Implementing a Glossary and Vocabulary Service in an Interdisciplinary Environmental Assessment for Decision Makers

Simon N. Gallant[1], Rebecca K. Schmidt[1], and Nicholas J. Car[2]

[1] CSIRO Land and Water Flagship, Canberra, ACT, Australia
[2] CSIRO Land and Water Flagship, Brisbane, QLD, Australia
{simon.gallant,becky.schmidt,nicholas.car}@csiro.au

Abstract. When delivering scientific information for decision makers, it is important to define and use appropriate terminology to ensure scientific credibility and good communication. A glossary with terms from authoritative sources for specific domains can increase the usefulness and reusability of information for decision makers as the information can be more easily used without adaptation or translation. Linked Data principles and semantic web-based vocabulary tools provide mechanisms for delivering formalised glossaries via vocabulary services for use in integrated products, both documents and information platforms.

Issues to consider when implementing a glossary and vocabulary service are covered: persuading stakeholders to accept standard external terms and gain agreement on unique terminology; requirements for gathering, controlling and maintaining terminology in a glossary to ensure transparency and persistence; formalising a glossary as a standards-based vocabulary; and efficiently implementing this glossary via automation.

Keywords: glossaries, web services, interoperable information systems.

1 Introduction

This paper explores the implementation of a glossary service in the Bioregional Assessment Programme (the Programme) [1,2]. The Programme provides information on the ecology, hydrology, geology and hydrogeology of specified bioregions with explicit assessment of the potential direct, indirect and cumulative impacts of coal seam gas and coal mining development on water resources. This scientific information will be available for all interested parties, including Australian and state government regulators, industry, community and the general public, when considering coal seam gas and coal mining developments.

The Programme is delivering over 150 products, mostly scientific reports, over the course of three years. A key requirement for these products is a high standard of scientific and editorial quality including the consistent use of terminology. For plain-English words, this is straightforward as existing literature can be used. For more technical language, standards must be agreed upon and recorded, including both negative and positive instruction ('Use *bore* not *well* in the context of groundwater.

R. Denzer et al. (Eds.): ISESS 2015, IFIP AICT 448, pp. 380–387, 2015.
© IFIP International Federation for Information Processing 2015

Use *well* not *bore* in the context of oil or gas.'). Through interacting with experts in many fields, a Programme-specific language list has been developed, which guides the writing, integration and quality assurance of content for all Programme staff.

The limitation of this list is its simplicity. When an author knows precisely what they wish to say, but not exactly which word to use, this list provides authority. What the list cannot do is help multiple authors agree on what they mean, nor inform readers as to that meaning. For the Programme to publish its products in a way that is truly useful to and accessible by the public, the way words and concepts are used needs to be discoverable by readers. For this, a controlled, authoritative glossary service is proposed.

This paper provides a short background on the Programme, Linked Data, ontologies and controlled vocabularies in order to establish the context of the work. The processes by which terminology, both individual words and entire sets of words from particular authorities, is agreed and governed are described, as is the architecture for automatically building product-specific glossaries. Finally, the costs and benefits of such a service are discussed, as well as the implications for multiple-use services such as this, with particular reference to the difficulties of (i) conflicting requirements, (ii) multiple-context reporting and (iii) doing something rather than nothing.

2 Background

2.1 The Bioregional Assessment Programme

A bioregional assessment is a scientific analysis of the ecology, hydrology, geology and hydrogeology of a particular geographic location, with explicit assessment of the potential direct, indirect and cumulative impacts of coal seam gas and large coal mining development on water resources [1,2]. The Programme undertakes these assessments for a range of stakeholders including the Independent Expert Scientific Committee on Coal Seam Gas and Large Coal Mining Development (IESC), Australian and state government regulators, industry, community and the general public. The outputs are a suite of scientific products for each of the geographic locations currently being studied, delivered both as documents and via an information platform.

The Programme team spans both scientific disciplines and research agencies with four main collaborators: the Australian Government Department of the Environment; the Bureau of Meteorology; the Commonwealth Scientific and Industrial Research Organisation (CSIRO); and Geoscience Australia.

2.2 Linked Data, Ontologies and Controlled Vocabularies

In this Programme, multiple agencies contribute and multiple fields of research are involved so information from a diverse range of sources must be integrated. Semantic web [3] technologies such as standardised vocabularies[1] and Linked Data [4], are

[1] See the W3C's listing of Semantic Web technologies including vocabularies at `http://www.w3.org/standards/semanticweb/`

designed with heterogeneous data integration in mind and are thus of great utility to this Programme. Terms from a range of authorities in a range of formats can be integrated for a single purpose, then placed within semantic web vocabularies. The delivery of them as Linked Data assists this greatly. By using Linked Data, terms become properties of objects that are identified using Uniform Resource Identifiers (URIs)[2] meaning they can be linked to and information about them 'dereferenced' (looked up) by following their URI. This allows the term owners (the authorities or acting on behalf of the authorities regarding their definition) to deliver them at a single point of truth and in both human- and machine-readable formats, enabling unambiguous references (links) to individual terms within text (documents, webpages). If standardised concept ontologies, such as the Simple Knowledge Organization System (SKOS) [5], are used for vocabularies, multiple properties for terms may be recorded, not simply textual definitions. SKOS allows the mapping of terms between vocabularies using a range of relationships such as *broad*, *close* and *exact*. This allows for nuanced relationships between the constructed glossary and other known, trusted vocabularies.

Software tools, such as the Spatial Information Services Stack Vocabulary Service (SISSVoc, [6]), deliver controlled vocabularies with formalised relationships between terms defined using SKOS as Linked Data. Other vocabulary delivery tools do exist, such as the Australian National Data Service (ANDS) Controlled Vocabulary Service [7], but controlled vocabularies are more commonly delivered in informal ways without standardised information models (ontologies) and without formal data formats. For example, the Australian Government's Interactive Functions Thesaurus (AGIFT) http://agift.naa.gov.au/ delivers its controlled vocabulary via regular webpages.

3 Glossary and Vocabulary Services

3.1 Persuasion and Approvals

Editorial quality is required to be of a very high standard for products of the Programme. Part of ensuring high editorial quality is ensuring that language is consistent both within and between products. This required consistency in language spans the choice of terminology to the way that concepts are used. The Programme's interdisciplinary nature has made this particularly challenging as experts from different disciplines have different ways of expressing similar concepts and different uses for the same terms. Two approaches have therefore been taken: (i) to use an external authority for definitions wherever possible, and (ii) to discuss, collate and socialise a language list which is governed by Programme management.

External authorities are valuable as they provide a point of truth, once contributors agree that it is appropriate. For instance, Programme members have agreed to use the *Australian Oxford dictionary* [8] avoiding many arguments over terminology (such as

[2] A 'URI' is similar to the more commonly known 'URL'. See `http://en.wikipedia.org/wiki/URI` for more details.

whether to use 'modeling' or 'modelling'). In the experience of the Programme, the following is necessary to gain agreement to use an external authority:

1. the majority of Programme staff have access to that authority
2. that authority includes a sufficiently volume of terms to make it worthwhile using
3. the majority of Programme scientific leaders already agree with the majority of terms.

Similar projects have devised related language lists, for example for the Sustainable Yields projects [9] and the Great Artesian Basin Water Resource Assessment [10]. In addition, the BA methodology [2], Australian Government's *Style manual* [11] and the *Australian Oxford dictionary* [8] were accepted as authorities. This provided the Programme with a sufficiently comprehensive language list to begin with, which was endorsed by Programme management as part of the development of the products, thus developing a first-pass list of approved terms.

As the Programme progressed it became clear that it was important to move beyond simply specifying a list of approved terms, but rather to give definitions. This is best practice when writing, particularly in interdisciplinary projects where readers and co-authors from different disciplines might have different meanings for the same word. This confusion needs to be avoided within the single context of a product and the broader context of the whole Programme. Some middle ground between two disciplines must be determined, or one discipline must use a different word. This problem cannot be solved by the imposition of a rule based upon personal preferences – it can only be solved by having conversations with the scientists involved and coming to agreement. For efficiency, the discussion ideally would start with determining external authorities that each discipline accepts so that glossaries from them could be adopted, then it would move to the task of defining individual terms that are not already defined in any external authority. External authorities that will be considered by the Programme include: the METOTERM database [12], the Australian Water Information Dictionary [13], and the Water Quality vocabulary developed for the Bioregional Assessment Framework [14].

A method of discourse based in participatory research methods was used to facilitate this agreement. The involvement in the decision-making processes of parties those decisions will affect has been encouraged since the 1950s and it has been long argued (for example, [15]) that increasing this involvement will improve outcomes. The persuasion work done is best described as a Partnership [15]: groups of scientists are given the power to negotiate, trade-off and come to agreement both with each other and with the editing and managerial teams, but the editing and management teams retain the power to formalise these decisions. Once decisions are formalised, they are presumed to be fixed.

3.2 Governance

What are the requirements for the process to gather, version control and maintain terminology in a glossary in order to ensure transparency and long-term persistence? The governance of terminology in the Programme is managed from the time an undefined term is identified right through until that term is published in the glossary.

At the first step, identification, the term is entered in the glossary as 'proposed, awaiting editing'. The definition of the term is then raised with those concerned (editors, authors, managers, subject matter experts), and a member of the glossary team facilitates the discussion. The glossary team member revises the definition in the glossary, changing its status to 'edited, awaiting approval'. Edited terms are then submitted to Programme management for acceptance at which point they gain a status of 'approved'.

The glossary will hold a large number of terms in various stages at any given time. Various status-based subsets of these terms will be shown through the use of different views. Only 'approved' terms will be available in the Public view, only 'edited, awaiting approval' terms will be visible in the ForApproval view, and all terms will be visible in the Management view. All views are able to show multiple definitions for a term, but this functionality is intended to be used infrequently. It is expected, that the Public view would only ever show the most recent definition.

The Public view can be used as an authoritative list of approved terminology, fulfilling the function of enforcing consistency. An agreed list of terms is a necessary part of publishing products with a high standard of scientific and editorial quality with transparency. Using the glossary as a way of tracking the changes and approvals of the way terms are used makes it possible to ensure that the products of the Programme are consistent in their terminology, with a transparent process for defining, approving and possibly redefining terms.

This process to submit and accept terms is informed by the Geographic Information standard ISO19135 [16] and future efforts will be made to harmonise the term stage naming with existing semantic web ontologies that handle resource lifecycles.

The management and governance of the glossary is important due to its multiple requirements: (i) the glossary is the complete list of terms for which rules on usage and spelling have been made and (ii) the glossary is an audited and controlled list of important terms and their definitions. While those terms that fall under (ii) can be included in (i), the reverse is not practicable. Therefore, there is conflict in the use, governance and maintenance responsibilities and requirements for the glossary as a whole. To resolve this issue, it is helpful to simplify the idea of 'the glossary': the glossary is a structured way of storing terms, some of which may have agreed definitions. The necessary additional maturity and complexity that comes from having multiple lists of terms can therefore be solved by using filter criteria for different views. An Authors view can then be presented, showing the complete list of approved words, without definitions even when they exist (addressing (i)), and the Public and ForApproval views show only those important, defined terms (addressing (ii)).

3.3 Formalising the Glossary as a Standards-Based Vocabulary

Once glossary terms have been identified and even before definitions are agreed upon, a URI for each term is generated. These are intuitive when designed well and take the following form for the Programme's terms:

```
http://{ProgrammsDataWebsiteAddress}/glossary/term/{term-
label}
```

With a URI and the text of the term (known as a *prefLabel* in SKOS) now known, the bare-minimum requirements of SKOS have been achieved. Once the term's definition is settled, that is added as a *definition* and SKOS relationships such as *broader*, *narrower* or *exact* can be determined. This information can easily be managed in the same media – a spreadsheet or database – as used to store the term's status and other information required by governance.

Once SKOS data are stored, a computer script can be used to automatically load the terms and their properties into a vocabulary service such as SISSVoc for delivery on the web.

3.4 Automated Implementation

The Public view of the glossary can also be used to generate product-specific glossaries in multiple formats.

The products are written using standard document preparation software which is easily available to all Programme contributors. For the traditional delivery of documents, a print-style glossary is required: a list of terms at the end of a product, with a definition for each. Where defined terms exist in the product, they should be linked to their individual location in the glossary through text indicating the reference. For the web-based delivery, the Public view of the entire glossary should be accessible through a link on any page and, where defined terms exist in these products, they should be hyperlinked to their individual location in the online glossary.

The production of both of these forms can be automated using computer scripts, which manipulate documents and can read information from web services. Post-processing of document files can identify defined terms used, inject referencing links and auto-assemble the print-style glossary. Similar processing can take place for marked-up files for web-based products.

4 Discussion and Conclusion

The primary cost involved in setting up a glossary service such as the one described here is time: a great deal of it is required in both system development and administration, as well as the discourse that is necessary to gain agreement on terms.

The benefits of having a glossary, however, should not be understated. The glossary is a binding context for Programme products. When a reader encounters a term they find ambiguous they are able to find its definition and can be sure that it will still be defined in the same way every time it is used. Thus, the clarity of written communication in Programme products is greatly increased and the trust that a reader places in Programme products is improved.

It is important to note that the Programme is large enough for such an activity to be worth undertaking. Relative to the total time for which the Programme will run (3 years) and the number of people involved (more than 160), the time and people required to develop and maintain the glossary system is reasonable. The same would not be true for a small project with few staff. However, the system and processes are

relatively easily transferable to other projects, thus expanding the benefits without much additional cost.

The Programme is reliant on credible, authoritative external sources of definitions: if external authorities for glossary terms are *not* available or are *not* accepted by the disciplinary scientists, the cost in developing and maintaining a glossary (by writing hundreds of definitions) may not offset the benefits.

Using one glossary service to fulfil many functions adds to the maturity of the Programme. In addition to improving communication and increasing trust, the problem of conflicting requirements is solved in part: instead of having a language list for authors and editors, a printed glossary for readers and a series of meeting minutes indicating managerial approval of terms – all of which must be aligned – the Programme is able to include all the necessary information in a single store and then provide the filtered information as subsets of the whole.

The difficulties of multiple-context reporting have not been fully addressed by the Programme. While the goal of having both complete, cover-to-cover reports *and* 'chunks' of online context is admirable, the shift of context both for authors and readers is difficult. It is believed that by providing a centralised service that ensures consistent definitions for terms, some of this context shift can be avoided.

Beginning development despite fluctuating and conflicting requirements has been valuable, and parts of the system can be implemented while others mature. The need to develop the glossary system incrementally has prompted the involvement of some Programme staff earlier than would have been anticipated, which should increase buy-in from many Programme staff. This is the most important outcome of the glossary development for if it is not valued it will not be used, and a glossary that is not used is not a glossary at all.

Acknowledgements. This work was funded by the Bioregional Assessment Programme, a scientific collaboration between the Australian Government Department of the Environment, Bureau of Meteorology, the Commonwealth Scientific and Industrial Research Organisation (CSIRO), and Geoscience Australia. For more information visit `http://bioregionalassessments.gov.au`.

References

1. Department of the Environment: Overview of the Bioregional Assessment Programme (viewed March 26, 2014), `http://www.environment.gov.au/coal-seam-gas-mining/pubs/overview-bioregional-assessment-programme.pdf`
2. Barrett, D.J., Couch, C.A., Metcalfe, D.J., Lytton, L., Adhikary, D.P., Schmidt, R.K.: Methodology for bioregional assessments of the impacts of coal seam gas and coal mining development on water resources. A report prepared for the Independent Expert Scientific Committee on Coal Seam Gas and Large Coal Mining Development through the Dept. of the Environment. Dept. of the Environment, Australia (2013), `http://www.environment.gov.au/coal-seam-gas-mining/pubs/methodology-bioregional-assessments.pdf`

3. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American Magazine (2013), http://www.sciam.com/article.cfm?id=the-semantic-web&print=true
4. Heath, T.: Bizer. C.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory Technology, Morgan & Claypool (2011)
5. Miles, A., Bechhofer, S.: Skos simple knowledge organization system - reference (2014), http://www.w3.org/TR/skos-reference/
6. Cox, S., Mills, K., Tan, F.: Vocabulary services to support scientific data interoperability. In: European Geosciences Union General Assembly 2013, Vienna, Austria. Göttingen, Germany, April 7-12, p. 1. Copernicus Publications (2013), http://meetingorganizer.copernicus.org/EGU2013/EGU2013-1143.pdf
7. Australian National Data Service: ANDS Controlled Vocabulary Service (2014), http://ands.org.au/services/controlled-vocabulary.html
8. Australian Oxford Dictionary, 2nd edn. Oxford University Press. (viewed March 26, 2014), http://www.oxfordreference.com/view/10.1093/acref/9780195517965.001.0001/acref-9780195517965
9. CSIRO (2014) Sustainable Yields Projects (viewed October 29, 2014), http://www.csiro.au/Organisation-Structure/Flagships/Water-for-a-Healthy-Country-Flagship/Sustainable-Yields-Projects.aspx
10. Ahmad, M., Schmidt, R.K., Marston, F., Cuddy, S., Mahoney, J.: Editing conventions for authors and editors. A document in the CSIRO Great Artesian Basin Water Resource Assessment reporting tools series. CSIRO, Canberra (2012), https://publications.csiro.au/rpr/pub?list=SEA&pid=csiro:EP1210495
11. Australian Government: Style manual for authors, editors and printers, 6th edn. John Wiley & Sons, Australia (2010)
12. World Meteorological Organization:METEOTERM database, http://wmo.multicorpora.net/MultiTransWeb/Web.mvc (viewed October 29, 2014)
13. Bureau of Meteorology: Australian Water Information Dictionary, http://www.google.com/url?q=http%3A%2F%2Fwww.bom.gov.au%2Fwater%2Fawid%2F&sa=D&sntz=1&usg=AFQjCNHtnJRpK23FADXkaB-uuWBjTSE_Ww (viewed October 29, 2014)
14. Simons, B.A., Yu, J., Cox, S.J.D.: Water Quality vocabularies for the Bioregional Assessment Framework. Water for a Healthy Country Flagship Report series. CSIRO, Canberra (2013) ISSN: 1835-095X
15. Arnstein, S.R.: A Ladder of Citizen Participation. Journal of the American Institute of Planners 35(3), 216–224 (1969)
16. International Organization for Standardization: ISO 19135:2005 Geographic information - Procedures for item registration (2005), http://www.iso.org/iso/catalogue_detail.htm?csnumber=32553