20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Defining a water quality vocabulary using QUDT and ChEBI

B.A. Simons[a], **J. Yu**[a], **& S.J.D. Cox**[a]

[a] *CSIRO Land and Water*
Email: bruce.simons@csiro.au

Abstract: Vocabularies of observed properties and associated units of measure are fundamental to understanding of groundwater, surface water and marine water quality observations. The ability to support annotation of observation values from the disparate data sources with appropriate and accurate metadata is crucial to achieving interoperability; that is, precisely describing metadata relating to the magnitude of a physical quantity denoted by a unit of measure in a machine readable format.

Previous projects have developed stand-alone water quality vocabularies, which provide limited support for cross-system comparisons or data fusion. We propose that relationships between the water quality concepts, the associated chemical entities and appropriate units of measure be represented formally using ontologies. As part of the development of a new water quality ontology for groundwater and marine domains, we have reused and aligned our definitions with existing ontologies. We use the 'Quantities, Units, Dimensions, Data Types' (QUDT) ontology to provide a consistent model of measurable quantities and units, and the Chemical Entities of Biological Interest (ChEBI) ontology for observations relating to chemical concentrations. In this paper we demonstrate reuse of these ontologies for describing observation values in the water quality domain. We show how QUDT can be used to define additional quantity kinds and units of measure relevant for the domain and how use of ChEBI enriches the water quality ontology, while maintaining separate ontology governance.

Keywords: *ontologies, water quality, QUDT, ChEBI, vocabularies*

1. INTRODUCTION

This paper describes a water quality ontology that has been developed for the Australian Bioregional Assessment Framework (http://www.environment.gov.au/coal-seam-gas-mining/bioregional-assessments, Moran, 2012) and eReefs (http://ereefs.org.au) projects that require access to environmental measurements of water quality. The requirement was that water quality measurement data stored in various and diverse State and Federal government surface and ground water databases be harmonized to allow for information exchange (interoperability).

The major problems with the existing water quality terms that needed to be resolved were:

a) Ambiguity –concepts are poorly defined, often with multiple concerns merged into one, such as the object of interest (e.g. nitrogen) and the parameter being measured (e.g. concentration) both labeled as 'nitrogen'. A common occurrence was that the parameter concept was included with the observation method in a single definition, and in some cases the unit of measure, containing medium and method (e.g. "*Total nitrogen, water, filtered, milligrams per liter*"; http://waterdata.usgs.gov/nwis/qw).

b) Inconsistent governance – various data providers manage the terms with spreadsheets and database look-up tables. Essentially the same terms appear in multiple vocabularies, but relationships between terms from different data providers, hierarchies and collections are limited or non-existent.

c) Modularity – specific scientific disciplines require access to vocabularies not restricted to or governed by that discipline, in addition to their domain-specific vocabularies. For example, water quality requires access to chemistry, unit of measure and biological vocabularies. These should be incorporated into the specific domain as required with minimum effort.

d) Not interoperable – the use of local, non-resolvable identifiers, lack of a formal definition and the lack of an ontology describing the relationship between concepts all precluded the re-use of the terms in a distributed computing environment and restricted their use by other domains or projects.

The challenge was to develop and apply a consistent methodology that would produce semantically rich water quality concepts. The methodology used here aims to address these ambiguity, governance, modularity and interoperability issues to enable interoperable information exchange.

A data provider or aggregator will typically start by collecting the terms and concepts used in specific projects, and organize them into a vocabulary, and make this available somehow to data providers and consumers. However, vocabularies relevant to the domain may already exist, which have had significant development, domain level agreement and are maintained by a recognized authority. Thus, where possible, the relevant authoritative definitions of these standard ontologies should be adopted and leveraged.

The water quality ontology developed here uses the concepts specified in the 'Observed Property' model presented in Open Geospatial Consortium *Observations and Measurements v1.0* (Cox, 2007), and *Quantities, Units, Dimensions, Data Types v1.1* (QUDT; Hodgson and Keller, 2011). The resulting ontology specifies the relationships between measurements on properties of objects of interest for water quality, the types of measurements, and the units permissible with those types of measurements.

2. BACKGROUND

In information science, an **ontology** formally represents knowledge as a set of concepts, the relationships between pairs of concepts and the set of logical assertions that apply. Ontologies provide the ability to reference the formal semantics of a model (i.e. a concept). This in turn supports a level of interoperability and information exchange between systems and representations of domain information.

A suite of semantic web standards is published by the World Wide Web Consortium (W3C) for to facilitate exchange between systems. The Resource Description Framework (RDF) provides the basic foundation. RDF Schema (RDFS) and the Web Ontology Language (OWL) provide a set of classes and properties that allow various levels of knowledge representation, including Description Logic, to be expressed using RDF. The Simple Knowledge Organization System (SKOS) is an OWL application for vocabularies and thesauri. Using SKOS, a vocabulary can be easily formalized with a syntax and logic compatible with the Semantic Web. The concepts are denoted by a Uniform Resource Identifier (URI), and 'labelled' with a term and its synonyms. Some basic taxonomic relationships are built into SKOS, such as broader/narrower relationships within a vocabulary, and mapping relationships to concepts in other vocabularies. SKOS/RDF allows members of a vocabulary to be explicitly connected to members of related vocabularies, and published as a set of web resources. This allows the vocabulary to be used to classify data in a web context.

Formalizing the water quality concepts using SKOS allows providing the standard vocabularies to web users. This required: a) asserting externally governed OWL classes as subclasses of the SKOS Concept, e.g. the QUDT ontology; and b) transforming externally governed OWL classes to SKOS Concepts, e.g. the ChEBI ontology.

Finally, we want to achieve the above requirements without disrupting the governance, requirements and formalisms of the respective vocabularies and ontologies, and facilitate a maintainable strategy of curating an aligned community vocabulary or ontology.

3. METHODOLOGY

For this application these existing vocabularies were used:

1. Observed water quality concepts derived from supplied by Bureau of Meteorology from the Queensland Healthy Headwaters database (Worley Parsons, 2012).
2. QUDT provides a unified model of units and quantity kinds required for observation metadata (http://www.qudt.org/; Hodgson and Keller, 2011).
3. Chemical Entities of Biological Interest (ChEBI) OWL is a large ontology with more than 30 000 fully annotated chemical entities that was developed with considerable domain expertise (http://www.ebi.ac.uk/chebi/; Degtyarenko et al. 2008, Hastings et al. 2013).

3.1. Quantities and units of measure

Vocabularies and ontologies of measurement, quantities and its set of units are fundamental components in any scientific, engineering, government, and trade domain for reasons of provenance, reproducibility, and transparency of the underlying observations and datasets. Precise metadata relating to the magnitude of a physical quantity denoted by a unit of measure in a machine readable format is crucial for achieving interoperability.

A number of formal units of measure systems exist:

- 'The Unified Code For Units of Measure' (UCUM; Schadow and McDonald, 2009),
- 'Ontology of units of measure' (OM; Rijgersberg et al., 2013),
- 'Measurement Units Ontology' (MUO; Berrueta and Polo, 2009),
- 'Ontology of Units of Measurement' (UO; http://code.google.com/p/unit-ontology/),
- Within the 'Semantic Web for Earth and Environmental Terminology' (SWEET v2.2; http://sweet.jpl.nasa.gov/2.2/reprSciUnits.owl and http://sweet.jpl.nasa.gov/2.2/quan.owl),
- 'Quantities, Units, Dimensions and Data Types in OWL and XML' (QUDT; Hodgson and Keller, 2011).
- Quantities, Units, Dimensions, Values (QUDV; de Kooning et al. 2009)

These use different modelling approaches and formalisms ranging from simple vocabularies enumerating units of measures, to alignment of measurement related concepts to upper ontologies. They also differ in their coverage of the set of unit of measures. Of the established and well-governed unit of measure ontology options, QUDT is well-aligned with our understanding of the relationships between measurements and units of measure. In the model described below, we focus on the QUDT *Unit* and *QuantityKind* OWL classes Nevertheless, alignment with the other vocabularies is desirable future work.

3.2. Observable properties model

Most existing water quality vocabularies conflate a number of concerns into each term or concept. These include the quantity-kind, such as concentration of an organism or substance, the specific substance or taxon being measured or counted, and often also the methodology and units of measure, leading to 'properties' such as "*Chlorophyll, total, water, fluorometric, 650-700 nanometers, in situ sensor, micrograms per liter*" (http://waterdata.usgs.gov/nwis). In the water quality observable properties ontology (Figure 1) a key distinction is between 'observable properties' such as '*nitrogen concentration*' that are formalized as instances of the class *wqop:ScaledQuantityKind),* and the associated 'identified objects' such as '*nitrogen*' which are formalized as instances of the class *wqop:SubstanceOrTaxon*.

Vocabularies for the units of measure (instances of the '*qudt:Unit*' concept) and the kinds of quantities (instances of the '*qudt:QuantityKind*' concept) were imported from QUDT, supplemented with units of measure from the existing water quality vocabularies and any additional required quantity kinds.

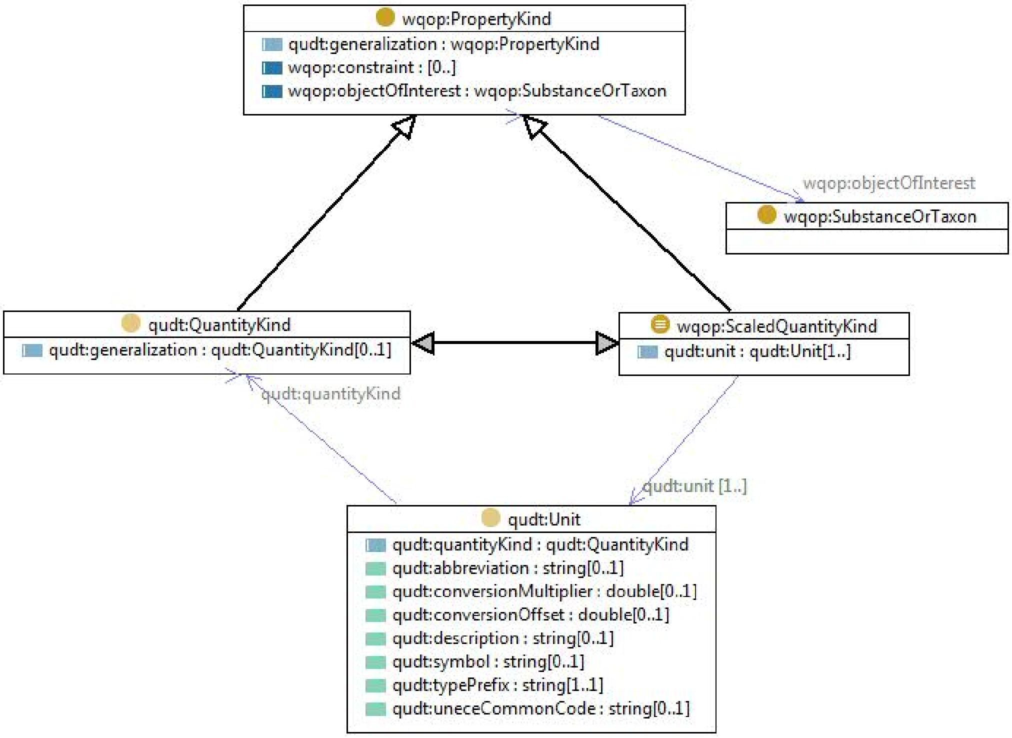
wqop:PropertyKind
qudt:generalization : wqop:PropertyKind
wqop:constraint : [0..]
wqop:objectOfInterest : wqop:SubstanceOrTaxon
wqop:objectOfInterest
wqop:SubstanceOrTaxon
qudt:QuantityKind
qudt:generalization : qudt:QuantityKind[0..1]
wqop:ScaledQuantityKind
qudt:unit : qudt:Unit[1..]
qudt:quantityKind
qudt:unit [1..]
qudt:Unit
qudt:quantityKind : qudt:QuantityKind
qudt:abbreviation : string[0..1]
qudt:conversionMultiplier : double[0..1]
qudt:conversionOffset : double[0..1]
qudt:description : string[0..1]
qudt:symbol : string[0..1]
qudt:typePrefix : string[1..1]
qudt:uneceCommonCode : string[0..1]

Figure1. Water quality ontology linked to QUDT ontology.

Two technical issues with adopting QUDT were identified:

QUDT semantic relations

QUDT uses *skos:exactMatch* (which is a type of *skos:semanticRelation*) property to relate members of the vocabulary to equivalent dbpedia concepts. However, *skos:exactMatch* has also been used to link OWL classes in QUDT, which is inconsistent with the use of the DL sublanguage to allow reasoning support and performance.

There are two potential responses to this:

1. Accept the use of SKOS relations, and rely on the 'punning' capability of OWL2-DL[1]
2. Modify QUDT by converting relationships between classes and individuals to use 'annotation properties' such as *rdfs:seeAlso*.

As part of this work the second approach was taken. The QUDT OWL classes were asserted as subclasses of the SKOS Concept with the dbpedia references remodelled as annotations using the *rdfs:seeAlso* property and imported into the QUDT SKOS ontology (Figure 2).

Incomplete QuantityKinds and Units

QUDT contains 1484 units of measure implemented as instances of *qudt:Unit*, and 236 quantity kinds implemented as instances of *qudt:QuantityKind* or its specializations. However, it is missing some units of measure and kinds of quantities required for the water quality data. To rectify this, we added 41 additional units of measure, and 17 quantity kinds, all in separate namespaces to the original QUDT. The conversion factors between these new units and the existing QUDT units may be defined using the QUDT mechanics.

[1] http://www.w3.org/TR/2012/REC-owl2-new-features-20121211/#F12:_Punning

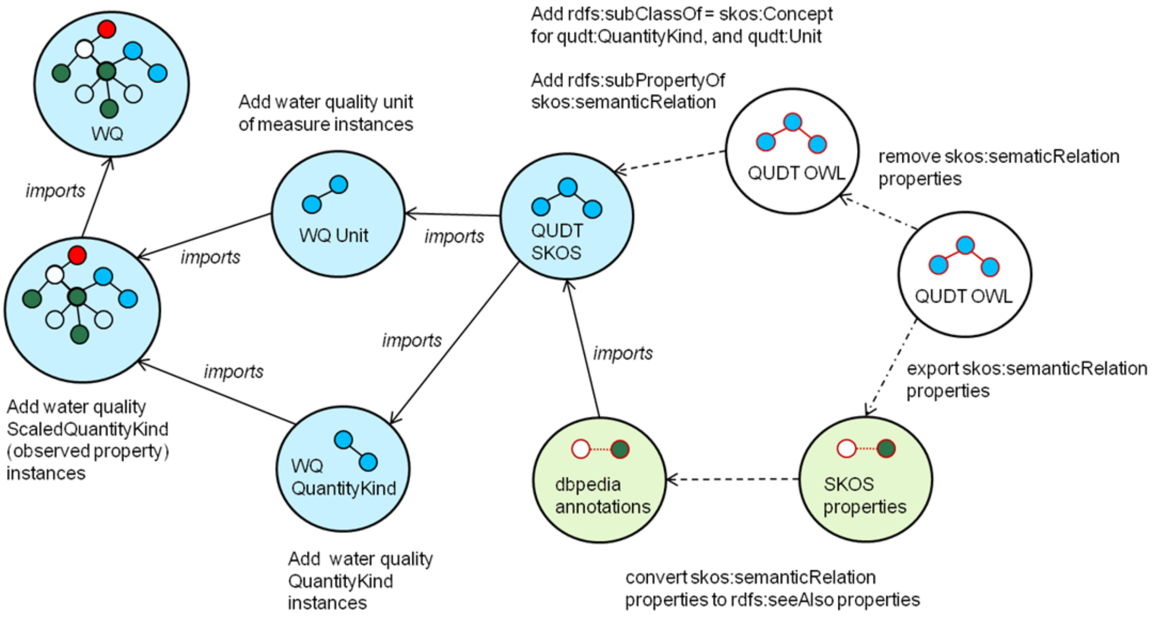

Figure 2 QUDT to water quality 'cleaning' process showing intermediate files.

3.3. Vocabulary concept instances

An initial vocabulary was derived from the Healthy Headwaters database (Ryan et al, 2011; http://www.nrm.qld.gov.au/water/health/healthy-headwaters/), supplemented by the CUAHSI Hydrologic Information System (Zaslavsky et al., 2007) and the Australian National Groundwater Data Transfer Standard (ANGDTS; Bureau of Rural Sciences, 1999). The vocabulary of identified objects associated with these observed properties was developed and informally mapped to concepts in the ANGDTS, the Water Data Transfer Format (WDTF; http://www.bom.gov.au/water/standards/wdtf/), the Australian Drinking Water Guidelines (NHMRC, NRMMC, 2011), the International Union of Pure and Applied Chemistry (IUPAC, http://www.iupac.org/), and the Chemical Abstracts Service (CAS; American Chemical Society, 2013). These vocabularies encompass several domains, such as chemistry, biology, and hydrology.

Most of the identified objects are chemical entities. It was therefore important to align these with chemistry definitions, which we took from the ChEBI ontology (Degtyarenko et al. 2008, Hastings et al. 2013). However, chemical entities and substances in ChEBI are defined as OWL Classes. As indicated above, this presents a challenge as our observable property model includes properties which are defined to apply at the level of instances or 'individuals', rather than classes. In order to manage this, the external ChEBI owl:Classes were mirrored as skos:Concepts, in a new namespace, using the following method:

- Base URI changed from the ChEBI namespace to the water quality namespace;
- *owlClass* converted to *skos:Concept;*
- *rdfs:subClassOf* converted to *skos:broader;*
- *obo2:Definition* converted to *skos:definition;*
- *rdfs:label* converted to *skos:prefLabel;*
- *prov:hadPrimarySource* (http://www.w3.org/TR/prov-o/) added to provide a reference to the URI of the original ChEBI *owl:Class*.

Other properties contained in the ChEBI OWL ontology, such as *obo2:Synonym*, were not transformed. This minimised duplication of data, with the *prov:hadPrimarySource* property providing a mechanism to navigate to the original source material to determine these properties.

Each of the water quality *wqq:IdentifiedObject skos:Concepts* was manually compared with the ChEBI *skos:Concepts*. Where a match occurred the *skos:exactMatch* property was used to map between the two concepts.

The resultant water quality ontology contains 396 instances of *wqop:ScaledQuantityKind* concepts, 332 *wqop:SubstanceOrTaxon* concepts linked to ChEBI concepts where appropriate, 1525 *qudt:Unit* concepts, and 253 *qudt:QuantityKinds*.

3.4. Observations and Measurements property-type model

The ScaledQuantityKind model for observable properties can be mapped to a model for Property Types described in Observations and Measurements v1.0 (Cox, 2007). The key concept from O&M for our purposes is the class **ConstrainedPropertyType.** This specializes a generic **PropertyType** by modifying a **base** with the addition of constraints. Each constraint typically specifies a value on some secondary axis (Figure 3).

In the implementation described here, 'base' is mapped to qudt:generalization, and 'singleConstraint' is implemented as wqop:constraint, with a single subProperty wqop:objectOfInterest. Future elaboration of the model may see additional subProperties of wqop:constraint.

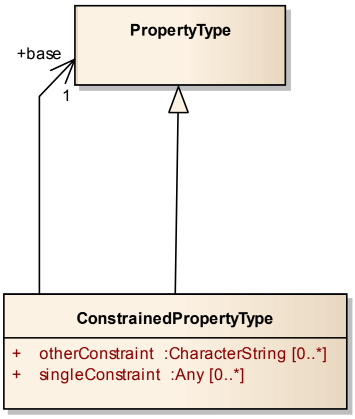
PropertyType
+base
1
ConstrainedPropertyType
+ otherConstraint :CharacterString [0..*]
+ singleConstraint :Any [0..*]

Figure 3 OM v1 model for derived property-type definitions

4. DISCUSSION AND CONCLUSIONS

Our water quality observable properties ontology is aligned with the previously-established Observation and Measurements and QUDT conceptual models, and linked to definitions in ChEBI and DBPedia. This provides a level of integration not readily achieved by stand-alone vocabularies. This work was particularly focussed on re-use of the QUDT Unit and QuantityKind concepts and instances. The ontology may be further improved by including the QUDT systems and dimension components to support unit of measure conversions.

Previous water quality vocabularies consist of terms that conflate separate concepts, such as the observed property, the observation method, the units of measure and the object of interest. Separating these concepts makes them easier to query and map to definitions in other domains, thus facilitating greater semantic interoperability.

Even when mapping concepts between ontologies is straight forward, there are a number of issues associated with adopting or making ontological commitments to other domain vocabularies and ontologies. By doing so, great care must be taken to examine the ontological commitments that these domain vocabularies and ontologies make to other ontologies, such as upper ontologies, as they bear implications for the formalisms adopted in the community vocabularies/domains.

There are situations where the representation language used to encode the respective vocabularies and ontologies may impede the ability to reference one to the other. The requirement to represent an OWL ontology as a SKOS vocabulary poses several challenges. Firstly, there is the DL-safe issue, where reasoning capabilities can be limited or inaccurate by naively mapping one to the other. Secondly, there is an issue of a mismatch in semantic granularity, i.e. are we able to translate SKOS Concepts represented in a community vocabulary equivalent to an OWL class? Lastly, there may be constraints on publishing the vocabularies or ontologies in a particular representation language, such as SKOS vocabularies.

Therefore, there is a two-fold requirement to:

a. Curate a mapping of local vocabularies to standard vocabularies and ontologies;
b. Ensure practical publishing requirements are satisfied.

The methodology outlined here for mapping between OWL and SKOS ontologies while maintaining separate governance, might be used for aligning other standard vocabularies and ontologies.

Aligning the water quality vocabularies with the chemistry domain via the Chemical Entities of Biological Interest (ChEBI) ontology, provided access to definitions, synonyms, relationships and properties created as a result of considerable domain expertise input. Linking to the ontology, rather than replicating it, ensures that governance and currency concerns are addressed.

Establishing a SKOS based vocabulary service enables users to query and navigate through the concepts. More importantly, it allows applications to use the service to establish controlled lists in user interfaces and provide improved query functionality on appropriately configured data provider services.

Further work is required to map the concepts presented in this ontology to similar concepts in other appropriately governed and related ontologies, along with incorporating the QUDT 'dimension' concepts.

Extending the ontology to include measurement procedures, the host medium and the related feature of interest is also desirable.

ACKNOWLEDGMENTS

The authors wish to thank Eloise Nation, Jill McNamarra and Peter Kennedy at the Australian Bureau of Meteorology for providing the initial Healthy Headwaters excel spreadsheet data and reviewing the water quality ontology. The review of this paper by Laurent Lefort (CSIRO) improved it considerably and is greatly appreciated.

REFERENCES

American Chemical Society (2013). CAS Chemical Abstracts Service. American Chemical Society. Online available: https://www.cas.org/; last accessed 06/2013.

Berrueta, D. and L. Polo (2009). Measurement Units Ontology. Online available: http://forge.morfeo-project.org/wiki_en/index.php/Units_of_measurement_ontology; last accessed 07/2013.

Bureau of Rural Sciences (1999). The Australian National Groundwater Data Transfer Standard. Online available: http://www.brs.gov.au/land&water/groundwater/; last accessed 07/2013.

Cox, S. J. D. (ed.) (2007). Observations and Measurements – Part 1 - Observation schema. OGC 07-022r1, Open Geospatial Consortium Inc. Online available: http://www.opengeospatial.org/standards/om; last accessed 06/2012.

de Koning, H.P., Rouquette, N., Burkhart, R., Espinoza, H., and Lefort, L. (2009). Library for Quantity Kinds and Units: schema, based on QUDV model OMG SysML(TM), Version 1.2. Online available: http://www.w3.org/2005/Incubator/ssn/ssnx/qu/qu; last accessed 09/2013.

Degtyarenko, K., P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj and M. Ashburner (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36**, D344–D350.

Hastings, J., P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams and C. Steinbeck (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*

Hodgson R. and P.J. Keller (2011). QUDT: Quantities, Units, Dimensions and Data Types in OWL and XML. Online available: http://www.qudt.org/; last accessed 07/2013.

Moran, B. (2012). Phase 1 Information Platform to support bioregional assessment of Australia's major coal basins. Project Management Plan. Australian Government, Bureau of Meteorology. 28 July 2012, Version 4.

NHMRC, NRMMC (2011). *Australian Drinking Water Guidelines Paper 6 National Water Quality Management Strategy*. National Health and Medical Research Council, National Resource Management Ministerial Council, Commonwealth of Australia, Canberra.

Rijgersberg, H., M. van Assem and J. Top (2013). "Ontology of units of measure and related concepts." *Semantic Web* 4.1 (2013): 3-13. Online available: http://www.wurvoc.org/vocabularies/om-1.8/; last accessed 07/2013.

Ryan, J., Rodrigues, K. and De Hayr, R. (2011). National Information Management Protocols for Water Quality Monitoring. Report A Water Quality Metadata Guidelines. State of Queensland, Department of Environment and Resource Management. Online available: http://www.bom.gov.au/water/standards/projects/waterqlty.shtml; last accessed 06/2013.

Schadow, G. and McDonald, C.J. (2009). The Unified Code for Units of Measure. Regenstrief Institute, Inc. and the UCUM Organization, Online available: http://unitsofmeasure.org/ucum.html; last accessed 07/2013.

Worley Parsons, (2012). Description of data tables in the integrated hydrochemistry database developed for Activity 1.2 of the Healthy Headwaters Coal Seam Gas Water Feasibility Study. State of Queensland, Department of Environment and Resource Management.

Zaslavsky, I., D. Valentine and T. Whiteaker (eds) (2007). CUAHSI WaterML. OGC 07-041r1. Open Geospatial Consortium Inc.