(Notes taken from Advanced Semantic Technologies class - April 30, 2013)

last week's (Apr 23) notes page:
https://docs.google.com/document/d/13FdBoqp1RYW19eH9bP2TsUbJwlXoCo-RBJyzVaux9Ig/edit
next week's (May 07) notes page:
https://docs.google.com/document/d/1VafTjB6_-_RsioTTU3RyXpBFOskYMVCmS--SwMF4zCM/edit

Attendees:
Katie
Zach
Brendan (a bit late)
Patrice
Deborah

Regrets:
Justin
Project Due Next Week
Web page of project

Katie Chastain - project update
http://tw.rpi.edu/web/SemantEco-Health-Facet

focus on symptoms and disorders
atsdr- agency for toxic substances and disease registry - she is filtering based on this.

Contaminant induces some Toxicity
Toxicity hasEndpoint some System
organ system to toxicity to contaminant
progress a bit lacking this week
question - is the ontology consistent? reasoner has been run and it was consistent previously - need to recheck

consider modularizing ontology to get the ASTDR information (this is the base level of information)

makes sense to have base ontology, then have other augmentation modules
take out:

- organ class
- health effects

organ partonomic relationships - from wikipedia....

organ class
Todo:

- finish flowchart
- have a paragraph and some protoSPARQL done by next week
- put latest version of ontology into github and patrice will load into an endpoint
- give general picture of ontology
    - What would you do tomorrow if I had time
    - Walk through an example (2 examples)
- write up an explanation about how SemantEco::Characteristics relates to your ontology
- once in endpoint, write SPARQL query's in object language to test that you can hit the endpoint

Could be subclassing chemical characteristics. Can infer a contaminant from a chemical. Don't need to define specifically.

Varying units seen for toxicity in ASTDR(sp?). Does this need to be considered? Patrice says probably not.

Missing one level of direction: need to update ontology.

Graph -> Put into Github!

Current Status of SemantEco: Can it be run locally? Not currently. No maven issues(I think).

Brendan has uploaded his batch to Github (SemantGeo)

Zach Fry - project update
http://tw.rpi.edu/web/EntityDisambiguation

freebase annotation focus
- keyword goes into freebase,
- given a keyword, search freebase, freebase returns a single type-ish term. patrice's service uses that to choose an ontology that is in the ncbo portal and then returns
a virtual ontology ID.
that is then used to use the ncbo annotator service

We are grilling Patrice:
Take string, match to freebase. We match on concepts (don't always exist). If there is a match, returns gax ontology that lines up with the match. Freebase doesn't have an ontology, just userbase tags that they use. You only get one or no response. Sometimes no concept is matched so you are left with nothing back.

E.g. Crystal Lake -> Crystal Lake or a lake matched to a geospatial identifier (“patrices”). ??
He has a simple ontology connecting Freebase categories into a specific ontology in GAZ.

If you submit a term to the annotator service but don’t specify an ontology to match against, you get back matches (bad ones) to really random ontologies. Just weird results. Sometimes only matches the longest one versus shorter results. The matching algorithm is odd.

Type something if patrice says it is an important point! (TYPE TYPE!)

Used other properties to help disambiguate terms. BioPortal is claiming something is subclass related but it isn’t.

I am confused, but I think Zach understands what he needs to do with this issue.

Need to compare subClassOf of with locatedIn. Can help with the parsing.

Might be cleaner to use SNOMED. Zach is a little unsure how to work with SNOMED.

Can’t match longest string - Is a bug. Only happens when you don’t specify ontology? Param: longestStringMatch. Wants to make it false, but if also don’t specify an ontologyID then there is an ERROR.

Two passes: Patrice’s pass. Also, ‘this’ pass: match everything you can in GAZ, see what hits.

Take greatest numbers of hits or the ontology to match against. Good idea here. *writes it down*

Read my notes as well!

Talking about weighting descendants. A side by side chart showing how things have (or have not) improved when applying this weighting / ranking logic. Write about these suggested improvements.

Let’s talk about next steps. Draft for submission to workshop. A website would be nice: for everyone.

Brendan Ashby - project update
http://tw.rpi.edu/web/SemantEco-SemantGeo

In my writeup, link to my deployscript and other files in the SemantGeo github.

Searched for organisms in Bioportal to find ontologies containing that organism. NCBI organismal classification ontology seems to be very helpful for this. It also provides a taxonomic hierarchy, which is useful for doing “related” searches.

This is useful to get more information about the measurements of species. A lot of the measurements in the data are counts of organisms in the sample.

Useful for figuring out which ones are invasive?

Useful to know what sister species are doing in relation to one species - eg, Are they both growing? Is one growing while others are shrinking? etc.

Among the species you can’t find in NCBI, see if the water people have any information/database data/papers, etc. about those species. There’s a standard vocabulary somewhere.

The data you have can be converted into RDF as another source for SemantEco. Map headers for species names to NCBI taxonomy classes. Then this can be added as more species data. In transforming to RDF, you can implement an enhancement for the mapping: eg, x column-header maps to y class in z ontology. (Using Tim’s RDF converter tool).

How to make use of “adult” vs. “immature” for species - this is not something in the ontology right now.

What do you want to see when you search for one of these species in SemantEco? Sites where measurements are available? Trend plotting for one site (over time? vs. some other characteristic?)? This would, at the very least, be able to display the data via RDF, along with provenance, which is an added capability for them. Right now, they only have a bunch of Excel files.

USGS geospatial data (.nt files)
Got a bunch of data on NY water features. Currently looking for watershed information, and other indications that water features may be “related”. Browsing through the data to look for watershed info, but nothing found yet. Features aren’t clearly named (might not have them?).

Look for labels, and look for the lakes that are also included in the Darin Freshwater data.

The Polygon data will be helpful for a search facet (part of a thing, next to a thing, etc.)

Todo:
- Be on a call with Lake George people. What can you show them as a demo? Graphs of things over time?
- Perform some RDF conversions of Darin Freshwater data with the enhancements discussed.

Make sure you document this process so that someone else can replicate the process in the future.
- Go through the USGS data and see which (if any) lakes are also included in the Darin Freshwater data. Find these coordinates (soon).
- Draft emails

For Everyone:

Final thing due on the 10th of May

Present of draft of website in next class. Get feedback and then do a final edit due that Friday. Make these websites in the TWC website. How long? not 100 pages. Enough to show understanding. ~20 pages.

Can demo code or other tech via pages linked on the website. Put a lot of graphics for clarity. Have a static webpage, with demos and graphics and summaries and such. Then you can link to your paper as a DOC on the site. Since we are all targeting a workshop, may want to write in the format of a submission to a workshop?

Can have the website, along with the paper in the form of a DOC.

Show the value of semantics and provenance. What tech are we leveraging and what is the value of them? Why use them. Give a generalization of the value.

Zach is a draft, Katie and Brendan are not. We are a paper? Point to what you are using. Have a discussion part on what was hard, and what is left to do.

Show you have insight in the limitations of the tech we are using and the “next steps”

In writeup, link to anything in the Github

Overall guidance for writeups

1 - we need your writeups on a web page that is on the twc site (this is so they exist after you have graduated
2 - your writeup must include a description of the background ontology/ies you are using and must include at least one example of how the semantics is being leveraged to answer questions that would be difficult to do without the background ontology
3 - your writeup must include a description of the provenance you are keeping and also include at least one example of how you are leveraging that encoding of the provenance to do something that would be difficult to do without the provenance encoding. (note this may be answering questions about where things came from but we want to see interesting uses of the provenance - for example filtering based on provenance aspects
4 - your writeup needs to include discussions of the general value of semantics and provenance. this can leverage your examples but it needs to address what is generalizable.