NSF Earthcube White Paper

Methods for Word-based Data in the Geosciences

Chris Jenkins [1], Doug Fils [2]

1. INSTAAR - U Colorado, Boulder CO (chris.jenkins@colorado.edu); 2. Consortium for Ocean Leadership, Washington DC (dfils@oceanleadership.org)

BACKGROUND

The word-descriptive data type is extremely common, even perhaps dominant in geoscience datasets. It describes the landscapes, soils, biota, features, processes and changes.

For example, in recent land/coast/marine seamless mapping projects (NSF supported), about 85% of all ground-site observations are presented as word-descriptive data. Furthermore, across all geosciences, the parameters for which even quantitative sensor data is acquired rest on concepts and definitions expressed by words. Word-based data is pervasive in the geosciences.

But word-based data is difficult to make operational in databases because of the indiscipline of vocabularies, the syntactic and semantic complexity of text, and frankly, lack of agreement about the destination for efforts at harmonization and integration. Solving these problems as we propose will make data base queries more complete and accurate, and will help the important efforts to integrate numeric sensor data and word descriptive observations.

This white paper proposes additional work with word-based geosciences data, experimenting with methods and algorithms that are new, or suggested by developments in linguistics in other sciences and commerce. The goal is for increased inter-workability between word-based and numeric data types.

PROPOSAL

Strategy

We draw on experience with the Integrated Ocean Drilling Program (IODP) SEDIS interdisciplinary and international data access system, the global seafloor mapping dbSEABED ("http://instaar.colorado.edu/~jenkinsc/dbseabed/"), and a recent Louisiana wetlands NSF project "Seamless over Strandline". Strong progress has been made in: (i) extracting meanings from word-descriptive text to the point where maps and homogeneous datasets are produced, and (ii) organizing vocabularies and semantics to exploit the new semantic web technologies including those from commerce.

Applications

Many geosciences databases have an unsolved problems with their parameter sets and descriptive data entries. For example:

(i) The Earthchem project ("http://www.earthchem.org/") recognizes the need to systematize its lithology/sediment/soil names so that database queries which are based on material type obtain full, yet focused results. "The whole truth, and nothing but the truth". For

this, semantic nets are needed to *cross-walk* between the various specialist domains that are contained in the database.

(ii) The PANGAEA database in Germany ("http://www.pangaea.de/"), perhaps the largest data warehouse of the geosciences, has over 90,000 separate parameters with data. Consequently, query results are fragmented and incomplete – a serious inefficiency in the conduct of the science. Our computed systematization of the parameter concepts has allowed a semantic net to be constructed which serves as an ontology to resolve this problem. The initial net was recently delivered to IODP-MI, but we understand the next steps that are required for increased sophistication and workability.

The problem of heterogenous data integration is an important contemporary challenge in computing. How can diverse datasets be merged, or at least made to be interoperable ? By attaching fuzzy set values to word concepts, the vocabularies are reduced to a common denominator – numeric data – accompanied by uncertainties. This is the key to integration of diverse formerly incompatible geosciences datasets. For example:

(i) At the data-divide between land and sea, seamless mappings are needed of ground conditions for input to models and decision systems. The onshore USDA and USACE soils, wetlands habitat classifications (NOAA), beach types (USGS) and offshore sediment types need to be harmonized so that a unified mappings of the composition, erodability, and bafflement are available to models.

Science / Technology Impacts

(i) More effective database searching, in particular, with semantic web and linked data (facet searching) technologies. Those technologies are still new and developing, but already provide the basic functionality to exploit integrations word-semantic geosciences data.

(ii) Greater cross-disciplinary workability of data. A knowledge of specialized vocabularies will not be as necessary for using the datasets from sister science disciplines.

(iii) The project is highly relevant to: groundwater and geothermal geology, land geologic mapping, resources exploration, marine environments, coastal and onshore soils, biogenic earth surface materials.

METHODS

The products will be data mediators ('cross-walks') between the domain-datasets of geosciences. Considering the scale of the problem, those products will need to be generated computationally.

These methods have been found to yield good results in terms of reliability and useability of results:

(i) Use of *domain specific languages* (DSL): natural language is not sufficiently reliable for science even though most geoscience data is in noun-phrase syntax – the most tractable linguistic structure; therefore a DSL like that of dbSEABED is needed; such languages are required to deal with and quantify terms for objects, modifiers, quantifiers and locators in unambiguous syntax.

(ii) Parameters have been repositioned in a *data-model*, where each is described in terms of *properties* such as "footprint", "units", "intended parameter", "target (material)". The same scheme, which has been reviewed and accepted by labs in several countries now,

may be highly useful across the geosciences, leading to linked-data/facet-search opportunities.

(iii) *Dictionary-thesaurus* (D-T) where terms and clichés in the DSL are given numeric or coded meanings, for instance in *fuzzy set theory* syntax; the dbSEABED D-T is over 12,000 terms long, and will be published. When convolved with the DSL data, the D-T is the key to production of regional landscape/soil/habitat mappings such as shown in Fig. 1.

(iv) *Computational linguistics* methods can efficiently generate reliable cross-walks between the geosciences' domain vocabularies. We have applied network theory including *entropy similarity* and *page-rank centrality* methods to build a hierarchical semantic net of the IUGS/BGS, IODP and GeoSciML rock vocabularies (Fig. 2). This is in the spirit of WordNet and Google methods. We have validated the results against manually written ontologies with very good *receiver-operator characteristic* (ROC) outcomes. This offers a solution to the problem of there being over 10,000 names for geomaterials, which need to be systematized.

For Earthcube, we want to become involved with data projects in a range of disciplines – marine, soils, geothermal/groundwater geology, geological mapping, vegetation studies and carbon repositories, coral reefs. We want to build an overarching set of methods which will bring their word-based data online: harmonized, statistically homogeneous, and quality assured.

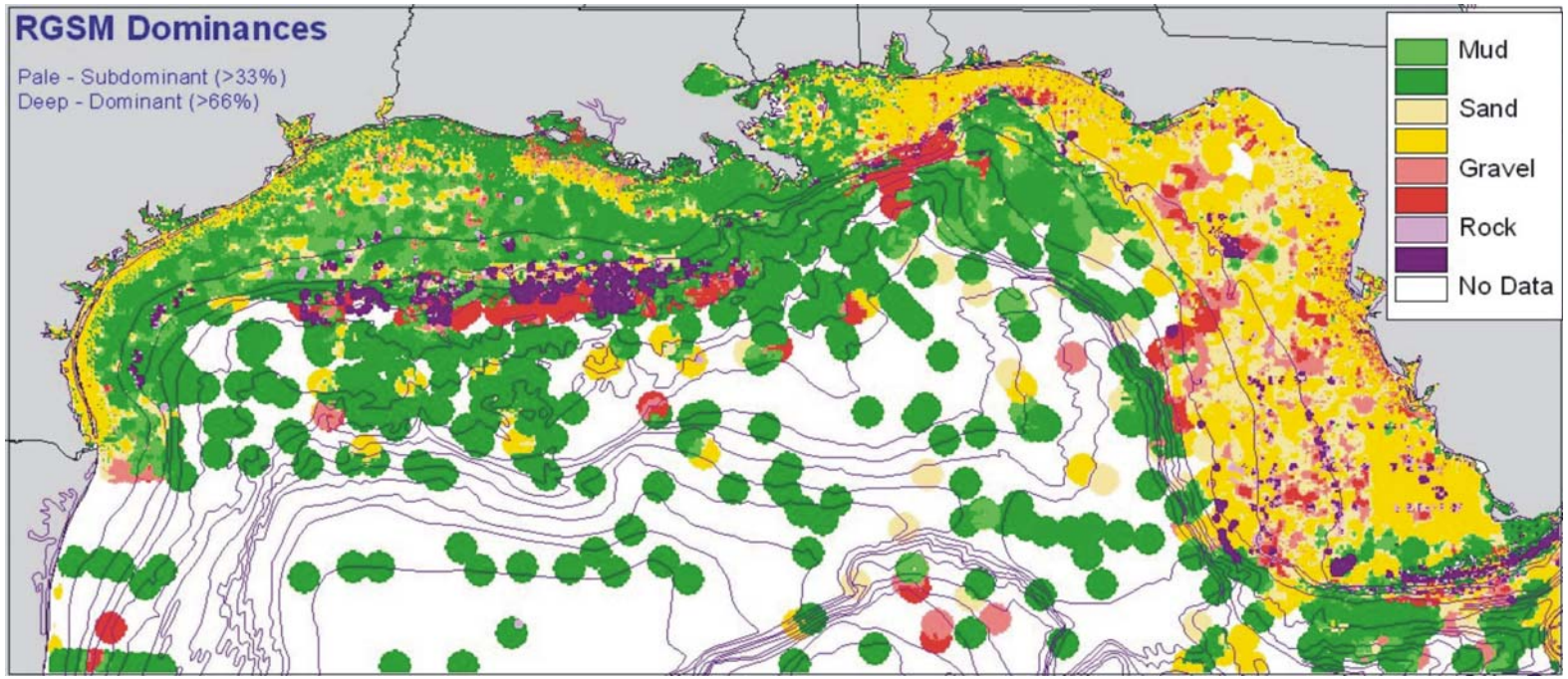

Fig 1. A regional picture of habitats – here marine, Gulf of Mexico – required the harmonization of diverse forms of data from ~10^5 sites – including data from geologists, biologists, surveyors, engineers, navy. The majority of the data was word-based descriptive – especially in the hard or biogenic areas – so operations on the word data were essential to creating the map. The manual work involved was minimal.

This product in more detailed form is now widely used in fisheries management, to support numerical models including ROMS, and to prepare against weather and pollution events (NOAA).

Fig 2. A prototype computed semantic net of lithologies, in effect an ontology to guide the machine-querying of complex word data on a large scale. The nets can be tuned to needs 'on the fly'. Machines can 'walk the net' or use the adjacency matrix to solve word-based queries. Here, lithology names from GeoSciML are shown in a computed heirarchical net (red, higher order concepts).