From “Digital Mapping Techniques ‘08—Workshop Proceedings”
U.S. Geological Survey Open-File Report 2009–1298
http://pubs.usgs.gov/of/2009/1298/

Vocabularies for Geoscience Information Interchange

By Stephen M. Richard[1] and David R. Soller[2]

[1]Arizona Geological Survey / U.S. Geological Survey
416 W. Congress # 100
Tucson, AZ 85701
Telephone: (520) 770-3500
Fax: (520) 770-3305
email: *steve.richard@azgs.az.gov*

[2] U.S. Geological Survey
926-A National Center
Reston, VA 20192
email: *drsoller@usgs.gov*

Introduction

Development of a digitally networked global community has progressed from simple text based interchange to progressively richer content, including audio, video, maps and imagery of all sorts. Metadata and semantic content descriptions are necessary for more effective search, discovery, and evaluation of these various types of information. In addition, the sheer volume of accessible content begs for more automation in the acquisition and analysis of data; the key to such automation is interoperability.

Information interoperability is built on a ‘stack’ of shared protocols and interchange formats (Figure 1). The hardware and network parts of this stack constitute the infrastructure of the Internet. There is a tremendous amount of ongoing work to develop file formats and schema to achieve the intermediate or schematic levels of interoperability—e.g., well-defined file formats (netCDF, SDTS, shapefile, XML…) and markup languages that implement particular domain-specific information schema. For example, GeoSciML (*http://www.geosciml.org/*) is a markup language developed for geoscience information interchange. Within this schema, there are various elements (like database fields) that are populated using geologic and other terminology lists. Semantic interoperability occupies the top of the information interchange stack and involves understanding the meaning of content transported via the underlying stack elements. Semantic interoperability requires agreement between data providers and data consumers on shared concepts and the mechanisms to represent concepts.

Interoperability is predicated on the idea that the data consumer and provider do not have to negotiate the format and content model for each information interchange individually. The engineering concept is to construct patterns or protocols (service definitions) for discovering, acquiring, and utilizing content that do not require the consumer to have any knowledge of how the provider is implemented. Semantic interoperability in such an architecture requires mediation between concept representations used by the provider and consumer if they do not use the same system; the simplest example of such mediation is language translation. Software tools for semantic mediation are still in their infancy, so the best way to know what someone else means is to use a shared vocabulary of controlled terms.

A controlled vocabulary is a collection of concepts. Each concept in the vocabulary has a definition and one or more assigned terms (e.g., names) that are, effectively, labels for the concept as used in everyday or scientific communication. Each of these terms has a scope—the community of users who use that term or label for the concept in the vocabulary. Typically terms are scoped by association with a language; for example, Spanish or French, or, if the word “language” is used in a more general but less familiar sense, “geoscience language.” Within any particular scope there should be a one-to-one mapping between terms and concepts. A controlled vocabulary may also include relationships between concepts (especially hierarchical relationships) (Richard and others, 2003). The identity of a controlled concept is based on its definition, not on the term used to label the concept.

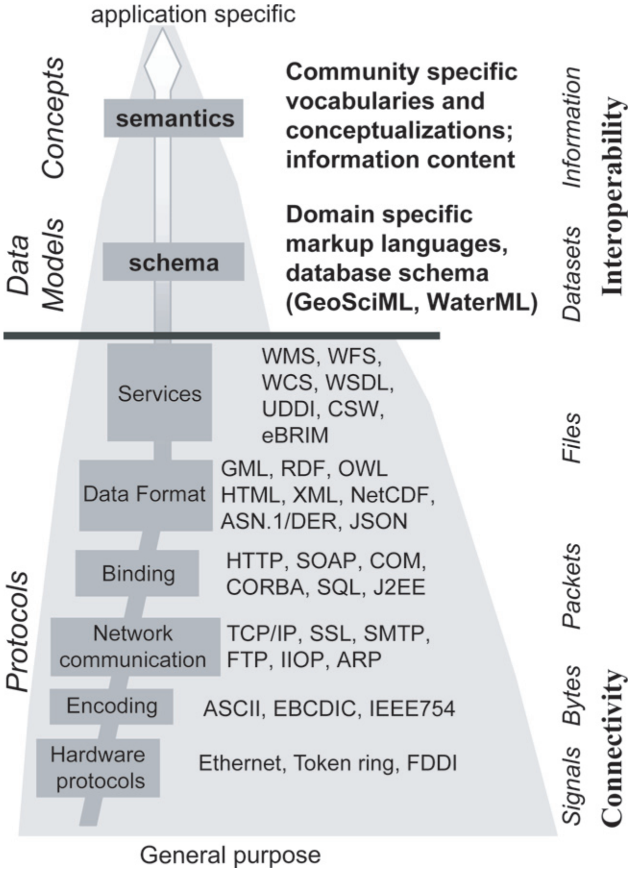
application specific
Concepts
Data Models
semantics
Community specific vocabularies and conceptualizations; information content
schema
Domain specific markup languages, database schema (GeoSciML, WaterML)
Information
Datasets
Interoperability
Services
WMS, WFS, WCS, WSDL, UDDI, CSW, eBRIM
Files
Data Format
GML, RDF, OWL HTML, XML, NetCDF, ASN.1/DER, JSON
Binding
HTTP, SOAP, COM, CORBA, SQL, J2EE
Packets
Protocols
Network communication
TCP/IP, SSL, SMTP, FTP, IIOP, ARP
Encoding
ASCII, EBCDIC, IEEE754
Bytes
Hardware protocols
Ethernet, Token ring, FDDI
Signals
Connectivity
General purpose

Figure 1. Interoperability stack. This diagram represents the collection of protocols and specifications that enable interoperability. General purpose, low-level specifications in the lower part of the diagram enable basic network connectivity, starting at the hardware level, working upward to services that specify collections of operations and basic transport protocols. These 'system' protocols are built on top of one another, with each layer dependent on the underlying layers and adding progressively more complex functionality, from basic signal interpretation ('0' or '1') to delivering digital files. Many of these protocols are so widely adopted and used that most users are not aware of them. The solid line in the upper part of the diagram separates these system protocols from more application specific data models that start to define domain-specific file content and structure; these have narrower applicability, but are necessary for computer-based automated content packaging and interpretation. At the top of the stack is the ultimate objective of interoperable system design—the conveyance of information between systems with only minimal human intervention.

A conceptual data model for the topic or domain of interest dictates the kinds of controlled vocabularies required. For example, the NADM-C1 (NADMSC, 2004) model includes 'WeatheringCharacter' as a property of a geologic unit. Because this property is specified by a term rather than descriptive, free text, a controlled vocabulary of terms that specify different weathering character values is necessary. The North American Data Model's (NADM-C1; *http://nadm-geo.org/*) use of a controlled term list was chosen to facilitate interoperability. Controlled vocabularies make possible the clear and unambiguous communication of content.

Vocabularies for Shared Use

The USGS National Geologic Map Database Project (NGMDB) has been supporting community development of standardized vocabularies for several years, mostly through participation in the NADM's Science Language Technical Team (NADM-SLTT, 2004; *http://pubs.usgs.gov/of/2004/1451/nadm/*), the GeoSciML Concept Definitions Task Group (*https://www.seegrid.csiro.au/twiki/bin/view/CGIModel/ConceptDefinitionsTG*), and at the project level (e.g. Richard and others, 2003). From this work, numerous vocabularies (approximately 42) either were defined or adopted for use by the NGMDB project. Most of these were compiled informally for project use and have not been published.

The development of these NGMDB vocabularies was coordinated with the project's efforts to (1) implement a federated (USGS – State geological survey) database from the NADM conceptual model; (2) design a data-entry tool for populating this database with geologic map information according to these standard vocabularies; and (3) convey a simplified view of the Nation's geology via a subset of the federated database and an interface, the NGDMB Data Portal. The Data Portal is described in Soller (this volume). For the purposes of the Data Portal, five of the compiled vocabularies were used – lithology, genesis, particle sorting, weathering character, and proportion (e.g. proportion of a geologic unit that is composed of a specified rock type). These five vocabularies are briefly described below. The NGMDB Data Portal vocabularies, and those developed in anticipation of a NGMDB federated database, are available online at *http://ngmdb.usgs.gov/Info/standards/NGMDBvocabs/*; previous versions of each standard vocabulary also are archived there.

Lithology

Science language technical teams formed under the auspices of the North American Data Model Steering Committee (*http://nadm-geo.org/sltt/*) developed vocabularies for sedimentary and metamorphic rocks, and adopted existing standards (Streckeisen, 1976) for use with igneous rocks (NADM-SLTT, 2004; *http://pubs.usgs.gov/of/2004/1451/nadm*). Our synthesis of these vocabularies into a single lithologic classification produced a vocabulary with 2027 terms. Over the past several years, project experience developing a user interface to utilize this vocabulary, and testing it with geologic users, demonstrated that this list is too large and the relations among terms too complex to be successfully utilized without considerable training.

Based on this experience, we determined that a smaller vocabulary would be necessary for integrating geologic maps displayed through the NGMDB Data Portal. The lithology category vocabulary for the portal will be used for searching and for online map services to report the composition of map units. Because the map services are to be accessible to a wide audience, we required that the terminology should be broadly understandable. Simultaneously, demonstration vocabularies for use with GeoSciML interchange documents were developed by the Concept Definitions Task Group of the Interoperability Working group of the CGI (CDTG). The senior author led development of both the NGMDB and CDTG vocabularies; they are identical except for some minor differences discussed below.

The CDTG vocabulary was assembled by a group of geologists from various countries, who discussed the kinds of lithology categories they thought should be included in a simple lithology vocabulary consisting of about 100 terms. As for the NGMDB Data Portal, the purpose of this vocabulary is data integration, not detailed scientific categorization of the full spectrum of materials found in the Earth. The initial list of terms was reduced and balanced in an attempt to include equivalent depth of detail for various families of rocks (igneous, sedimentary, metamorphic). Generalized category names had to be added in some cases where there is not a commonly used lithology term in order to allow construction of a hierarchy of categories (e.g., composite genesis material, fault-related material). The resulting vocabulary contains 146 terms, and is available at *https://www.seegrid.csiro.au/twiki/bin/view/CGIModel/ConceptDefinitionsTG*.

The NGMDB Data Portal lithology vocabulary has some minor differences with what has emerged as the CDTG (version 200811) vocabulary. These differences are discussed here. The NGMDB vocabulary does not include *Foidite* and *Foidolite*. These rocks, which consist of greater than 60 percent feldspathoid mineral, are distinguished in the CDTG 200811 vocabulary by grain size (phaneritic versus fine-grained), following LeMaitre and others (2002). For NGMDB purposes, these unusual rocks are not differentiated based on grain size, and so they are aggregated into one category, *Feldspathoid rich igneous rock*, to denote any igneous rock with more that 60 percent modal feldspathoid. The CDTG 200811 lithology vocabulary includes *phyllonite*; NGMDB does not include this because it is an unusual rock type that is sufficiently represented by the *Mylonitic rock* or *Phyllite* category. Several categories not included in the CDTG lithology vocabulary are included in the NGMDB lithology vocabulary. A generic *Compound material* category represents any sort of rock or unconsolidated material that is part of the Earth. NGMDB lithology also includes *Rock formed in surficial environment*, *Weathered rock*, and *Residual Material* categories to allow composition description of units that are mapped/defined based on presence of these sorts of materials. CDTG 200811 did not include such categories based on the argument that protolith or precursor terms should be used. This produces a potential incompatibility in that composition specified by one of these categories would have to map to CDTG 200811 *Unconsolidated material*, which may not be a very accurate mapping.

We have tested the NGMDB lithology vocabulary by using it to categorize lithology for State geologic maps of Arizona, Idaho, Oregon, and Washington, as well as the Geologic Map of North America. Our conclusion is that the vocabulary has worked well for this map synthesis, and we plan to continue using it. Variations with the CDTG vocabulary with the CGI Interoperability working group are being discussed and hopefully will resolve discrepancies between the vocabularies.

Genesis of Earth Materials

The purpose of this vocabulary is to define categories that may be used to specify the geologic origin, setting, and processes by which geologic units or materials were formed. The implementation of these aspects or properties of genesis is somewhat different in the GeoSciML v.2 model and the NGMDB Data Portal schema. The NGMDB portal follows the GeoSciML v1.1.1 scheme by associating a genetic category property (GrossGenesisTerm in GeoSciML v1.1.1) with a geologic unit. In GeoSciML v.2 the genesis of a geologic unit is disaggregated into a collection of one or more events, each with process and environment properties. The genetic categories in the NGMDB Portal vocabulary can be parsed into implied process or environment properties to map into the GeoSciML v.2 schema.

Particle Sorting, Weathering Character

Vocabularies for characterizing the particle sorting and weathering character of geologic units were included in the NGMDB data-entry tool software and were tested during the process of parsing into the Data Portal the geologic map descriptions on selected national and State geologic maps. Not unexpectedly, particle sorting and weathering character were seldom found to be generalizable for regional map units and so were not used in the Data Portal. They are provided here because we anticipate they will be more useful for detailed map descriptions in local areas. Regarding comparable vocabularies in GeoSciML, the NGMDB vocabularies were compiled before the CDTG work had advanced, so these term lists were submitted as contributions for the CDTG members to consider. When the CDTG completes its work, we anticipate adopting their vocabularies for future use.

Proportion

This vocabulary provides terms that may be used to qualitatively express the abundance of a rock type in a geologic unit. It is a simple list including Dominant, Present, Subordinate, Minor, and Rare.

Summary

An international community of geoscientists is working to develop shared vocabularies for information interchange. The advantage of using shared vocabularies is that a participating agency only has to do one mapping—to and from their agency's vocabulary to the standard, shared vocabulary. The downside is that information may be lost when specific agency terms must be mapped into generalized or non-equivalent terms in the shared vocabulary. This is offset by the substantial benefit for users, because they aren't required to interpret and understand the different terminologies in use by each data source. The NGDMB project has long supported this international effort and provides numerous science vocabularies at the website *http://ngmdb.usgs.gov/Info/standards/NGMDBvocabs/*. These vocabularies are relatively stable in their content, but some of them are still evolving. Therefore, they are here provided informally, and have not been fully critiqued and edited in order to meet USGS and other agency standards for editorial consistency. However, we anticipate they might be found useful by individual agencies and by the international standards-development community, as a resource and possibly for incorporation of the terms and definitions.

References

Le Maitre, R.W., ed., Streckeisen, A., Zanettin, B., Le Bas, M.J., Bonin, B., Bateman, P., Bellieni, G., Dudek, A., Efremova, S., Keller, J., Lameyre, J., Sabine, P.A., Schmid, R., Sorensen, H., and Woolley, A.R., 2002, Igneous rocks: A classification and glossary of terms: Recommendations of the International Union of Geological Sciences Subcommission on the Systematics of Igneous Rocks: Cambridge, Cambridge University Press, 236 p.

NADMSC (North American Data Model Steering Committee), 2004, NADM conceptual model 1.0, A conceptual model for geologic map information: U.S. Geological Survey Open-File Report 2004-1334, 60 p., *http://pubs.usgs.gov/of/2004/1334*.

NADM-SLTT, 2004, Report on progress to develop a North American Science-Language Standard for digital geologic-map databases, *in* Soller, D.R., ed., Digital Mapping Techniques '04 — Workshop Proceedings: U.S. Geological Survey Open-File Report 2004–1451, p. 85-94, *http://pubs.usgs.gov/of/2004/1451/nadm/*. [Includes all vocabularies as appendices to this report.]

Richard, S.M., Matti, Jonathan, and Soller, D.R., 2003, Geoscience terminology development for the National Geologic Map Database, *in* Soller, D.R., ed., Digital Mapping Techniques '03 — Workshop Proceedings: U.S. Geological Survey Open-File Report 03–471, p. 157-168, *http://pubs.usgs.gov/of/2003/of03-471/richard1/*.

Streckeisen, A., 1976, To each plutonic rock its proper name: Earth Science Reviews, v. 12, p. 1-33.