A Structured Vocabulary for Geomaterials - Technical Guide

Chris Jenkins,
INSTAAR, Univ. Colorado, Boulder USA

(chris.jenkins@colorado.edu)

(In roll-out stage March 2014)

Supporting projects:

(i) *NSF EAGER Collaborative Research: Bringing Together Computational and Linguistic Methods to Extract 'Dark' Geosciences Data for the EarthCube Framework (Award 1242909)*

(ii) *NSF RAPID: Seamless Marine-wetlands-coastal Soils Database to Support Urgent Decision-making Against the Deepwater Horizon Coastal Oiling (Award 1047776)*

(iii) *dbSEABED under many awards, with the goal of hugely increasing the pace of importing seabed data from text sources, but also tapping new data for zones of the ocean such as the coast and inshore where data is dominantly word-based.*

Introduction

Word-based data are pervasive in the geosciences. Parameters, materials, processes, events, physical arrangements and many other features are identified linguistically. A structured vocabulary is the key that opens word-based data to system logic and machine knowledge.

A comprehensive vocabulary of earth materials – geomaterials - is presented here. Geomaterials include sediments, rocks, soils, biogenic buildups, ice and snow, and man-moved and man-made landscape materials. The vocabulary is presented as a structured vocabulary, as a semantic net and associated components. Rather comprehensive information on concepts, relationships and certainties can be presented in this way. An ontology - a specialized subset of the structure - is computed from the primary products.

This vocabulary is being served to the community through the NSF Community Earth Surface Dynamics Modeling System (CSDMS) – an appropriate site for software, supporting data, community resources relevant to the earth's surface. A paper on the vocabulary is being finalized, where methods will be described in detail. Collaborations using the data and background software are welcomed.

Building the Vocabulary

The vocabulary is computed from a corpus of glossaries, dictionaries, thesauri, ontologies, classifications. It was necessary to compute it because of the great number of geomaterials terms now available – estimated to be 10^4. There could be ten times as many useful relations between them. Manual efforts to create a structured vocabulary through ontologies have encompassed only ~300 terms with rudimentary relationships (Geosciml 2012) in several years of work. By computing the vocabulary many times that number can be documented and related each time a new corpus is added. Interrelationships in text, and quantitative linguistic measures of concept distance and scope can also be computed.

The corpora used here were sourced from authoritative institutions such as US Geological Survey (USGS), Society for Sedimentary Geology, CSIRO Australia, US Federal Geographic Data Committee, Center for Deep Earth Exploration (CDEX) in Japan, British Geological Survey, US National Aeronautical and Space Agency (NASA), and the World Meteorological Organization (WMO).

The methods for extracting relationships follow those of lexical extraction Wordnet (Millar 1995), decomposition of names similar to CSDMS Standard Names ('http://csdms.colorado.edu/wiki/CSDMS_Standard_Names'), and distributional semantics. Since all sections of the incoming corpus had some misspellings and irregular formatting, considerable work was required on text cleaning. But given that, the resources are now available for this project and more sophisticated mining in the future.

The purpose of having separate litho and cryo runs (lithological and ice-surficial materials sub-corpora) as seen below was to examine processing results for different sets of input texts.

The tallies on components of the vocabulary are:

(i) Used 16 corpus documents. These had 836 lithological and 325 cryological-surficial materials nodes;

(ii) There were 2308 strong words of which 854 and 493 occur in the litho and cryo node descriptions, respectively. Clearly, the strongword collection is much bigger than the concept collections being analysed. There were 260 weak words from general English.

(iii) For litho 918, 1270, 1696 relationships from ontology, lexical and logical (i.e., the merged) methods respectively (not counting inverses). The relationships were: synonym, related, broader or narrower (i.e., skos:altLabel, skos:related, skos:broader, skos:narrower). For cryo there were 35, 264 and 269 onto, lexi and logi edges recognized.

(iv) In litho and cryo 151 and 46 undocumented (undefined) nodes from links directly in the ontologies of the corpus – an example of the problems with manually constructed ontologies.

(v) Concepts that had calibrated, very close linguistic relations, i.e., <0.2 relative entropy (the Kullback–Leibler divergence; Wikipedia 2014), were accepted as 'related' in 147 cases.

The significance of the detailed figures above is that: (i) computational methods do greatly extend the semantic network above manual methods; (ii) applying computer methods to the manually constructed ontologies reveals many shortcomings.

Naturally, the project now goes to a phase of reviewing the accuracies of the computed relations, which is the purpose of this data release, including the RDF-TTL ontology subset.

Components

In the zip file are two collections of data: “litho” and “cryo” for geological lithologies, and cryological ice and soil materials. In each is a set of files showing data and figures. Additionally, there is general system-wide data on the strong and weak word listings.

The streams of processing are identified throughout: ontological, for relationships *extracted* directly from resources; *lexical* for relationships inferred from text following Wordnet methods; *logical* for these combined; *statistical* for relations suggested by entropy measures of linguistic distance and generality. With the relationships in place concept *network rankings* were also measured using networkx (Python 2.7) methods (Hagberg et al. 2008).

i. A table of concepts (nodes) with their names, definitions, relationships, various metrics, and metadata. [*_anlzdNodeInfo.txt]

ii. Tables of ‘strong words’ and ‘weak words’ (a ‘stop list’) that are used to describe the concepts. The strong words are accompanied by frequency metrics and the sets of other strong words which they associate with. Strongwords are those that occur in the names of materials concepts and are not in the weak-words lists. They may refer to materials, minerals, parameters, processes, structures, properties. [all_SWordInfo.txt; _weakWords.txt; _wordnetstopWords.txt]

iii. Adjacency matrices for ontological, lexical, combined (‘logical’) and statistical relationships [ontoMatrixA.asc, lexiMatrixA.asc, logiMatrixA.asc, statMatrixA.asc; with corresponding PNG images]

iv. (TBA) A semantic net of subsumption relations, and also quantitative strengths on the links between them

v. A formal ontology of subsumption relations (i.e., synonym, related, broader, narrower) expressed using SKOS and RDF logic systems in TTL syntax. []

vi. Graphical nets written in GML (Graph Modeling Language), verified using networkx. [litho_onto.gml, litho_lexi.gml, litho_logi.gml, litho_stat.gml, litho_all.gml]

vii. This documentation. [GeomaterialsVocab_1.pdf]

Use case scenarios

The vocabulary components provide a large resource which are needed for downstream software applications such as query mediation, semantic crosswalk, disambiguation, databasing.

(i) A query could be launched using a set of terms (e.g., “feldspar-bearing sediments with glauconite”). The query is using local vocabulary and could alternatively we written “feldspathic sediments with verdine”. A ‘smart search’ (‘concept search’) drawing on a semantic net resource is able to search for both expressions – and also narrower ones such as “glauconitic albitic sands”. This is ‘query mediation’ and ‘query extension’.

(ii) Semantic crosswalks relate and compare two concepts. How close are they, do they subsume, what are their neighbours ? For example, “sandstone” and “arenite” are two close terms, but how close are they compared to others and do they have different associates ? The question arises for example when comparing map units from different states or nations as in One Geology (Jackson 2010).

(iii) Disambiguation is a similar concept: given a homonym like “plastic”, state of ductility versus hydrocarbon material can be distinguished by their typical word-associates in the text, with the patterns defined in a structured vocabulary like that served here. This is an essential operation in all operations described above.

To Do.

The vocabulary needs to be simplified for presentation. A format that would be very useful to develop would be that of ConceptNet v5.

All the products need to be subject to peer review and use. Only through use and application will they achieve validation and acceptance. The existence of the vocabularies needs to be publicized, for which there are detailed plans involving Earthcube and the CSDMS.

Acknowledgements.

I thank NSF for the funding support on these projects. I also thank D. Fils, T. Wever, R. Duerr, W. Ryan, J. Collier, K. Asch, K. Lehnert, S. Peckham and S. Richard for interesting discussions on geosciences semantics during the work.

Figures

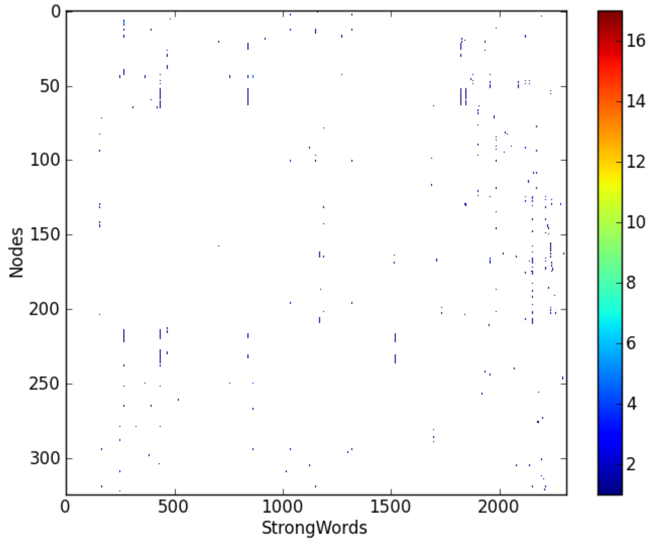

Fig. 1. Image of the distribution of strong words in the cryo sub-corpus of ice-snow and surficial landscape materials.

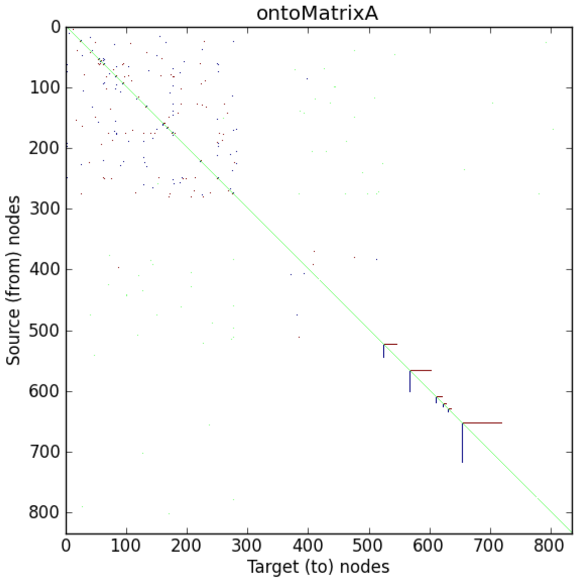

Fig. 2. Image of the ontological relations between pairs of concepts, litho sub-corpus. Red - broader, green – related, blue – narrower.
The density of relations is very sparse indeed. All relations have a nominal certainty of 100% from ontologies.

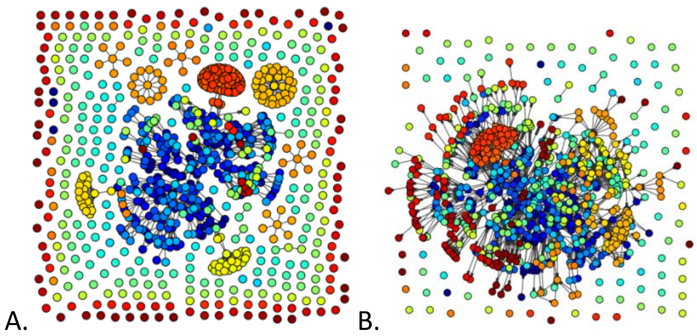

Fig. 3. Simple plots of the network connectivity of the concepts and relations at various stages of processing. A. Just the ontological relations (manually constructed). B. After the addition of lexically mined relationships – clearly improving the overall connectivity.

The ontology relations set seems to have been given special attention in particular sections.

Note the Orphan concepts, both plots. litho sub-corpus.

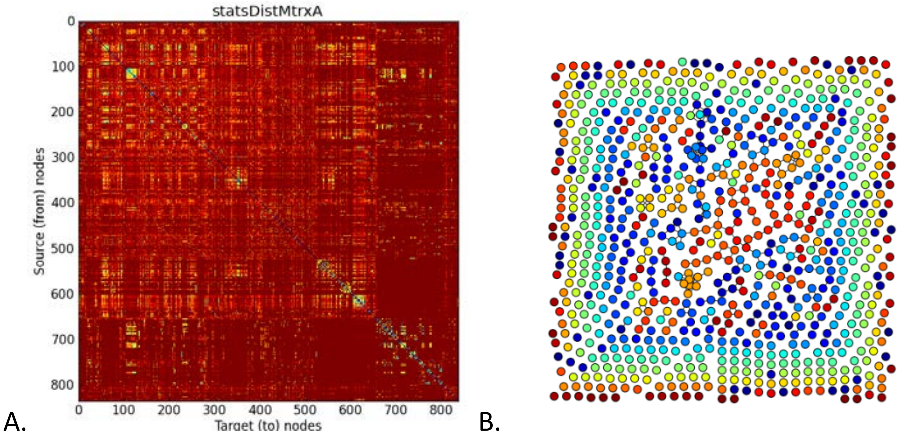

Fig. 4. Results of the search for statistically very close relations. A. The matrix of two-way statistical distances (relative entropies). Only a very small set of these were taken as 'related' assertions for the final network. Red, high distance. B. The network connections are sparse, but when combined with the onto and lexi relations form a comprehensive structure.

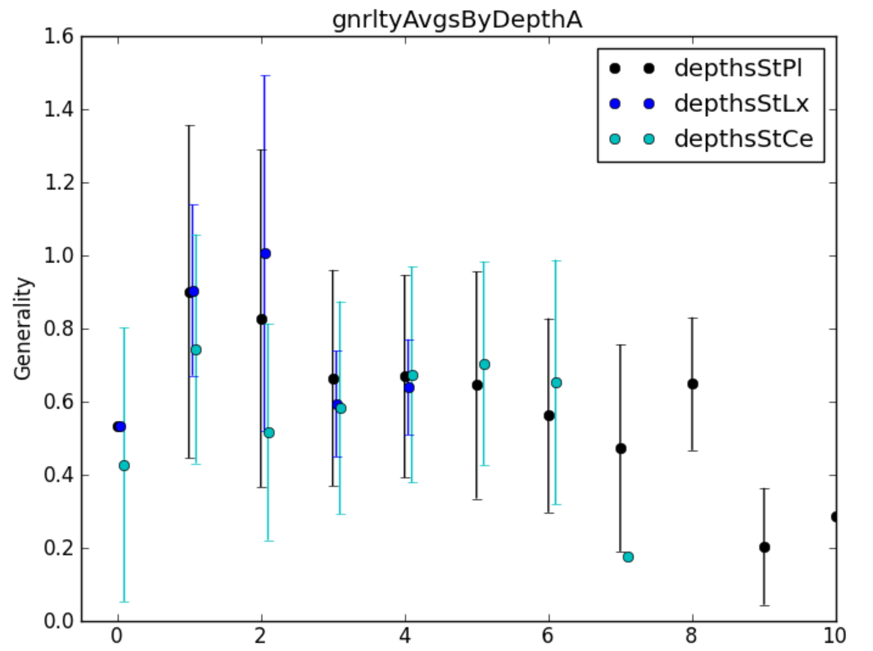

Fig. 5. Results for analysis of Generality metrics vs network rank ('depth', 0-10) on the 'logi' network. The metrics, which are entropy based, decline overall with increasing depth in the network away from node one. Despite the trend, generality may not be accurate enough to be applied corpus-wide. Concepts near the top (node one) of the structure are affected by tapered frequencies of use though the corpus.

I. Alkali Igneous Rocks

Name	ontoR	logiR	gnrlty	dijkstra
igneous_rock	2	2	0.48	1.6
volcanic_rock	-	-	-	-
basalt	-	4	0.93	2.8
alkali_basalt	4	-	-	-
sodic_basalt	5	5	-	-
rhyolite	-	-	0.85	3.7
comendite	-	9	0.05	5.4

II. Clastic Sediments

Name	ontoR	logiR	gnrlty	dijkstra
sedimentary_material	1	1	0.34	0.8
sediment	2	2	0.71	1.2
clastic_sediment	3	3	0.59	1.6
mud	4	4	0.88	2.1
clay	5	5	0.05	2.6

Fig 6. Example of type of calibration employed.

As a test of the local operation of the computed ranking and generality metrics, two expert colleagues were asked to supply a hierarchy of lithological names in order of ranking in classifications, one for alkali igneous rocks, the other for clastic sediments. The results above suggest that the network ranking methods and generality metrics do work on branches of the vocabulary, but not across branches.

The various measures are: 'ontoR', 'logiR' – ontology and logical lexical rankings by networkx; gnrlty – the entropy based generality measure; dijkstra – networkx ranking based on distance from node one, weighted by the relative entropy distances along each path.

Red values are out of order. Generality measures suffer from volatility in the numbers and choices of words, especially near the top – near node one.

Tables

A. Parameters for the Concepts Information

Parameter code	Parameter details
Index	Unique for concepts(nodes) of the run
UniqueCode	Unique code like "Label.[Subset].Corpus.Institution"
OriginalName	Concept name in original corpus
Definition	Text definition (if any) in original corpus (cleaned)
prefLabel	Concept Label as in 'UniqueCode'
altLabel	Alternative labels (i.e. synonyms)
related	Equivalent concepts, perhaps overlapping or different scope
broader	Subsuming concepts
narrower	Subsumed concepts
type	Flag to indicate class of concept, e.g. whether soil, rock, ice, or (for strongwords only) process, parameter, mineral.
source	Indication of the source of the name and/or definition; e.g., a paper citation.
comment	Comment
Lexical broader	Mined, lexically subsuming concepts
tidyText	Text which has been cleaned and tidied (e.g., "less than 33 percent" to "<33%"). Verbosities removed.
strongWord	The 'Bag of Words'
depth[lexi;stat;djkp]	A ranking of concepts from top node ('material') at zero, increasing numerically to the branched extremities of the network. Three measures are given: lexi - networkx path to top node, stat – linguistic (relative entropy) distance to top node; djkp – networkx Dijkstra Minimum Path Length to top node. The network distances were calculated on the Directed Acyclic Graph.
gnrlty[entrpy]	Entropy-based matric for the generality of the concept based on strongwords, compared to the total corpus.

B. Parameters for the Strong Words

Parameter code	Parameter details
strongWord	A word that occurs in any of the concept names of the corpus, but not in the stop-list. Candidates for the Bag of Words.
strongWordIndx	Numerical index
globalCount	Frequency of the word in the run (e.g., for litho)
posTags	Parts of Speech Tags from Python NLTK; not always unique. Used for stemming.
levenshtein(term:idx:editsCnt:editsPct)	Nearest Levenshtein morphs.
associates(term:idx:count)	Associated strong words through the run (e.g., for litho)
parts(..|pfxs<stem>sfxs|..)	A stemming performed on suffixes (combined NLTK methods) and prefixes. The results, as with all stemming, are not always as expected.

C. List of Files in Zip and Commentary

File Name	**Contents**
general	
i. all_SWordInfo.txt	i. The strong words collection, with morphs and metrics; used to make the 'BagOfWords' texts
ii. _weakWords.txt, _wordnetstopWords.txt	ii. Stop list files – from external sources
*** (folder – cryo or litho)**	
i. *_lithNet.ttl	i. TTL serialization of the proposed ontology
ii. *_all.gml	ii. GML encoding of the graph of the ontology
iii. *_anlzdNodeInfo.txt	iii. Listing of the concepts with all text and metric entries
iv.	iv. JSON listing of the nodes and edges as a multigraph with attribute data for each
where * is either cryo or litho	
figures	
i. termsInDocsA.png	i. Image of terms distributions through the corpus items (ie per glossary entry)
ii. statsDistMtrxA.png	ii. Computed pairwise entropy divergences between the concepts
iii. adjcncyMtrcsA.png	iii. Images of the onto, lexi, logi (i.e., onto & lexi merged) and stat (used only) relations
iv. onto_Graph.png	iv. The networkx simple graph of connectedness for onto
v. lexi_Graph.png	v. The networkx simple graph of

vi. logi_Graph.png vii. stat_Graph.png	connectedness for lexi vi. The networkx simple graph of connectedness for logi vii. The networkx simple graph of connectedness for stat
adjacencyMatrices	
i. pathLenMutualA.asc ii. logiMatrixA.asc iii. lexiMatrixA.asc iv. ontoMatrixA.asc v. statsDistMtrxA.asc vi. statMatrixA.asc	i. Shortest path lengths between all concepts in Logi graph (all lexi & onto relations) ii. Onto and Lexi adjavencies merged; conflicts resolved iii. Lexi relationships iv. Onto relationships v. Entropy measures of distance based on BagOfWords vi. The stat relations chosen to be represented in the final ontology
litho_HTM	
*.htm	A small HTM file for each concept, useable when a network diagram becomes click-able.

D. Formats of the Network Graph Products

JSON Multigraph	

{
 "directed": false,
 "graph": [
  [
   "project",
   "litho"
  ]
 ],
 "nodes": [
  {
   "code":
"material.keystone.lithnet",
   "id": 0,
   "rank": "0.0;0.0;0.0"
  },
 	
  {
   "source": 412,
   "relation": "lex:rel",
   "target": 412,
   "weight": 0.80000000000000004
  },
--- 	
	Is Graph Directed ?
	
	Project Name
	
	Start of NODES Listing…
	
	Node Code
	Node Id
	Node Ranks data
	
	Start of EDGES Listing…
	
	Source node Id
	Type of Edge Relationship
	Target Node Id
	Edge Weight
	
	------------------------ 	
	Note: Although this is an undirected multigraph serialization, in fact the edges have Src→Tgt directions.

References

Hagberg, A.A., Schult D.A. and Swart, P.J. 2008. Exploring network structure, dynamics, and function using NetworkX. In: G. Varoquaux, T. Vaught, and J. Millman (Eds), *Proceedings of the 7th Python in Science Conference (*SciPy2008*), (Pasadena, CA USA),* pp. 11–15.

Geosciml 2012. *CGI Simple Lithology Categories*. [URL: “http://resource.geosciml.org/static/vocabulary/cgi/201012/Vocab2011html/SimpleLithology201012.html”]

Jackson, I. 2010. Acting Locally, Thinking Globally: One Geology? *Eos, Transactions American Geophysical Union, 91(5),* p. 43.

Millar, G.A. 1995. *WordNet: A Lexical Database for English*. Communications of the ACM 38(11), 39-41.

Wikipedia, 2014. Kullback–Leibler divergence. [URL: “http://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence”]

INSTAAR 27 Mar 2014