ClearEarth: Preparing a Science Domain for NLP/ML, drawing on Biomedical Semantic Technologies

Chris Jenkins, Jim Martin, Martha Palmer, Ruth Duerr, Anne Thessen,
Skatje Myers, Jenette Preciado, Sarah Ramdeen

Smithsonian

Computational Language and EducAtion Research
C L E A R

This project is funded by NSF - Award ACI 1443085.

For the NSF DIBBS Meeting, 11 January 2016

An NSF-Funded project at INSTAAR and CLEAR at the University of Colorado Boulder, the Smithsonian Institute, and Ronin Institute for Independent Scholarship

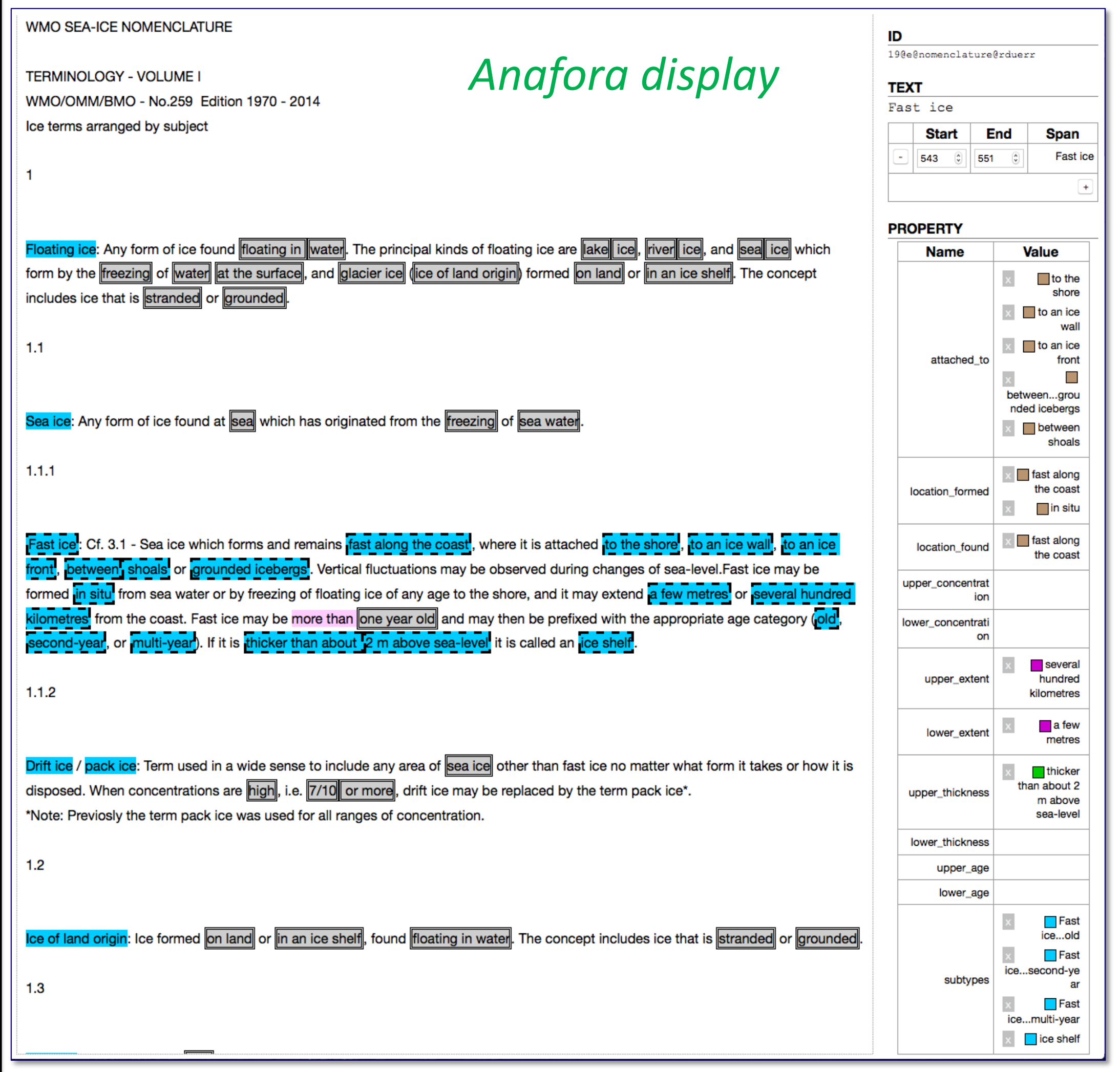
What is ClearEarth ?

C1F21 DIBBS: Porting Practical Natural Language Processing (NLP) and Machine Learning (ML) Semantics from Biomedicine to the Earth, Ice and Life Sciences

- To develop new semantic software tools and resources for the earth science fields of geology, biology and cryology - specifically earthquakes, ecology, sea-ice.
- Achieve this with high efficiency and effectiveness by porting resources, tools and methods from the biomedical field.
- Focus on Machine Learning (ML) and Natural Language Processing (NLP) and their data inputs and outputs.
- Develop a well-organized process for extending the products to other earth-science domains covering materials, processes, organisms, structures. To disseminate the products and methods in a variety of steps including conferences, papers, hackathons

Project progress:

- A more systematic approach to the major hurdle of seeding ML/NLP processes with annotated texts. The methods are now documented, schematized, the software made more robust, and publicized. (i) Semantic schemas which are not academic and abstract, but pragmatic – to work for annotators. (ii) The first set of annotation guidelines has been produced: "Sea Ice based on Reference Ontologies" now available at "http://bit.ly/2eaSVrG"
- Adoption of a pre-annotation step which (at this early stage) automatically marks up named entity terms with appropriate simple types and properties. This speeds up the annotation. Annotators will validate the pre-processed markup as they add further, deeper annotations.
- The results are now achieving sufficiently high scores on adjudication. We have completed double annotation and adjudication of Semantic Role Labels for 41300 instances in sea-ice. This is verification that the annotation faithfully captures the content of "gold standard" ontologies.
- The rest of the CLEAR TK pipeline will be adapted to produce extracted information (IE) databases and ontologies.



Anafora display

WMO SEA-ICE NOMENCLATURE

TERMINOLOGY - VOLUME I
WMO/OMM/BMO - No.259 Edition 1970 - 2014
Ice terms arranged by subject

1

Floating ice: Any form of ice found floating in water. The principal kinds of floating ice are lake ice, river ice, and sea ice which form by the freezing of water at the surface, and glacier ice (ice of land origin) formed on land or in an ice shelf. The concept includes ice that is stranded or grounded.

1.1

Sea ice: Any form of ice found at sea which has originated from the freezing of sea water.

1.1.1

Fast ice: Cf. 3.1 - Sea ice which forms and remains fast along the coast, where it is attached to the shore, to an ice wall, to an ice front, between shoals or grounded icebergs. Vertical fluctuations may be observed during changes of sea-level.Fast ice may be formed in situ from sea water or by freezing of floating ice of any age to the shore, and it may extend a few metres or several hundred kilometres from the coast. Fast ice may be more than one year old and may then be prefixed with the appropriate age category (old, second-year, or multi-year). If it is thicker than about 2 m above sea-level it is called an ice shelf.

1.1.2

Drift ice / pack ice: Term used in a wide sense to include any area of sea ice other than fast ice no matter what form it takes or how it is disposed. When concentrations are high, i.e. 7/10 or more, drift ice may be replaced by the term pack ice*.
*Note: Previosly the term pack ice was used for all ranges of concentration.

1.2

Ice of land origin: Ice formed on land or in an ice shelf, found floating in water. The concept includes ice that is stranded or grounded.

1.3

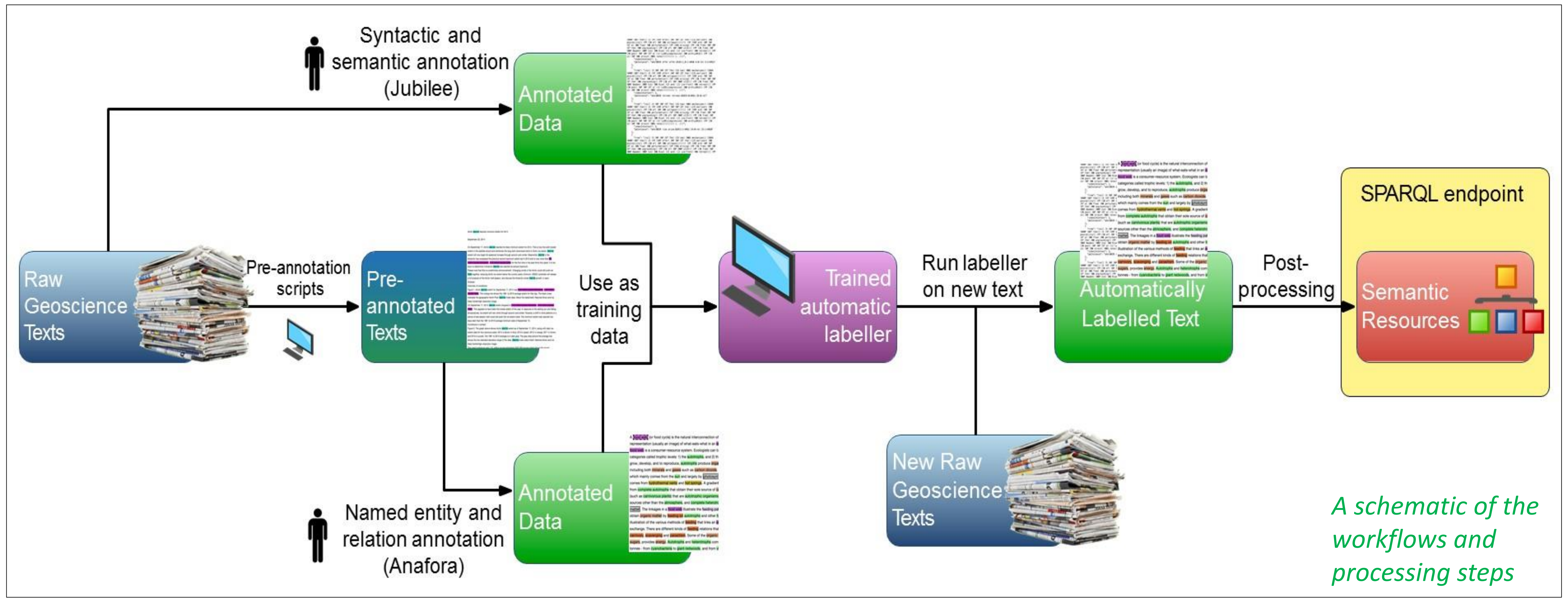
ID
19@e@nomenclature@rduerr

TEXT
Fast ice

	Start	End	Span
-	543	551	Fast ice

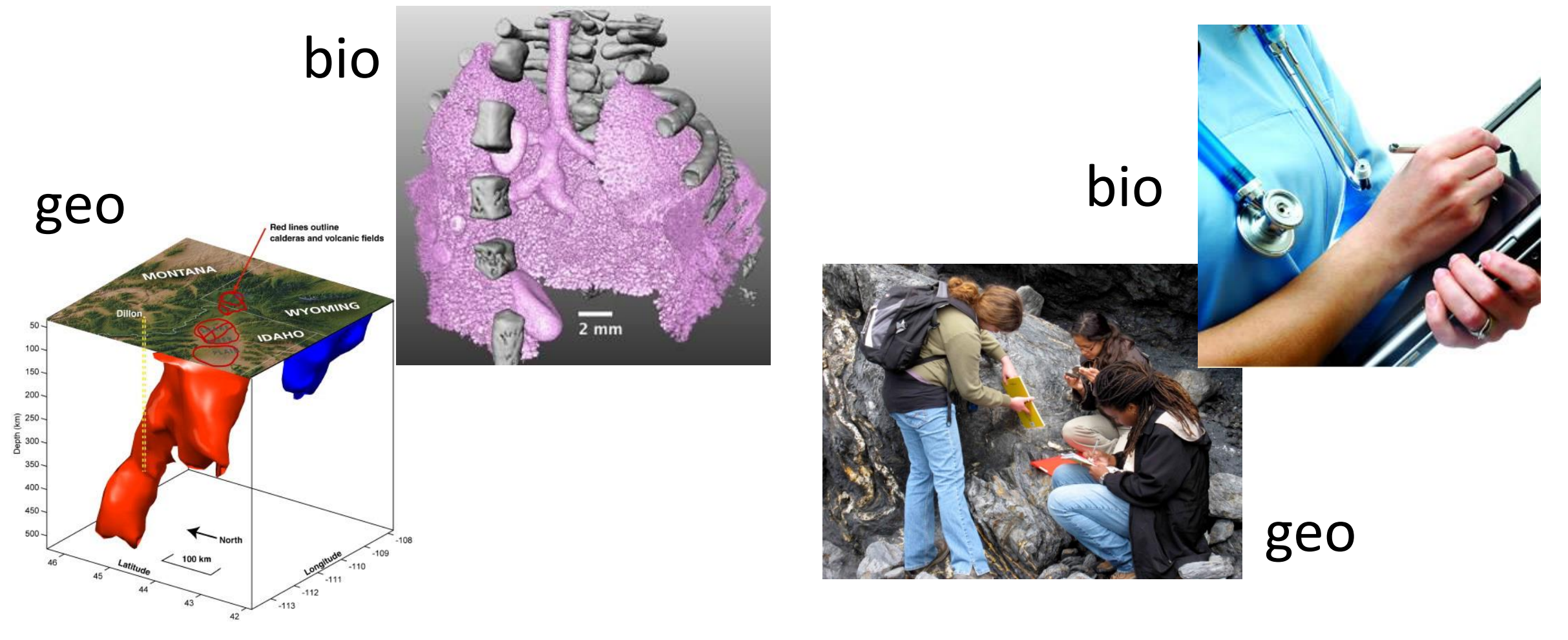
PROPERTY

Name	Value
attached_to	x to the shore x to an ice wall x to an ice front x between...grounded icebergs x between shoals
location_formed	x fast along the coast x in situ
location_found	x fast along the coast
upper_concentration	
lower_concentration	
upper_extent	x several hundred kilometres
lower_extent	x a few metres
upper_thickness	x thicker than about 2 m above sea-level
lower_thickness	
upper_age	
lower_age	
subtypes	x Fast ice...old x Fast ice...second-year x Fast ice...multi-year x ice shelf



A schematic of the workflows and processing steps

There are many similarities between the earth-ice-life science information requirements and those of biomedicine: organs and tissues (formations, reservoirs, with fabrics, shapes and locations); events (e.g., earthquakes, volcanic eruptions, ocean and climate events); data collection and diagnosis (e.g. fieldwork, hypothesis)



Anafora Annotation of Corpus:

- Software has been implemented to extract corpus-type texts from many sources. A highlight is the extraction from online biological databases such as the Encyclopedia of Life (EoL).
- For handling text inputs to annotation and analysis we now distinguish 'genres' - different types of text such as science paper, glossary, ontology, magazine article, news report, etc. The annotation difficulties and scores differ.
- The annotation (Anafora) stage is extremely important. The Machine Learning (ML) software requires sufficiently large annotated corpus. Once a threshold is reached, virtually all texts in the domain are treatable by the ML-NLP methods.

Named Entity Recognition system:

- Conversion of the Anafora-based annotation data to a token-by-token IOB-tag scheme that is amenable to training sequence-labeling systems.
- Development of a baseline set of features to extract from the data for use in training.
- Experiments with both support vector machine (SVM) and conditional random field (CRF) systems to gain initial performance data.

Semantic Role Labelling system:

- Development of a baseline set of features, including tree-based features, for using in building/training models.
- Experiments with SVM-based classifiers to get initial performance data.
- Experiments with re-training the Berkeley constituent-based parser with our data from the sea ice domain.

Lessons Learned:

- The project is giving insights on the best way to proceed for adapting existing ML/NLP methods and resources to new specialist domains. (i) Harness existing ontologies, glossaries and other semi-structured data for the training phases. (ii) Recognize the impact of different levels of text expertise – the Genres. (iii) Systematize the annotation process – provide a framework to the annotators. (iv) Training the SRL process with domain texts as opposed to general texts gives large increases in performance. (v) Monitor the performance metrics and achieve target levels at each step.

Example – SRL using Proposition Bank

- This well established process helps to quickly begin annotation of document sets. It employs tree-structured frames of the semantic roles. We ran an automatic syntax-based parser for the cryo texts. Manual annotators corrected egregious errors for the final automated annotation results. Using this approach, the annotators completed double annotation and adjudication of 413 tasks – an enhanced annotation rate.
- We also ran evaluations with a state-of-the-art automated SRL system. Systems that had been trained on general news (Ontonotes), were tested on the general news + the cryo data and just the cryo data. Best performance was by training solely on the cryo text -- 67.72 vs 60.45. When using hand-corrected syntactic trees, the results were 77.88 vs 68.39.
- Clearly, training the SRL on the domain texts rather than general text produces the best performance. This suggests that work should be focused on development of domain-specific semantic resources.