Semantic e-Science and Geology

Clinton Smyth[1], David Poole[2] and Rita Sharma[3]

[1] Georeference Online Ltd., Vancouver, Canada cpsmyth@georeferenceonline.com
[2] Department of Computer Science, University of British Columbia, Vancouver, Canada poole@cs.ubc.ca
[3] Georeference Online Ltd., Vancouver, Canada rsharma@cs.ubc.ca

Abstract

e-Science, as implemented for the study of geology with Geographic Information Systems over the Internet, has highlighted the need for standardization in the semantics of geology, and stimulated international action in this regard. This standardization is not intended to replace free thinking and expression, but to supplement it. Standardisation of these semantics is a difficult and costly exercise, which requires prototyping in order to be successful. Similarity ranking exercises provide a broadly relevant class of prototyping activity, and these are being actively pursued in a number of geologically related fields of scientific endeavour.

Introduction

Geology, perhaps more than any other science, has advanced with the aid of pictures. The pictures which are so important to geologists are the drawings they produce to show the distribution of different rock types in space – most commonly on the surface of the earth. These pictures are called geological maps, and current efforts to improve them have spawned a global e-science-based initiative to review the semantics of geology - the very foundations on which our knowledge of the earth depends.

With time, this initiative will likely lead to changes in the way earth scientists view the vocabulary they use and the way they conduct their scientific investigations.

Language and Science

It is self-evident that language – the assembly of words into sentences which convey information – is a fundamental building block of all scientific knowledge. On the contrary, however, inconsistencies in language are not self-evident, and tend to be noticed only in exceptional cases, unless they are specifically being sought. As such, they often lead to errors, or oversights.

When inconsistencies in language are sought, as, for example, in comparing how the US Department of Agriculture Forest Service (Haskins 1998) and the British Columbia Ministry of Environment, Lands and Parks (Howes 1997) describe and classify the same set of geomorphological phenomena, the differences may be so great as to make the non-initiated wonder how any effective geomorphological science is carried out by either august body. Figures 1 and 2 below illustrate how fluvial features are represented in the two classification systems.

Science Goals

To understand how good science is, in fact, conducted by our two example institutions, we need first to establish, in broad terms, the goals of scientific endeavours.

One useful statement of the goals of science is that science seeks to understand the processes operating in our universe, and how to interact with them in such a way as to improve the quality of life for all human beings.

Standardisation and Freedom of Expression

Freedom of investigation and freedom of expression have proven to be essential to the progress of science during the pre-e-science era. In times of repression, the rate of scientific advancement has slowed. By definition, in pre-e-science times, only human beings were doing the thinking. In today's world, with advances in computer technology and the understanding of linguistics (Pinker 1994), researchers are programming computers to do some of the "thinking". Many human beings resent the imposition of standardised structures and vocabularies on their scientific findings or models, which are necessary, at this time, to take advantage of thinking computers. (It is likely that in the future we will be able to program computers to be as tolerant of inconsistencies in vocabulary and semantics as human beings are.)

The dawn of the e-science era has changed this paradigm, and with it, the relative importance of freedom of expression. While freedom of expression will remain essential for many breakthroughs in science, breakthroughs will also be made by the adherence to standards of vocabulary and description when submitting descriptions

LandformClassification

LandformType	LandformTypeCom
root	from: A Geomorphic Classification System (United States Department of Agriculture)
Coastal Marine Landform	The collection of geomorphic processes occurring along or inland of, but related to, marine shorelines.
Eolian Landform	Geomorphic processes and landforms pertaining to the wind; esp. said of such deposits as loess and dune sand, of sedimentary structures such as wind formed ripple marks, or of erosion and deposition accomplished by the wind. (Bates and Jackson, 1995)
Fluvial Landform	Geomorphic processes pertaining to a river or rivers; produced by river action. (Bates and Jackson, 1995)
Fluvial Basin Landform	A collection of fluvial processes which occur in basins.
Bolson	(a) A term applied in the desert regions of the southwest U.S. to an extensive flat alluvium floored basin or depression, into which drainage from the surrounding mountains flows centripetally with gentle gradients toward a playa or central depression; an interior basin, or a basin with internal drainage. (Bates and Jackson, 1995)
Semi Bolson	A wide desert basin or valley that is drained by an intermittent stream flowing through canyons at each end and reaching a surface outlet (such as another stream, a lower basin, or the sea); its central playa is absent or poorly developed. It may represent a bolson where the alluvial fill reached a level sufficient to permit occasional overflow across the lowest divide. (Bates and Jackson, 1995)
Fluvial Slope Landform	Fluvial erosional and depositional processes resulting from overland flow or unchannelized flow on slopes.
Stream Landform	The collection of fluvial processes which occur or are directly related to streams.
Glacial Landform	Pertaining to distinctive processes and features produced by or derived from glaciers and ice sheets. (Modified from Bates and Jackson, 1995)

Figure 1. Part of the Landform Classification System of the USDA Forest Service.

to computer "thinking" processes, which would otherwise have taken much longer to be made. Computers processing very large consistently described data sets makes possible the recognition of patterns otherwise obscured by non-standard vocabulary.

Geologists' pictures are a fine example in this regard.

Geological Goals

Geology is the study of the solid, inanimate earth, and the processes that operate on it. Research ranges in scale from sub-microscopic to galactic.

When working at relatively large scales, geologists are invariably dependent on data and information gathered and reported by other geologists regarding smaller-scale features that they may never see in reality. They will see only the maps produced by their co-workers, and will attempt to build new knowledge from these.

e-Science and Geology

A consistent subject of scorn or jocularity in geology is the omni-present "map-boundary fault" - an apparent geological discontinuity along the boundary of adjoining maps, which is invariably nothing more than a difference in opinion or nomenclature between the individuals who mapped each sheet.

GeomorphologicalProcess

xValue	Description
root	root
Deglacial Process	
Erosional Process	
Fluvial Feature	
Anastomosing Channel	A channel zone where channels diverge and converge around many islands. The islands are vegetated and have surfaces that are relatively far above mean maximum discharge levels. Some channels are dry at moderate or low flows.
Braiding Channel	Active channel zone is characterized by many diverging and converging channels separated by unvegetated bars. Many channels are dry at moderate and low flows, but during major floods, the entire channel zone may be occupied by flowing water.
Irregularly Sinuous Channel	A clearly defined main channel displaying irregular turns and bends without repetition of similar features; backchannels may be common, and minor side channels and a few bars and islands may be present, but regular and irregular meanders are absent.
Meandering Channel	A clearly defined channel characterized by a regular and repeated pattern of bends with relatively uniform amplitude and wave length.
Hydrologic Process	

Figure 2. Part of the British Columbia Terrain Classification System

As our study of the planet earth has become more holistic in approach, these inconsistencies across provincial and national boundaries have become a hindrance to the advancement of geological understanding. This hindrance is receiving more attention (BBC 2007) because e-science, as delivered by a combination of Geographic Information Systems (GIS) and the Internet, has made it so much more visible, and provides some of the tools to remove it.

Semantic e-Science and Geology

The first concerted international effort to address these problems from an e-science perspective was initiated by 15 national geological surveys at a meeting in Edinburg, Scotland, in 2003, co-hosted by the British Geological Survey and the Geological Survey of Canada (Laxton 2004).

At the meeting, entitled "The Geological Data Model International Collaboration Inaugural Meeting", the 15 countries agreed to develop together a standard model for the expression of geological knowledge, particularly as it pertains to geological maps.

Three subsequent years of very active collaboration has produced a well-documented and working version of GeoSciML, which is defined as "a geoscience specific XML-based GML (Geography Markup Language) application that supports interchange of geoscience information" (CGI 2007). One of the first science goals to which GeoSciML is being applied is a complete semantically consistent Internet-delivered geological map of the world (PlanetEarth 2007).

Vocabularies need Ontologies

The collaborators have now turned their attention to vocabulary - the words that they need to populate the knowledge structures they have built (Richard 2006). These knowledge structures may be considered as "incomplete ontologies", and are essential pre-requisites for the production of vocabularies and semantics suitable for e-science. Without them, it is not possible to efficiently prototype candidate terms, taxonomies, partonomies and more complex semantic structures than have already been formalized.

Prototyping Semantics

Prototyping is typically an expensive activity. Prototyping standardized languages is no exception in this regard, and its high cost has held up the development of standardized terminologies across institutions, national boundaries, and across natural language groups.

Nevertheless prototyping is essential, with the consequence that it is important to identify semantics prototyping activities that have broad relevance. There is little sense in spending large amounts of money on semantics that satisfy niche activities, to the exclusion of broad scientific interests.

We have identified the related activities of "comparison" and "similarity ranking" as scientific activities of broad relevance, and hence high potential for semantics prototyping (Smyth and Poole 2004).

Comparison and Similarity Ranking

One of the most fundamental activities of an expert in any field is that of comparison. An example is a doctor comparing the symptoms of a sick person to her knowledge base of illnesses when seeking a diagnosis. In general, an expert compares one thing (the "in focus" thing) to a number of other things, and ranks the latter collection of things based on their similarity to the "in focus" thing.

The faster and more reliably this job can be done, the better the expert. There is often a tradeoff between speed and reliability.

There is a fundamental difference between "models", which are abstract concepts, and "instances", which are generally physical entities or events whose attributes can be measured, and about which True/False statements can be made. Equivalent statements about models, on the other hand, need to be qualified with probability-type qualifiers, such as "always", "sometimes", or "rarely", rather than with True/False truth-status qualifiers.

There are four types of comparisons that experts need to make:

Comparison Type	Example
Instance to many Models	Classification of a mineral deposit
Model to many Instances	Generation of a landslide hazards map
Instance to many Instances	Development of a land-slide classification system
Model to many Models	Evaluation of competing classification systems

Using ontologies which include standardized vocabularies it is possible to automate the similarity ranking activity for all four of the above comparison types.

Technologies for doing this in any domain are still in their infancy, despite significant growth of interest in semantic networks, and in ontologies over the last five years (Poole and Smyth 2005). While technologies for semantic descriptions exist and technologies for matching with probabilities and qualitative probabilities exist, their integration has been elusive.

Achieving success in this field depends on the solution of a number of related scientific and technological problems. These problems include the following:

(1) How expert descriptions of instances and models should be represented on the computer (issues attend both description structure and vocabulary);
(2) How the user interface to such representations should function so as to encourage expert and non-expert use of the system;
(3) How the similarity rankings should be derived (there are both user-dependent issues and user-independent issues remaining to be addressed here);
(4) How to explain to the user, in an easily understandable form, how similarity rankings were derived (i.e.: to provide an audit trail of results);
(5) How to provide advice to the user on the most salient information to gather to improve or reduce the quality of a similarity measure, as is necessary when a user is trying to distinguish between two close matches to his instance or model of interest.

All of these are challenges facing semantic e-science.

Vocabulary Prototyping Applications

Our research group is prototyping vocabularies for use in e-science by developing web services which can be used to store, compare, and rank on similarity, descriptions of models and instances of mineral deposits (Smyth 2004), landslides and landslide hazards (Smyth 2005a), and plutons (Smyth 2005b). We are also prototyping the

matching of optimum extractive metallurgical processes to ore types. The progression of these prototypes to thinking computer systems which aid thinking human beings in their pursuit of a better quality of life currently depends on standardized vocabularies and ontologies, one of the key current challenges to e-science.

Conclusion

There exists, in the use of e-science technologies, a tension between the need to standardize and the need for freedom of expression. Large scale geological mapping presents a compelling example of the benefits of standardization. The potential rewards of standardising how knowledge is represented on computers are greater than they ever were before, and this is causing much costly effort to be invested in developing ontological standards, including vocabulary standards. Comparison and similarity ranking activities are appropriate contexts within which to refine prototype semantics for use in various aspects of e-science.

References

BBC 2007. http://news.bbc.co.uk/2/hi/science/nature/6434011.stm

CGI 2007. http://www.cgi-iugs.org (2007)

Haskins, Donald M..; Correll, Cynthia S.; Foster, Richard A.; Chatoian, John M.; Fincher, James M.; Strenger, Steven; Keys, James E. Jr.; Maxwell, James R.; King, Thomas. 1998. A Geomorphic Classification System, version 1.4. Washington, DC: U.S. Department of Agriculture, Forest Service. 130 p.

Howes, D. E. and Kenk, E. 1997 A System for the Classification of Surficial Materials, Landforms and Geological Processes of British Columbia (Version 2) http://ilmbwww.gov.bc.ca/risc/pubs/teecolo/terclass/cove1.htm

Laxton, J. 2004. Geological Data Model International Collaboration: Report on the Inaugural Meeting http://www.bgs.ac.uk/cgi_web/tech_collaboration/data_model/docs/inaugural.html

Pinker, S. 1994. Chapter 4: How Language Works in *The Language Instinct* published by HarperCollins ISBN: 978-0060976514

Poole, D. and Smyth C. P. 2005. Type Uncertainty in Ontologically-Grounded Qualitative Probabilistic Matching *Proceedings of the Eighth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (ECSQARU-2005) Barcelona, Spain

PlanetEarth 2007. http://www.bgs.ac.uk/onegeology

Richard, S. 2006. Geologic Vocabularies and Reference Systems http://www.seegrid.csiro.au/twiki/view/CGIModel/GeologicVocabulary

Smyth, C. 2004 MineMatch http://www.georeferenceonline.com/ref/MineMatch2004/MineMatch2004.html

Smyth, C. 2005a The Role of Geological Terminology in Working with Landslides and Landslide Models *Presentation to the Geological Survey of Canada* Ottawa, Canada. http://www.gomatcher.com/ppt/HazardMatchOttawa02/HazardMatchOttawa02.html

Smyth, C. 2005b Researching Continental Evolution with Computer-Assisted Comparisons and Similarity-Rankings of Plutons *Abstracts of the Geological Society of America Annual Meeting*, Salt Lake City, Utah. http://gsa.confex.com/gsa/2005AM/finalprogram/abstract_94831.htm (includes slides)

Smyth, C. P. and Poole, D. 2004. Qualitative Probabilistic Matching with Hierarchical Descriptions in *Proceedings of the Ninth International Conference on Knowledge Representation and Reasoning* (KR2004) http://www.aaai.org/Library/KR/kr04contents.php