Establishing vocabularies for the exchange of geological map data—how to herd stray cats

Linda Bibby
GeoScience Victoria
1 Spring St, Melbourne
Victoria 3001
Linda.Bibby@dpi.vic.gov.au

SUMMARY

GeoScience Victoria (GSV) is introducing a new service delivery model to improve capture, storage and delivery of geoscience information. This major project has prompted significant effort to establish a data model with an accompanying vocabulary to govern what data to store in the model, and how to arrange and group the data.

The principles behind the vocabulary are that it must be robust, flexible, expandable and hierarchical. The vocabulary, like the data model itself, must satisfy both internal business requirements and international efforts with respect to information exchange in a global environment.

There are three components to the vocabulary: the terms themselves with their definitions/synonyms; arranging the terms into schemes; and the grouping of those schemes to provide meaningful information (handled by the data model). Our solution manages the vocabulary in an open source technology, with terms (and synonyms) arranged in hierarchies.

The process demonstrated that while computers may require rigid terminology from us, in some cases this is a real challenge. As an evolving science, many areas of geology have terms that mean different things to different people, with no single correct meaning.

Common vocabularies are essential to the exchange of geoscience information at national or international level, which is increasingly demanded via the internet, and GSV strongly encourages their development.

Key words: controlled vocabularies, information exchange, data management, geoscience information.

INTRODUCTION

GeoScience Victoria (GSV) has a major strategic project underway to upgrade its geoscience map-making processes and related applications, in order to overcome internal inefficiencies imposed by legacy data structure and software constraints. The project aims to:

- remodel the geological data into a data structure compatible with international data exchange standards;
- enable electronic exchange of data with other jurisdictions and exploration companies;
- build the necessary data repositories and processes for generating digital data and map products;
- deliver a greater range of structured, searchable earth science data products to meet varying client requirements;
- produce a better range of quality maps more efficiently;
- provide improved science quality and business processes compatible with international standards; and
- position GSV to take advantage of emerging web-based technologies, such as Web Feature Services, eScience, Grid technologies and standard data exchange languages, such as GeoSciML.

The project centres on a new data model, with most fields populated via controlled vocabularies. The internal benefits are significant, including improved workflows, better data quality and shorter time between data collection and provision. Internal users will have access to more powerful query capabilities, improved methods of data collection and interrogation, and higher quality, more consistent data. The project is due for completion by 2008.

External considerations

New scientific insights and improved understanding come from recognising patterns and connections, an effect that can be greatly enhanced when combined with a collaborative approach. With the increasing need for collaboration on complex geological problems, whether with internal or external parties, comes an increased need for data integration based on a set of quality, structured data.

There is also a push for online access to data from within and without GSV. Our clients, many of whom are located outside of Victoria, have consistently expressed their information needs in terms of straightforward access to data that is on-line, up-to-date, high quality, cheap, web-based, and integrated using exchange language technologies. However, geological problems do not recognise government jurisdictions, and once organisations want to use data from more than one source, challenges rapidly arise with respect to communication and exchange of digital data. A common data model and exchange language is part of the solution, but this is insufficient without a supporting vocabulary.

International work on data modelling and an exchange format, such as GeoSciML (Simons *et al.*, 2006) will move us to a new way of exchanging data, while at the same time presenting a new set of problems. Traditional hardcopy media, such as maps, reports and papers, often incorporate personal or house styles, and as such rely on the skill of the user to interpret the intent of the author. Access to digital data

lets a user see the raw data, but the interpretive qualities of a map or journal article, replete with human-readable language, are lost. Added to this, different government agencies, academics and industry use different preferred terminologies, making it difficult to compare different datasets. Consequently, there will be a need for scientists to be more rigorous and consistent with the language they use.

Collaboration and exchange of data requires structured descriptive information. A seemingly simple solution would be to set up a single vocabulary that all geoscientists would use. But this is at odds with the way geology has evolved for the past 200 years: the guidelines and codes we have are largely voluntary, and based around common usage. For example, in 1976 the IUGS advised that use of the term 'adamellite' was to be discouraged, but even a few years ago it was appearing in formal unit names on Australian geological maps. Most geologists can still interpret the meaning of the term and can understand the map as a result—and the mapmakers would probably argue that the term conveyed a meaning they felt would be lost if they used another term. Similarly, a debate over the use of dolomite/dolostone in the early 1990s raised blood pressures around the country and led to a new Stratigraphic Names Committee policy statement. In 1998, the question of whether to Suite or not to Suite provoked four years of vigorous debate, culminating in the first guide to nomenclature for igneous rocks in Australia (Brakel, 2005). Looking further afield, Cox and Allibone (1995) argued that naming rocks on lithology alone in southern Victoria Land hinders the understanding of the area, and that structural terms should be available to geologists when naming Antarctic rocks; a view at odds with the SCAR Working Group on Geology (Ricci *et al*., 1993; Krynauw *et al*., 1995).

What these examples demonstrate is that geoscientists constantly test geological concepts and how they are communicated, a process fundamental to our science. From the process of geological investigation will come new ways to name units, describe faults and classify fossils. The earth sciences are constantly evolving, and any system must be able to accommodate both advances in understanding and geological language, irrespective of the outcomes of the arguments.

The problem for exchange of data with other organisations is not one of technology, but science. It is not enough to have a place in the database for information about a rock: in order to compare data, the same rocks must be described in the same way. GSV has an internal solution that looks towards future developments, but the broader problem of data exchange requires consensus from the geoscience community at large, which must determine for itself what terms are used to communicate concepts.

UNDERLYING PRINCIPLES

If you gave any geologist a lump of rock, they could probably assess it and provide some sort of description for it. That description would vary, though, depending on the background of the individual and the intent and detail of the description (Figure 1). Mafic rock, basalt and tholeiite are all valid names for the same rock, and it may also be described as scoria. Others would analyse it, then tell you it was remanently magnetised, a basanite with CIPW normative olivine, or that it was a rock with an elemental suite prospective for gold and base metals. All are equally valid—and valuable—descriptions, but without standard vocabularies to organise the data, all would be separate entities in a computer system, with no relationship between them. All tholeiites are basalts, but not all basalt is tholeiite, or scoria. Add to this that the IUGS (Le Maitre, 2002) suggests the term tholeiite should be replaced by tholeiitic basalt, and any system will struggle to capture the meaning of this information.

A rock that might be described as:

- MAFIC ROCK (name based on colour index; observation)
- BASALT (type of volcanic rock; observation)
- SCORIA (textural term; observation)
- THOLEIITE (type of basalt; thin section examination) = THOLEIITIC BASALT (IUGS, 2002)
- BASANITE (type of volcanic rock; chemical analysis)
- ROUNDED COBBLE (particle size and shape)

Figure 1. An example of a rock that can be named in multiple ways (Photo: D. Bibby).

The legacy GSV system has a single field containing a flat list of terms for basic rock type. If basalt, tholeiite, scoria and basanite are all in the same list, the data entry is immediately ambiguous (multiple valid choices) and one search will not find all the relevant information. Solving these types of problems is a major driver behind remodelling our science language and work processes, to ensure we have a structured vocabulary within which to work.

As a guide, we decided that any vocabulary we came up with had to be:

- able to be understood by a wide range of users, with explicit definitions of terms;
- unambiguous;
- flexible enough that we could capture a diverse range of geological concepts;
- hierarchical, so terms that were subsets of other terms were grouped as 'children' of the parent term;
- expandable, so that new terms and concepts could be added as they were needed;
- robust enough that any changes would not break the whole system;
- based on user requirements and international standards where available, to ensure we don't end up with the 'perfect' scheme that no one uses.

METHODOLOGY

Translating even seemingly basic geological knowledge to a digital database is not a simple task. Every system has different fields, with different terms in those fields. Individual geologists and organisations all approach their information management with varying levels of experience, areas of expertise and interests. Geochemists want different information than do physical vulcanologists or mineral explorers, and ideally, as a government organisation, we should try to cater for as many different end-users of our data as possible.

The initial stages of the broader project included extensive surveys of internal staff and collation of external requirements that highlighted the good, the bad and the ugly within our current system, and what different people saw as the future of geological data management. Analysis of the parts of the information that related to science language formed the basis of a plan to erect a standard vocabulary for GSV, and showed that if we could satisfy the needs of our internal users, we would also meet most of the external requirements.

There are three separate components to the vocabulary:

- the terms themselves with their meanings/synonyms;
- the arrangement of terms into schemes; and
- the grouping of schemes to provide meaningful information (handled by the data model).

Some important international guidelines are already available, such as the IUGS systematics for igneous rocks (2002) and the International Stratigraphic Guide (ISSC, 1994, 1999). However, wherever there are guidelines, controversy usually follows close behind. Many hours of debate with internal users helped to strike a balance between common usage, international standards, preferred terminology, practical schemes, and creating a system that was not too complex to implement. This inclusive process also served to give the internal users a sense of ownership of a science language based on their requirements.

Terms

We gathered many of the terms from international standards and dictionaries where these were available (eg. Le Maitre, 2002; Jackson, 1997). All terms are linked to explicit definitions and referenced, and synonyms or alternate spellings are captured (eg. aeolian/eolian; euhedral/idiomorphic; tourmaline alteration/tourmalinisation/ tourmalinization). The effect of this is quite striking: for example, our existing data contain sixteen different permutations of ‘sandstone’, much of it within free text fields that limited search capabilities (Figure 2A). There are no relationships between these terms: wacke and quartz wacke are unrelated terms in the database. The data highlighted how critical it was that we store and deliver data in a meaningful way.

Organising the terms into hierarchies has also been a powerful improvement to our data (eg. Figure 2B). This allowed us to capture the different ways to describe the basalt–tholeiite–basanite–scoria example (Figure 1), instantly improving the quality of our data. Simple hierarchies and synonyms solved all of the ambiguous language problems, while allowing us to maintain the necessary flexibility of geological terminology.

Schemes

The schemes presented some additional challenges. Some terms seemed to naturally group together (eg. anticline/syncline), but others could easily fit into multiple schemes depending on the context (eg. ‘igneous’ can describe a texture, a process or a genesis). A long process of research and group debate highlighted what our geoscientists had to have to do their jobs and what they would like to see in the vocabulary. Some basic questions guided us in some lively debates: ‘what terms belong in this scheme?’ and ‘where would you expect to find this term?’ Another important part of the process involved researching how other organisations are dealing with the classification, storage and exchange of geological data, in particular the geological surveys of Britain, the U.S. and Canada (eg. SLTT, 2004; Struik *et al*., 2002). We also consulted IT specialists to assist with implementation, and incorporated external comments from companies and individuals as to what they wanted from our data.

Data model

The new GSV data model is an implementation of the NADM C1.0 conceptual data model (NADM Steering Committee, 2004), the structure of which influenced many of the natural groupings of terms. GSV representation on the IUGS Commission for the Management and Application of Geoscience Information (CGI) data modelling subcommittee has also helped us to improve our version of the data model and associated vocabularies.

The CGI charged the data model working group with developing the GeoScience Mark-up Language (GeoSciML), a harmonised geoscience data model and exchange format to enable the delivery of web-based data (see Simons *et al*., 2006). However, as outlined above, with information in a totally digital environment, scientists cannot afford to be as free with their language as traditional media has allowed.

GeoSciML is designed around particular data types, including terms from controlled vocabularies. The most recent meeting of the CGI Working Group in Toronto identified that the lack of a set of standard vocabularies will be a limitation to the usefulness of GeoSciML. It is not clear where the work to establish these vocabularies will come from.

DISCUSSION

The new GSV vocabulary satisfies all of the original requirements:

- it is the result of rigorous debate and incorporates international guidelines where these are available;
- it is built on a list of ‘GSV preferred’ terms, with synonyms, that are arranged in hierarchies for unambiguous data entry and more powerful searching;
- new schemes and terms are easily accommodated if required, and both schemes and terms can be updated;
- processes are in place to keep the system flexible and trackable; and,
- the internal users have participated in the entire process, and signed off on all schemes.

Short term benefits

GSV has implemented a new geology data model, supported by the vocabulary we have erected. The vocabulary currently contains terms related to geological units and structures, with more to follow in line with the incremental implementation of the GSV data model.

Many areas are working well and have provided some instant improvements, but there are still challenges for us to meet. For instance, the lithology hierarchy is working well with basic terms, but multiple values and qualifiers need more work, for example attaching 'quartz–biotite–actinolite' (in that order) to the root name 'schist' to generate a human-readable, searchable, rock description.

Introducing a hierarchical vocabulary was a significant departure from the in-house technology available to us. We have implemented an open source ontology manager that allows for synonyms and hierarchical arrangement of both schemes and codes for our new vocabulary. The new system knows that an arenite is a type of sandstone, so both specific ('find all the arenites') and general ('find all the sandstones') searches will return all relevant data. The vocabulary also provides us with better flexibility for definitions. For example, we can define 'igneous' with respect to the context, and a search for 'igneous' will return all the relevant data.

Immediate benefits include:

- higher quality science through the increased use of controlled vocabularies over free text fields;
- improved data management and querying capabilities; and
- a platform for data exchange and collaboration.

The process has also put us in a strong position to enter into external debate, but we are only one voice in a global community. There will be other ways to erect and implement vocabularies, and as such we need to have national and international consensus on the language we wish to use to exchange digital data.

National and international initiatives

Australia is working towards a national set of standard vocabularies, in order to influence international developments. We have agreement on schemes for an early version of the National Geological Data Model (NGDM), which will form the basis for future work. The Chief Government Geologists' Information Policy Advisory Committee (GGIPAC) has established the Controlled Vocabularies working group to address this on a national level. Australia hopes to use this group to influence international developments in data exchange. At the very least, we hope to establish a common set of national vocabularies for Australian use, which can also form a sound base from which to contribute to international activities. The working group includes representatives from all state and territory geological surveys.

One important controlled vocabulary already accepted nationally is the database of Stratigraphic Names administered by Geoscience Australia on behalf of the Geological Society of Australia for all Australian geoscientists, based on the International Stratigraphic Guide (1994, 1999; see also Staines, 1985). The simple act of allowing an external committee to control stratigraphic nomenclature is an excellent start, and an achievement to which very few other countries can lay claim. In Australia, we can be confident the Eumeralla Formation is a volcanolithic sandstone in the Otway Ranges of Victoria, and that adjacent states do not contain different units with the same name. Likewise, a new South Australian formation should not be named Broadhurst because the Broadhurst Formation already exists in Western Australia. It is not an onerous or strict process for a scientist to erect a stratigraphic unit, but the system works.

The management of stratigraphic names demonstrates that it is possible to implement standard vocabularies in Australia, and the approach represents one possible way to develop and endorse national vocabularies for other areas of the earth sciences.

Agreement on vocabularies to support the exchange of geological information on a national or international level is a long-term project in its infancy, and requires many more voices to succeed. By working towards a national solution, we can position ourselves to influence international efforts so that our requirements are recognised.

Our experience indicates the biggest threat to any standard comes from within: balancing internal business requirements with the need to collaborate and conform to external standards is not always a smooth process. History has shown that if the standards don't meet our needs, we won't use them. However, what may be possible is an exchange vocabulary, where each organisation can map their preferred terms to common schemes and definitions, with the potential for growth as the earth sciences evolve. GSV sees this work not only as an obligation we have as a geological data repository, but also as essential to enabling global information exchange.

CONCLUSIONS

Getting geoscientists to agree to a single, rigid language is a lot like herding stray cats: it's difficult, unlikely to succeed, and is probably not worth the effort. The challenge is to accommodate diversity and to allow advances in understanding, but this is asking more of our computer systems than ever before, and requires careful consideration and international consensus to be effective. If we get our language organised, information systems will be able to accommodate the geological diversity that we need them to.

Our new system and the accompanying vocabulary are based on collecting data with a view to how it may be used. This will provide us with consistent data entry, which will improve the quality of our data irrespective of external vocabulary developments. Our work practices have changed, and the data meets corporate requirements for data collection. By looking outside for standards rather than setting them up ourselves, we are looking to the future.

Replacing free text fields with controlled, easily searched, hierarchical vocabularies is a powerful improvement. The geologists that built the hierarchies are those who will use them, so they are both specifically geared to their work requirements and compliant with international standards. This will deliver an early benefit to us, via our web mapping

services, but it also allows for future machine to machine exchange.

The top-down search capabilities offered by the new hierarchy will allow more powerful querying. A search for a general term will return all instances of specific terms as well, and a specific search will return specific data—even if the search term is a synonym for a GSV-preferred term. Much richer data mining will be the result.

Consistent vocabularies are an evolving field, and GeoScience Victoria has established principles for developing and managing its own based on flexible, expandable concepts, implemented in an open source technology. The implementation of our classification manager and hierarchical schemes will improve digital data interrogation, but represents only one possible solution to the problem. Although we are considering possible future international developments, the process needs to involve the wider geological community.

We believe our solution shows that rich datasets can be constructed and interrogated without the need to enforce inflexible terminology on the geoscience community, but we recognise that this is only one answer to a problem with many possible solutions. The only 'real' answer is to promote discussion and debate—and through this move towards some practical standards with wide acceptance.

ACKNOWLEDGMENTS

Much of the work is the result of extensive debate with GSV geoscientists. The author also benefited from discussions with Bruce Simons and Alistair Ritchie. Ollie Raymond compiled much of the data for the GGIPAC Controlled Vocabularies working group. Alan Willocks and Fons VandenBerg reviewed drafts of the paper.

REFERENCES

Brakel, A., 2005. Of dolomite, dolostone and adamellite. *The Australian Geologist* **136**, p. 8.

Cox, S.C. and Allibone, A.H., 1995. Naming of igneous and metamorphic rock units in Antarctica: recommendation by the SCAR Working Group on Geology (Discussion). *Antarctic Science* **7(3)**, pp. 303–304.

Hallsworth, C.R. & Knox, R.W.O'B., 1999. BGS Rock Classification Scheme Volume 3: Classification of sediments and sedimentary rocks. British Geological Survey Research Report, RR 99-03.

International Subcommission on Stratigraphic Classification (ISSC), 1994. *International Stratigraphic Guide—A guide to stratigraphic classification, terminology, and procedure (Amos Salvador, ed.), 2nd edition.* The International Union of Geological Sciences and The Geological Society of America, Inc., 214 pp.

International Subcommission on Stratigraphic Classification (ISSC), 1999. International Stratigraphic Guide—An abridged edition (Michael A. Murphy and Amos Salvador, eds.). *Episodes* **22(4)**, pp. 255–271.

Jackson, J.A., 1997. Glossary of geology, 4th edition. American Geological Institute, Alexandria, Virginia.

Krynauw, J.R., Ricci, C.A., Herve, F. and LeMasurier, W.E., 1995. Naming of igneous and metamorphic rock units in Antarctica: recommendation by the SCAR Working Group on Geology (Reply to discussion). *Antarctic Science* **7(3)**, pp. 304–306.

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., Sørensen, 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 University Press.

North American Geologic Map Data Model Steering Committee Science Language Technical Team (SLTT), 2004. Report on Progress to Develop a North American Science-Language Standard for Geologic-Map Databases. Digital Mapping Techniques '04—Workshop Proceedings. *U.S. Geological Survey Open-File Report* **2004–1451**.

North American Geologic Map Data Model Steering Committee, 2003. NADM Conceptual Model 1.0 for Geologic Map Information: preliminary website release under the auspices of the Geological Survey of Canada, the U.S. Geological Survey and the Association of American Geologists.

Ricci, C.A., Herve, F., Krynauw, J.R. and LeMasurier, W.E., 1993. Naming of igneous and metamorphic units in Antarctica: recommendation by the SCAR Working Group on Geology. *Antarctic Science* **5**, pp. 103–104.

Simons, B., Boisvert, E., Brodaric, B., Cox, S., Duffy, T., Johnson, B.R., Laxton, J.L. and Richards, S., 2006. GeoSciML: enabling the exchange of geological map data. *Proceedings of the Australian Earth Sciences Convention 2006, Melbourne, Australia.*

Staines, H.R.E., 1985. Field Geologist's Guide to Lithostratigraphic Nomenclature in Australia. *Australian Journal of Earth Sciences* **32**, pp. 83–106.

Struik, L.C., Quat, M.B., Davenport, P.H. and Okulitch, A.V., 2002. A preliminary scheme for multihierarchical rock classification for use with thematic computer-based query systems. *Geological Survey of Canada Current Research* **2002-D10**, 9pp.

Wentworth, C.K., 1922. A scale of grade and class terms for clastic sediments. *Journal of Geology* **30**, pp.377–392.

A

Old scheme (flat list and free text)
arenaceous
arenite
arkosic
psammite
psammitic
quartzwacke
quartzwackes
sand/sandstone
sandstone
sandstone/quartzite
sanstone
subgreywackes
sublitharenite
sublitharenites
wacke
wackestone

B

New scheme: 'Descriptive rock name'				
	Name	**Description**	**Reference**[1]	**Synonyms/ comments**
1	**sandstone**	clastic sedimentary rock composed predominantly of fragments 0.032-2 mm in diameter	Wentworth	psammite
1.a	arenite	sandstone with less than 15% matrix	BGS3	
1.a.i	quartz arenite	sandstone in which more than 95% of the clasts are quartz grains	BGS3	orthoquartzite quartzite
1.a.ii	subfeldspathic arenite	sandstone in which clasts are 5-25% feldspar, 75-95% quartz and less than 25% rock fragments	BGS3	
1.a.iii	feldspathic arenite	sandstone in which clasts are 25-100% feldspar, 0-75% quartz and less than 50% rock fragments	BGS3	arkose
1.a.iv	sublithic arenite	sandstone in which clasts are 5-25% rock fragments, 75-95% quartz and less than 25% feldspar	BGS3	
1.a.v	lithic arenite	sandstone in which clasts are 25-100% rock fragments, 0-75% quartz and less than 50% feldspar	BGS3	
1.a.vi	calcarenite	a limestone consisting of more than 50% sand-sized carbonate grains	Jackson	
1.b	wacke	sandstone with 15-75% matrix	BGS3	wackestone
1.b.i	quartz wacke	wacke in which more than 95% of the clasts are quartz grains	BGS3	quartzwacke
1.b.ii	feldspathic wacke	wacke in which clasts are 5-100% feldspar, 0-5% quartz and up to 50% rock fragments; feldspar > rock fragments	BGS3	arkose
1.b.iii	lithic wacke	wacke in which clasts are 5-100% rock fragments, 0-5% quartz and up to 50% feldspar; rock fragments > feldspar	BGS3	greywacke

1. Wentworth: Wentworth grainsize scheme (Wentworth, 1922); BGS3: British Geological Survey guidelines for naming sedimentary rocks (Hallsworth et al.*, 1999); Jackson: Glossary of Geology, 4th edition (Jackson, 1997).*

Figure 2. Comparison of how 'sandstones' are managed in old (A) and new (B) GSV models. In the first table, a user can choose only one of any of the terms. In the second table, 1.b.i is a child term of 1.b, so the system knows a wacke is a type of sandstone. Synonyms and children can be added as required. If another scheme for naming sandstones is required, such as a scheme of names based on environment of deposition, it can be incorporated by the system and both names could be captured, eg. 'lithic wacke' and 'turbidite'.