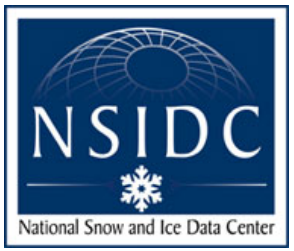
New NSF Project: Porting Biomedical Semantic Technologies to Earth, Ice & Life Sciences

Research team: Chris Jenkins (INSTAAR), Ruth Duerr (NSIDC), Martha Palmer (CLEAR), James Martin (CLEAR), Katja Schulz (SI)

Presenter: Chris Jenkins, CU/INSTAAR, CU Boulder

ESIP IT&I webinar 6 Nov 2014

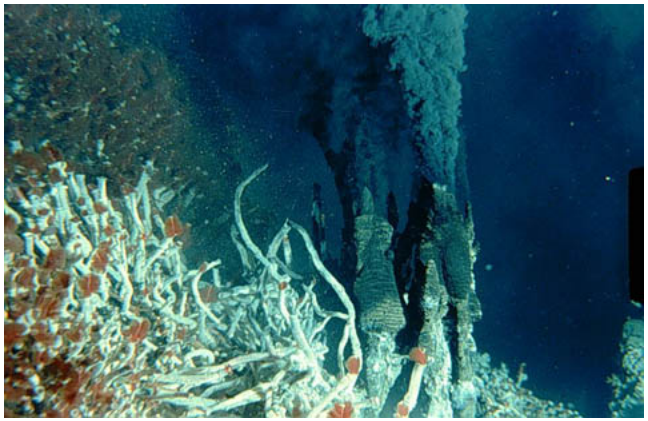


Motivations

- Earth sciences have put long effort into semantics, with small result. The process of building ontologies is too slow.
- But the requirement for smart query (query expansion), data crosswalks, ontologies, knowledge systems, and reasoners remains
- Need to use automated ontology-building (statistical and/or lexical)

Natural sciences

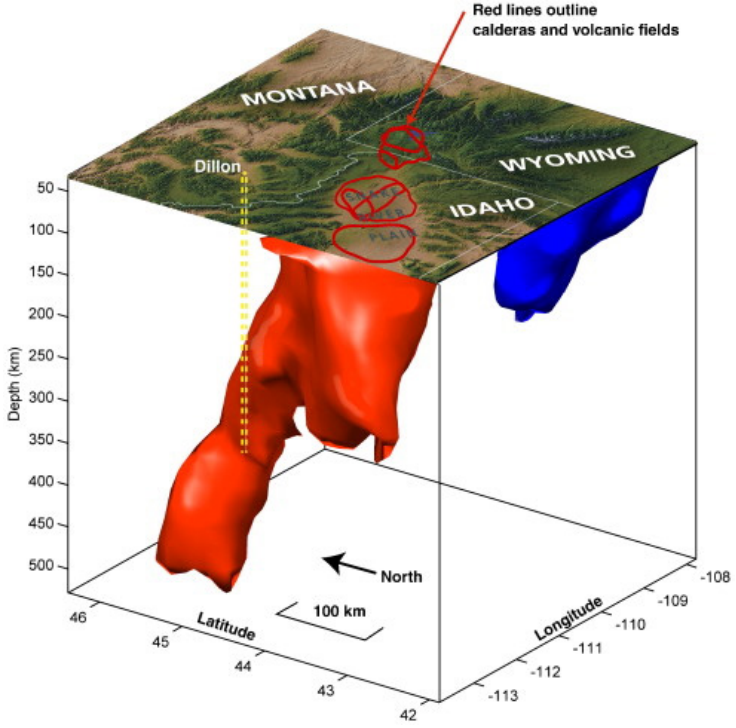
- Global seafloor (dbSEABED)
- Ice types and behaviour (SSIII)
- Organism traits (TraitBank)

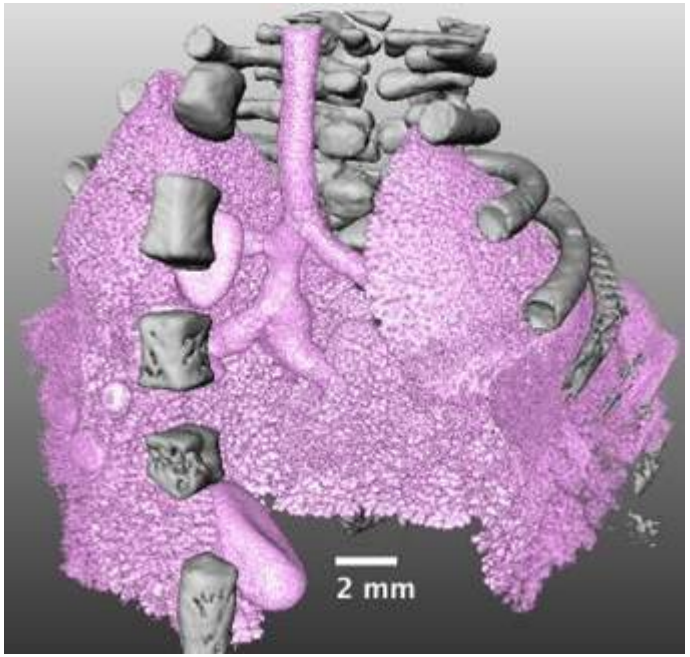


Why the Biomedical link ?

- Biomedical information has received huge research investment
- There are many similarities to the earth-ice-life science information requirements
  - organs, tissues – formations, reservoirs, with fabrics, shapes and locations
  - events – eg earthquakes, volcanic eruptions, ocean and climate events
  - diagnoses – hypothesis
- Collaboration: CU is advanced in biomedical semantic software through CLEAR TK Natural Language Processing pipeline.

Information
Structure
Similarities





- The UMLS (Bodenreider 2004) ontology for instance, includes many relations that are much sought after in the sciences but till now have been thought almost unachievable - relationships like events, processes, causality, geometric arrangement, dependency, result, parts within and hypothesis.

Computational Language and EducAtion Research

C L E A R

- CLEAR TK is a state-of-the-art combined NLP + ML system that also has associated essential tools for annotation, validation, document tagging, and event extraction.

- Is part of the framework for the document semantics system, Apache cTAKES (https://ctakes.apache.org), which used in over 60 institutions and networks such as the Mayo Clinic and Harvard Childrens' Hospital.

- Annotation, annotation assistant (active learning), validation
  - Automated (Machine Learning)

- Text normalization, sentence boundary detection, tokenization, part of speech tagging, stemming, entity identification, dependency parsing, coreference resolution, verb disambiguation, event recognition, semantic role labelling
  - NLP Pipeline

- Writing collected assertions to structured vocabularies / databases with quantitative truth assessments
  - (Developing)

Examples

POS tagging. In "Peat is a deposit consisting of decayed or partially decayed humified plant remains", 'peat' is marked as NNP (proper noun) and 'is' as VBZ (3rd person singular present verb).

Coreference resolution. In "Peat is a deposit consisting of decayed or partially decayed humified plant remains. It forms in wetland conditions.", 'peat' is grouped with 'it', which allows the system to infer that "peat … forms in wetland conditions"

Semantic role labeler. "Peat is a deposit consisting of decayed or partially decayed humified plant remains." Both nouns "peat" and "deposit" depend on the verb "is". Specifically, "peat" is the nominal subject (nsubj) of "is" and "deposit" is the attribute (attr). The past participles "humified" and "decayed" are treated as adjectival modifiers (amod) that depends on the nouns "plant" and "remains" respectively. Our SRL system identifies two states (i) "peat is deposit" and a "deposit consisting of …" and (ii) a process "plant is humified".

Event detection. In "Peat is a deposit consisting of decayed or partially decayed humified plant remains." our event and relation extraction models identify the process P1=<decay : agent=humified plant remains>, and the relations <is-a: peat, deposit> and <is-result-of: peat, P1>. These relations are then stored in text or database for checking and re-use.

Research Program

- Explore how to bring that technology over to the sciences, and perhaps also some of the semantic frameworks.
- Devise efficient, accurate methods of Human Annotations leading to, and multiplied by, Machine Learning – Natural Language Processing.
- Apply the technology to textbooks, glossaries and open text to mine factual assertions on a large scale.
- Build triple-stores and databases of the mined assertions across the cryo-, geo- and bio-Sciences.
- Sponsor development of applications to foster uptake of these semantics – Summer Hackathons

How to interface with the project

We are at early stages !!

- Do you have earth-ice-life research applications that require rich semantics ? Email us
- Do you have students who would like to write programs or apps for E-I-L semantics – apply for ‘hackathons’
- Do you have corpus texts that would be good to push through the pipeline to extract assertions ? - Email us

- End