# Machine-Assisted Buildout of a Structured Vocabulary, including Ontology, for Geomaterials

Chris Jenkins, INSTAAR, Univ. Colorado Boulder

***Abstract***

*Machine assisted construction of a structured vocabulary for geomaterials terms was recently carried out on a large scale for over 1400 terms from 20 corpus documents – glossaries, classifications, ontologies, catalogs. The processing steps are described, and some conclusions from the study are given.*

*The structured vocabulary is now available on a collaborative basis for studies to optimize semantic structures (e.g., ontologies) in the geosciences, and also to develop software applications for databases, such as semantic query and crosswalk.*

**Introduction**

- Semantics are central to geosciences information. They describe the materials, structures, processes, dispositions, events and causes - and all the parameters. Given this, an important question is: "How can existing geosciences vocabularies, which are large and complex, be placed in a systematic structure so they can be used computationally ?"
- For this , effective, efficient, and accurate ways are needed for bringing vocabularies into data systems, to a point where they have enough depth and scope to be useful in research. This means that must have non-trivial relations - beyond 'Geology 101' assertions and small controlled vocabularies. They should allow for homonym disambiguation, for attaching quantities to concepts, for measures of concept similarity and breadth, and for uncertainties.
- Sufficient research on suitable structures has already been conducted in other sciences, particularly biomedicine and the information sciences. Geosciences should adopt the findings from those fields. In this study we applied techniques from other sciences to a broad linguistic corpus on geomaterials, and extracted the beginnings of a structured vocabulary (including ontology) on the subject.

**Geomaterials**

- The vocabulary for describing geomaterials – the materials that form the earth's surface – is huge. The estimated number of distinct terms and usages is >10,000. And it is complex, perhaps because geomaterials have always been closely involved with humans' engineering and agriculture activities. Many sciences have their own distinct and entrenched terminologies for geomaterials: in geology , geotechnics, geomorphology, ecology, pedology, hydrology, oceanography.
- Examples of geomaterials terms include: granite, shale, mud, chernozem, peat bog, aleurite, lamprophyre, glacial moraine, sand dune, ice, methane hydrate, bitumen, coal, sulphide ore, seawater, coral reef. These are materials – usually bulk, composite materials – that compose earth-surface features. As we see, often the material and the feature ('geoform') are conflated in use.

**Motivations**

- The immediate aim of the study was to explore how this vocabulary, which is central in the geosciences, could be systematized using machine-assisted (i.e., software, not manual) techniques.
- For the longer term the goal was to help with data import and activation in the dbSEABED database (e.g., Goff & others 2008) of ocean-floor and coastal substrates. For example, seamless mapping of substrates from offshore through coastal to onshore requires semantic crosswalks between the marine, shoreline and agricultural vocabularies. To match the sophistication of the Fuzzy Set Theory methods already in dbSEABED (Jenkins 1997) semantic tools will need to be comprehensive and quantitative, hence the interest in this project in quantifying term relationships and actual numerical values for concepts.

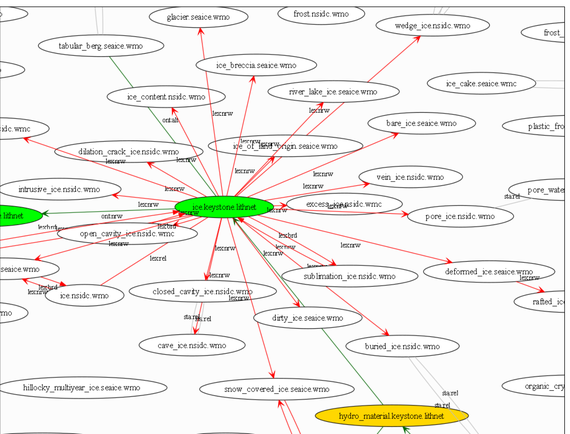



*Fig. 1. The geomaterials vocabulary is large, complex and disorganized. This makes it very difficult to show visually. Here we see just a part - for ice and snow that was built with WMO, NSIDC USGS, NASA corpora.*
*The green and orange nodes are 'keystone' nodes – essential for classification. Arrows for the network edges point to the narrower concept.*

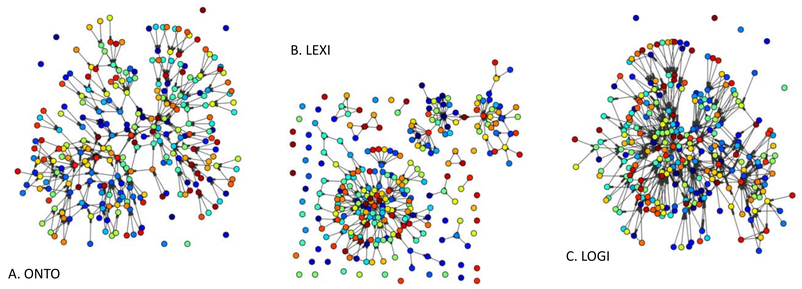



*Figure 2. Illustration of the contribution of the lexical mining procedures to a part of the geomaterials' semantic network. A. only extracted, existing ontology relations ('ONTO'); B. only mined lexical relations ('LEXI'); C. the combined result ('LOGI' – all logical relations).*
*The drawn relations include SKOS related, broader and narrower types, with arrows directed to the broader concept. Notice the prevalence of unconnected nodes, referred to in texts but not related. For clarity, these networks represent only a small part of the whole geomaterials' vocabulary that was processed (see Fig 4).*

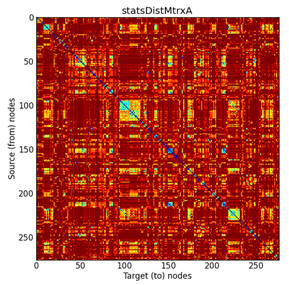
**Corpus Methods**

- Over 20 different glossaries, classifications, taxonomies, dictionaries, ontologies, thesauri and database schema were aggregated to form a multi-disciplinary domain corpus of concepts and terms. They were from institutions such as British Geological Survey, US National Aeronautical and Space Agency (NASA), US Geological Survey (USGS), US Army Corps of Engineers (USACE), Society for Sedimentary Geology (SEPM), Center for Deep Earth Exploration (CDEX) in Japan, and the World Meteorological Organization (WMO). The sources are already used in research and operationally, and are authoritative.
- The data were entered into a table format including the following columns: UniqueCode (unique identifier); rdf:prefLabel (clean name); Definition (verbose formal description); rdf:altLabel (other clean names); skos:related, skos:broader, skos:narrower (taxonomy assertions); and metadata. This had to be a manual process, but was table-based and therefore relatively efficient.
- The data were cleaned into two formats: (i) 'natural language' form, but with abbreviations, quantifications and other small items resolved; (ii) 'bag of words' form, using only strong word terms. The first was used for lexical analysis, the second for statistical analysis.
- For the statistical results the text was prepared to raw terms, spell-checked terms, stemmed terms, or n-grams which are n-sized fragments of the text. After considerable experimentation, stemmed text was preferred. NLP stemmers do not treat prefixes nor highly technical forms of text. The project created a stemmer in Python that is specialized to geosciences texts, for instance for 'pseudo-', 'meta-', etc.

**Lexical Extracted Relationships**

- The potential of automated lexical analysis is very great for successful mining of ontology from the corpus and also from texts in general. The present study exploited the potential only in a small way, not extending yet to Natural Language Processing (NLP).
- (i) some corpora already contain SKOS-type relations (W3c 2014) which were simply extracted.
- (ii) expressions such as "A is a B that has …", conveying that A is subsumed by B, were mined. This method follows those of WordNet (Millar 1995).
- (iii) The fact that geomaterials names themselves are often modified or refined by adding extra words, for example: "alkali basalt" as a refinement of "basalt" yielded extra subsuming relations.
- The relations extracted in this way were organized into a heirarchy using network links calculated by the Python package NetworkX (Hagberg & others 2008).

*Fig. 3. Statistical methods have the advantage that all concepts can be related, but the disadvantage of being unable to define any formal logic. This matrix depicts (in no particular order) the similarities of a subset of the geomaterials concepts based on term occurrences. Red are the more significantly related cases,*




**Statistical Extracted Relationships**

- Statistical word-distribution methods are widely used in computational linguistics for measuring the relatedness and generality of texts, and for concept disambiguation. Obviously, these methods cannot be used to strictly define subsumption relations, but they do perform the valuable function of associating terms - pending more precise lexical analyses. In this study entropy methods were used, for example the Joint Entropy Cosine Method of Tardelli & others (2004) which was found to give consistent and good results for similarity. For concept generality (breadth), entropy was again used - the Concept Generality measure of Benz & others (2011).

**Metrics**

- Once the geomaterials vocabulary was organized and analysed, the performance of various statistical and network metrics could be investigated with a view to later using them in the formal structures to assert relations and generalities.
- Of particular interest is how the lexical and statistical measures of relatedness behave together. This was put in terms of the probability of a degree of lexical relatedness for a value of statistical similarity (Fig. 3).

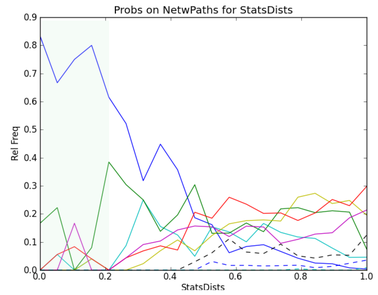



*Figure 3. Investigation of behavior between lexical and statistical measures of relatedness. Specific to the entropy-based similarity measures of this study, it appears that statistical similarities <0.2 (green shading) are highly correlated with lexical distances of just 1 (blue line). This could be a useful relation for construction of a structured vocabulary.*

- Geomaterials classifications and ontologies are generally drawn with a Level-of-Detail (LoD) heirarchy (though sometimes genesis-hypotheses are employed). To test LoD statistical concept-generalities were compared with lexical network rankings (from NetworkX). Only weak dependences were observed. However the results are compatible with the existence of localized, not absolute, LoD heirarchies in the geomaterials vocabulary.

**Conclusions**

- The structured vocabulary in its first iteration is available at "*tinyurl.com/dbseabed/resources/geomaterials/*", for purposes of critical review and feedback, and collaborative technical work. In particular, it is posted so that research can be conducted: (i) on how to optimize and present/visualize the structure, and (ii) building applications in query and crosswalk based on the structure.
- The postings include data in semantic web format (ConceptNet 5), adjacency matrices, and RDF SKOS ontology.
- Several significant technical hurdles remain, and the study has shown how these are likely to arise in any similar works on geosciences vocabularies. (i) That it must be decided whether to keep nodes that are repeated in the corpora separate (the 'granular' solution; as here), or merge them; (ii) That the statistical methods appear to perform better and more stably the larger the total corpus; (iii) The statistical results become less reliable with small sized glossary entries, suggesting that deeper forms of analysis drawing on synonyms may be needed.
- Overall, the results from the lexical ('Wordnet') forms of analysis proved highly satisfactory while the statistical results could only be accepted as useful indicators of relations. This is in accordance with many other studies. However, the statistical relations allow for discovery of relations and as such may be the most useful arm of analysis in advanced research applications.

**Acknowledgements**

Thanks to Doug Fils (Consortium for Ocean Leadership), Ruth Duerr (NSIDC, Boulder) and Thomas Wever (FWG, Kiel) for discussions on the topic. The funding support from NSF for the work is appreciated, under awards 1242909 ('Dark Data') and 1047776 ('Seamless Strandline').

**References**

Benz, D. & others 2011. One Tag to Bind Them All: Measuring Term Abstractness in Social Metadata. In: *The Semantic Web: Research and Applications, Lecture Notes in Computer Science*, 6644, pp 360-374.

Goff, J.A. & others, 2008. Seabed mapping and characterization of sediment variability using the usSEABED data base. *Continental Shelf Research*, 28(4-5), 614-633.

Hagberg, A. & others 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In: *Proc. 7th Python Sci. Conf. (SciPy 2008)*, 11-15.

Jenkins, C.J. 1997. Building Offshore Soils Databases. *Sea Technology*, 38(12), pp. 25-28.

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

Tardelli, A.O. & others 2004. An implementation of the trigram phrase matching method for text similarity problems. *Stud. Health Technol. Inform.*, 103, 43-9.

W3c 2014. *Introduction to SKOS*. [URL: "http://www.w3.org/2004/02/skos/intro"]