat	oceanic pelagic zo	ENVO:01000033	sea water
CAM_SMPL_000814	Lassen Volcanic Nationa	Sediment	sediment from an acidic flood	Boiling Springs Lake	sediment	freshwater lake bi	ENVO:01000252	lake sediment
CAM_SMPL_000814	Shane Seep; Coal Oil Po	Marine Sedir	Anaerobic subsurface sedimer	Pacific	sediment	marine benthic bio	ENVO:01000024	marine sediment
CAM_SMPL_000814	Placed near a Riftia pat	Diffuse Flow	Pacific: Gulf of California	Pacific: Gulf of Californ	saline water	marine benthic bio	ENVO:01000024	sea water
CAM_SMPL_000814	http://4dgeo.whoi.edu/	sediment	Marine methane seep	Eel River	sediment	marine benthic bio	ENVO:01000024	marine sediment
CAM_SMPL_000814		seawater	Temperate marine estuary	Estuary	marine habitat	estuarine biome	ENVO:01000020	estuarine water
CAM_SMPL_000814	Santa Monica Basin, off	Marine sedir	Carbonate mound formed by l	Pacific	sediment	marine neritic ben	ENVO:01000025	marine sediment
CAM_SMPL_000814	Mesocosm experiment	PFGE bands	Costal atlantic seawater	Raunefjorden	saline water	neritic pelagic zon	ENVO:01000032	coastal water
CAM_SMPL_000814	Placed near a Riftia pat	Diffuse Flow	Placed near a Riftia patch nea	Pacific: Gulf of Californ	saline water	marine benthic bio	ENVO:01000024	sea water
CAM_SMPL_000814	ice flow 50 miles NW of	Sea ice			marine habitat	marine biome	ENVO:00000447	ice
CAM_SMPL_000814		Saline water	uncharacterized, within perma	Arctic: Permafrost Sha	saline water	terrestrial biome	ENVO:00000446	saline water
CAM_SMPL_000814	Sea ice: Barrow, Alaska	Saline snow	polar coastal sea; saline snow	Arctic:Chukchi Sea	saline water	marine biome	ENVO:00000447	saline water|ice
CAM_SMPL_000815		seawater	1000 m deep in Atlantic Ocear	near BATS	saline water	oceanic mesopela	ENVO:01000036	sea water
CAM_SMPL_000816		Freshwater	Surface waters of an Oligotrop	Lake 239, Experimenta	freshwater habitat	freshwater lake bi	ENVO:01000252	fresh water|surface v
CAM_SMPL_000818		Freshwater	Surface waters of an Oligotrop	Lake 239, Experimenta	freshwater habitat	freshwater lake bi	ENVO:01000252	fresh water|surface v

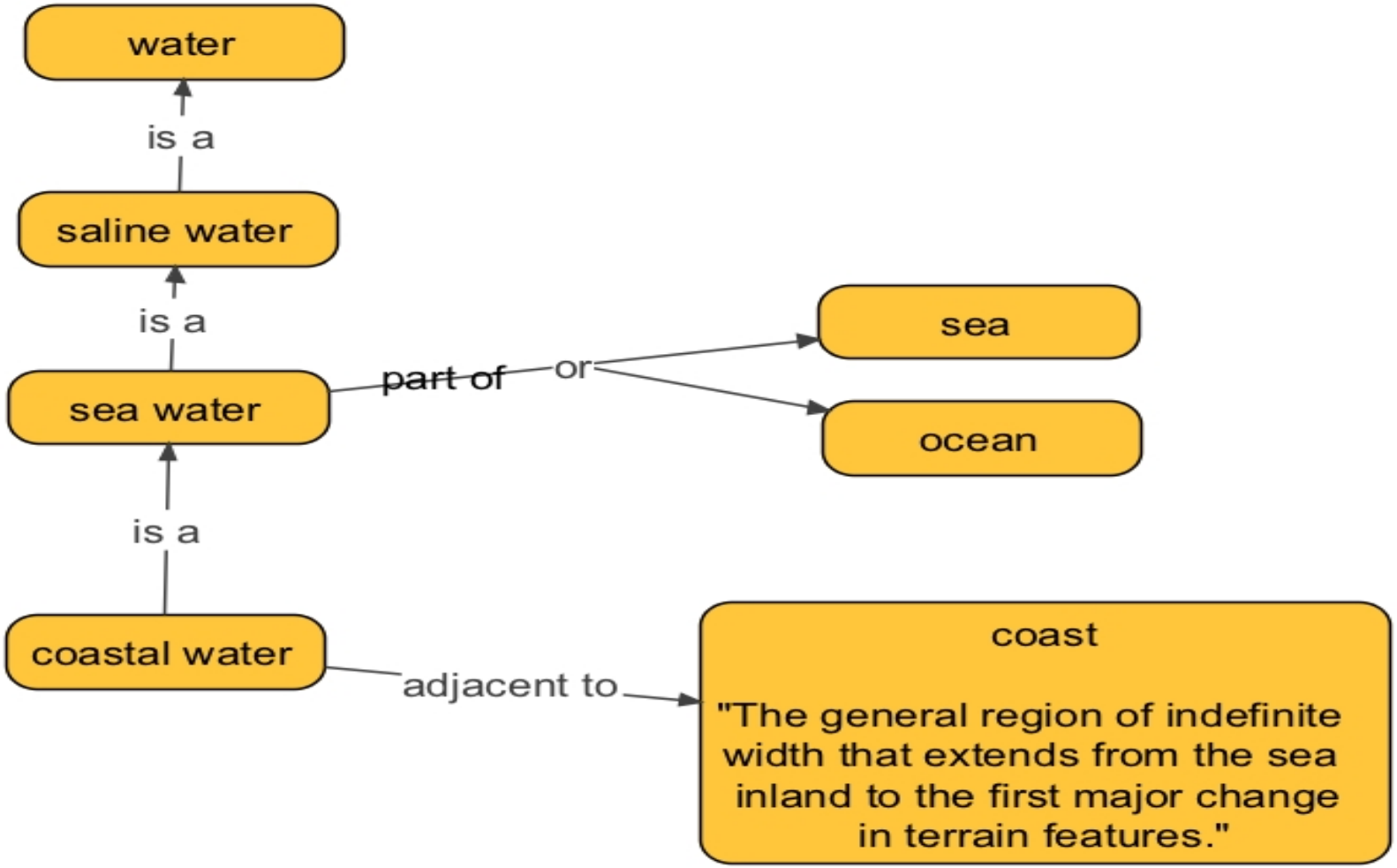
https://github.com/hurwitzlab/imicrobe-lib/blob/master/docs/mapping_files/CameraMetadata_ENVO_working_copy.csv

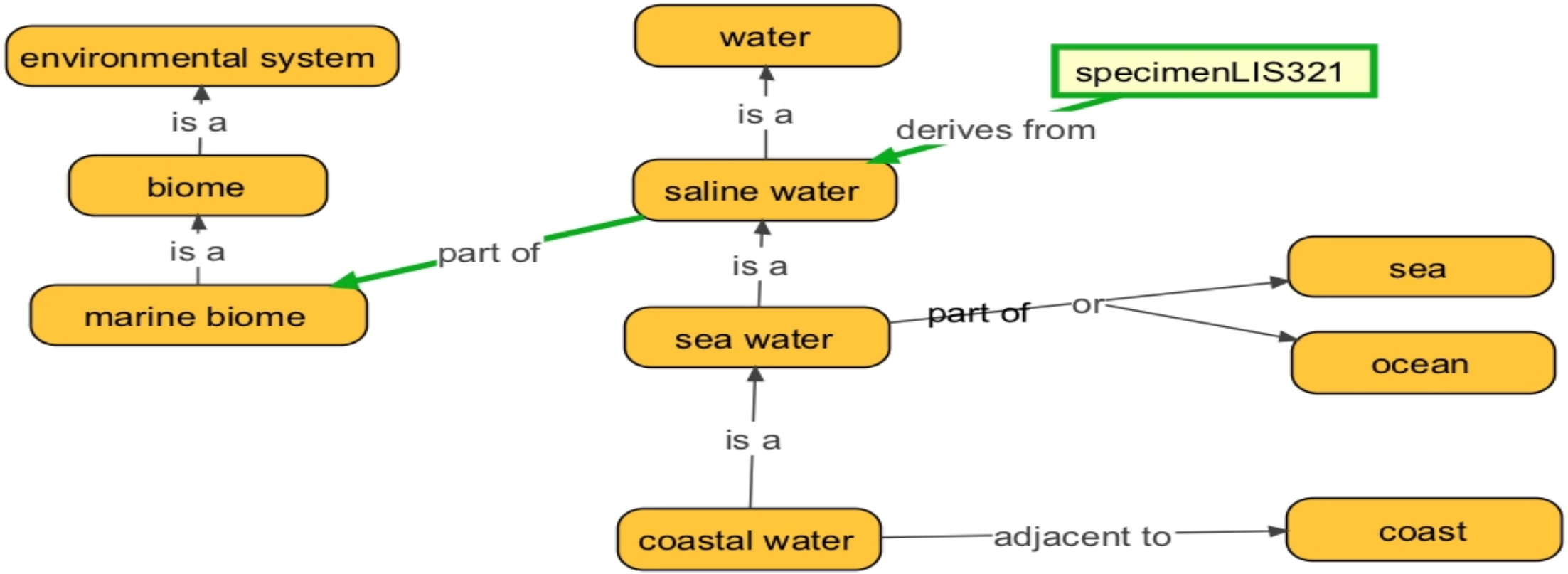
Use case 1: find all samples from coastal water

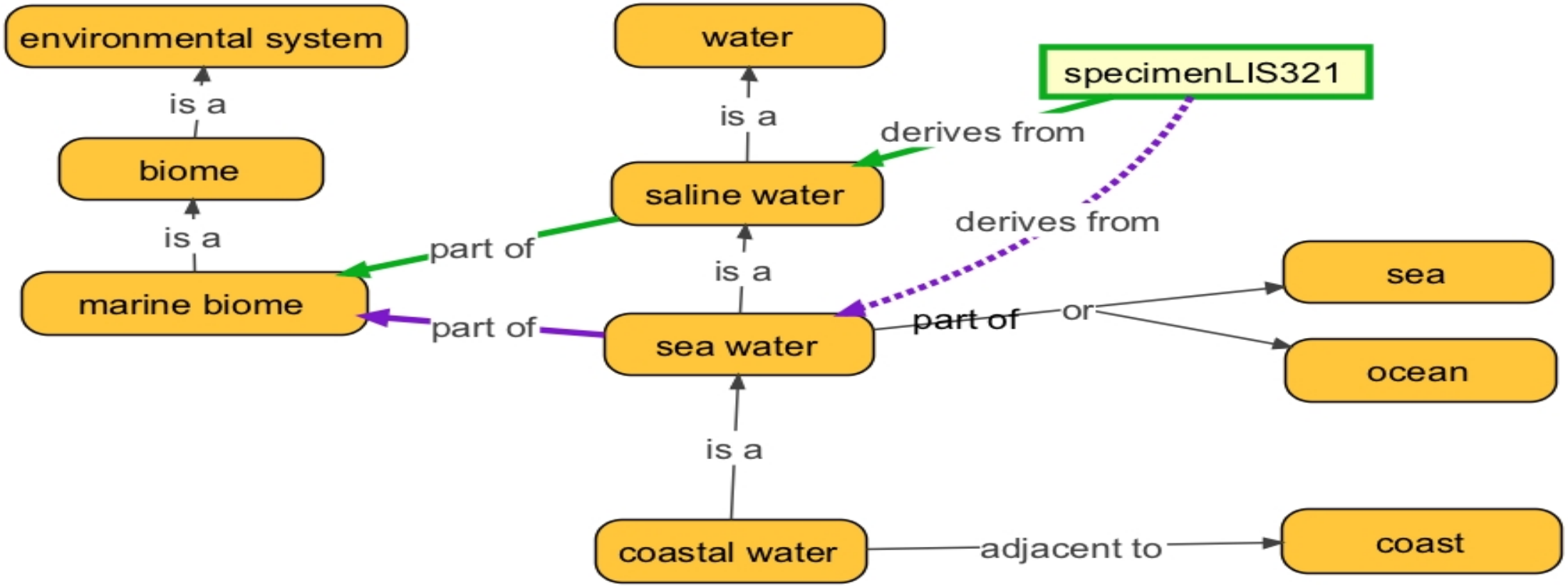


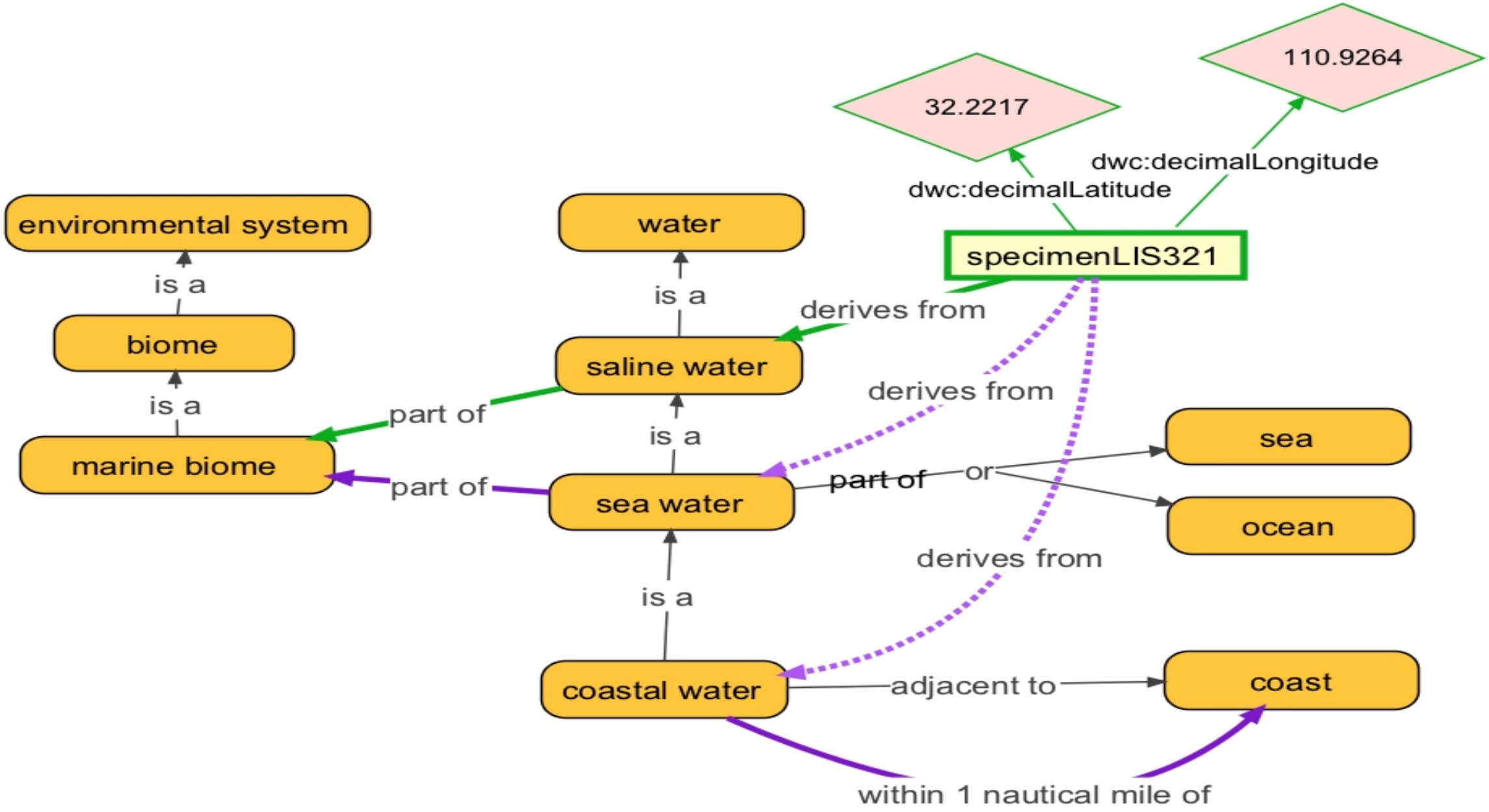
* specimen collection point


environmental system
is a
biome
is a
marine biome


water
is a
saline water
is a
sea water
part of
or
sea
ocean
is a
coastal water
adjacent to
coast
"The general region of indefinite width that extends from the sea inland to the first major change in terrain features."


environmental system
is a
biome
is a
marine biome
part of
water
is a
specimenLIS321
derives from
saline water
is a
sea water
part of
or
sea
ocean
is a
coastal water
adjacent to
coast

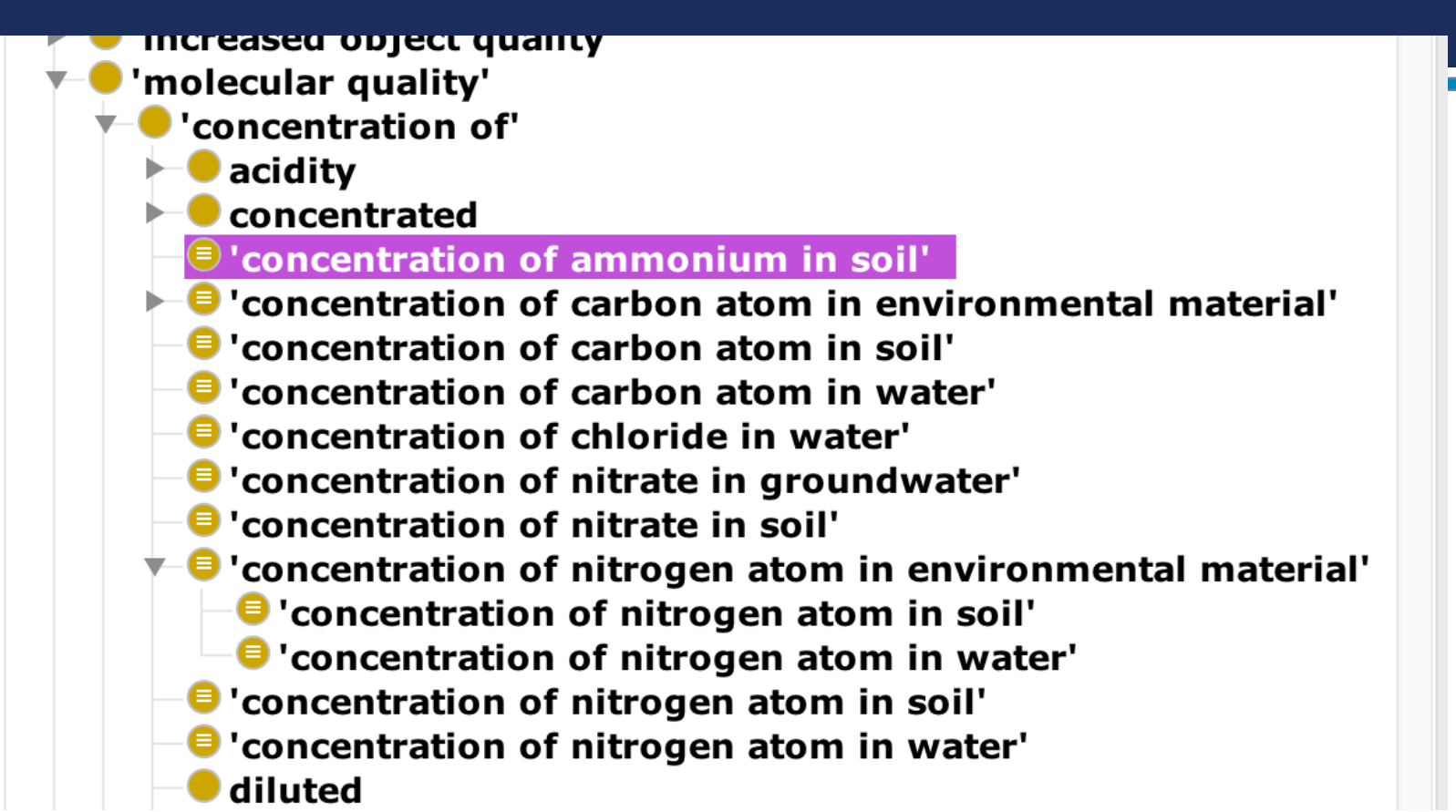

environmental system
is a
biome
is a
marine biome
water
is a
saline water
part of
is a
sea water
part of
is a
coastal water
specimenLIS321
derives from
derives from
part of
or
sea
ocean
adjacent to
coast


110.9264
32.2217
dwc:decimalLongitude
dwc:decimalLatitude
specimenLIS321
derives from
water
is a
environmental system
is a
biome
is a
saline water
derives from
part of
is a
marine biome
part of
sea water
part of
or
sea
ocean
derives from
is a
coastal water
adjacent to
coast
within 1 nautical mile of

Use case 2: find data associated with high nitrogen environments

Steps:

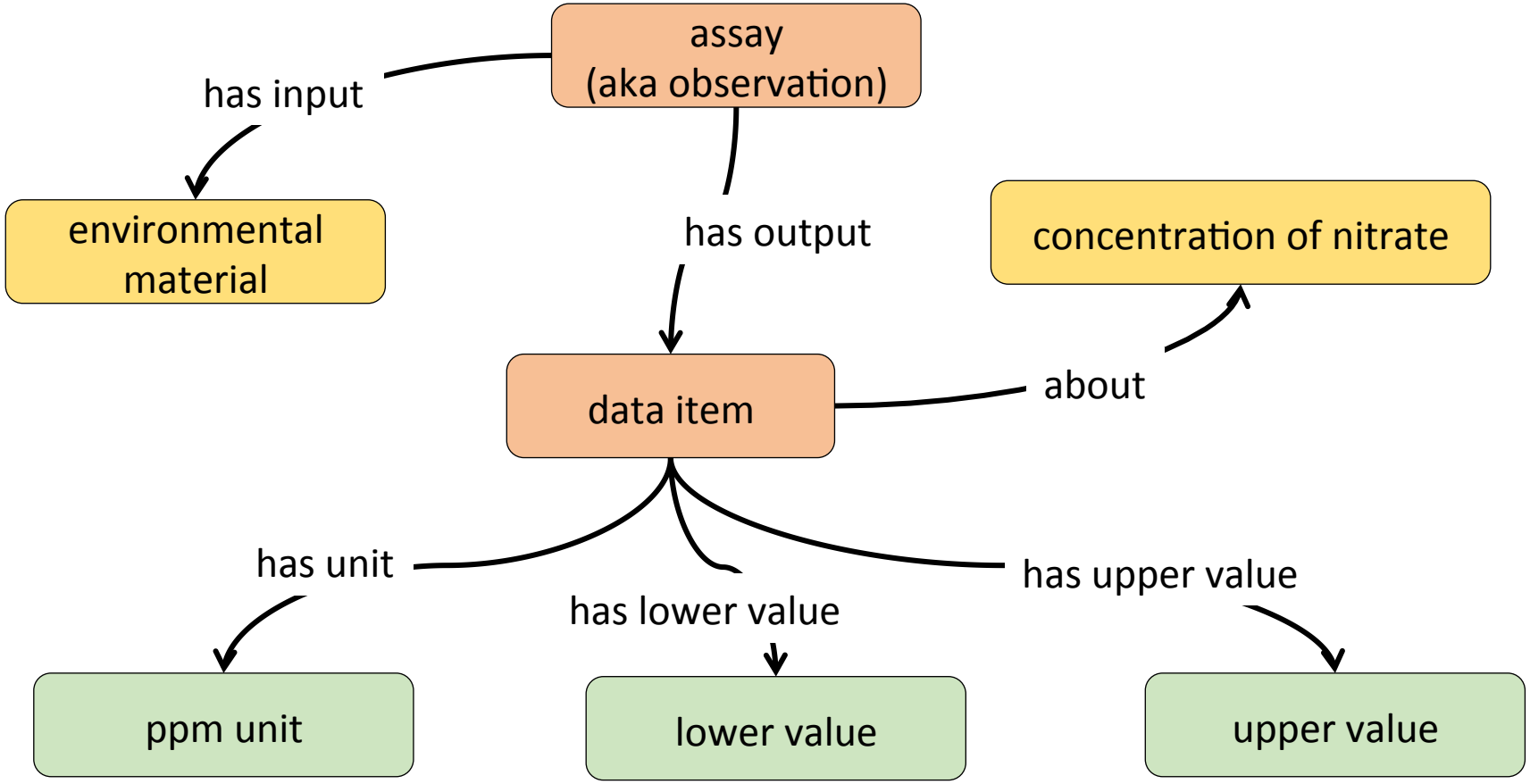
- Specify nitrogen containing molecules in ChEBI
- Specify environmental materials in ENVO
- Define classes for concentrations of chemical in environmental materials
  - Need to set boundaries here
  - Good use for post-composition/anonymous classes
- Define data limitations for “high nitrogen environment”



'concentration of'
and ('inheres in' some
(ammonium
and ('part of' some soil)))

How then to define a high nitrogen environment?

- Ingest the data as is, let the user decide.
- e.g.:
  - Find all sequences associated with samples where the total soil N was greater than Y
  - Find lists of species from sites where nitrate concentration is between X and Y



Obs. ID	Type	Variable	Value	Unit	Date	Lat	Long
CAM354-1	water	nitrate	2	ppm	20160923	25.2456	98.6892
mtep-873	soil	nitrate	0-10	ppm	20121108	12.4134	92.2754

Conclusions

http://freedesignfile.com/

- Data clean up trumps ontologies, for legacy data.
- Don’t try to force parameter categories (high, low, near, far, etc.) into ENVO or other reference ontologies.
- Use data driven queries to flexibly handle searches.