International Conference on Hydroinformatics

8-1-2014

Enhancing Water Quality Data Service Discovery And Access Using Standard Vocabularies

Jonathan Yu

Bruce A. Simons

Nicholas J. Car

Simon J.D. Cox

Recommended Citation

Yu, Jonathan; Simons, Bruce A.; Car, Nicholas J.; and Cox, Simon J.D., "Enhancing Water Quality Data Service Discovery And Access Using Standard Vocabularies" (2014). *CUNY Academic Works.*
http://academicworks.cuny.edu/cc_conf_hic/165

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

ENHANCING WATER QUALITY DATA SERVICE DISCOVERY AND ACCESS USING STANDARD VOCABULARIES

J. YU (1), B.A. SIMONS (1), N. CAR (2) & S.J.D. COX (1)
(1) Environmental Information Systems, CSIRO, Highett, Australia
(2) Environmental Information Systems, CSIRO, Brisbane, Australia

There is a growing need for increased integration across the publication, discovery, access and use of scientific datasets, including water related datasets. Scientific datasets have varying formats and are published using a variety of methods, ranging from physical media to sophisticated web service interfaces. The Network Common Data Form (NetCDF) is both a software package and a data format for producing array-oriented scientific data, which is commonly used to exchange data, including water quality data. NetCDF datasets are published through service interfaces using the THREDDS data server. Alternatively, water quality datasets can be encoded with standard XML formats such as WaterML 2.0, which can be published with the Open Geospatial Consortium (OGC) community's Web Feature Service interface standard (WFS) and Sensor Observation Service standard (SOS). However, appropriate interpretation of the content, discovery and interoperability of data depend on common models, schemas and vocabularies. We have developed a water quality vocabulary which is encoded and published using semantic web technologies. We present a general approach to enhance existing metadata that accompany scientific datasets using the water quality vocabulary. This approach is a first step in a broader approach to enabling better integrated discovery, integration and access to existing scientific datasets using standard vocabularies that are encoded and published using semantic technologies.

1. INTRODUCTION

More than ever before, an unprecedented number of scientific data repositories and services are being published and made accessible. Vast amounts of historic and real-time observations are being made by remote and in-situ sensors as well as modelled simulations in increasingly finer-grained temporal and spatial resolutions. It is anticipated that such increased capacity to capture earth observations “will deepen our understanding of natural phenomena” [20]. However, with the increase of complex and heterogeneous data services, there are increased discovery, access, understanding and integration challenges.

Multiple service interfaces, data formats, and methods of publishing data exist. In a recent survey of 100 scientific data repositories (SDRs) available online, Marcial and Hemminger found a wide variety of publication methods used by SDRs ranging from publishing data as PDF documents, where the data is presented as a report, to publishing data through standards-based web services, where the data may be queried using standard interfaces [17].

The need to conduct research across scientific domains, communities and disciplines is increasing. Supporting multidisciplinary research requires increased integration and interoperability between services and datasets, which is critical for earth system science and earth observation [20]. An example of this is the eReefs project, which seeks to establish an interoperable coastal information platform for the conservation of the Great Barrier Reef [5, 7, 6]. The eReefs information platform aims to enable integration of a number of datasets across scientific repositories in separate agencies working across multiple domains. A core requirement for interoperable systems, such as, the eReefs information platform, is that datasets curated and managed by the different organisations can be used together in a consistent and integrated way. Thus, streamlining the process of discovery, access, integration and use of the data to provide researchers and client applications with relevant and integrated data with little or no manual intervention is crucial.

However, data providers and users of scientific datasets adopt service interfaces and data formats due to a number of factors such as maturity of tooling, popularity, community expectations, and available client applications to consume it. An example is the NetCDF format [22, 21] which has become a de-facto standard for model outputs in many scientific communities including the climate and atmospheric community. This is due to available supporting tools, a simple data model and efficient data transfer. Another example in the water domain is the WaterML 2.0 XML data exchange standard for time series of hydrological observation data [24, 25]. Both NetCDF and WaterML 2.0 are used for publishing data using standard service interfaces, e.g. NetCDF using OpenDAP [9] and WaterML 2.0 using the OGC's Sensor Observation Service (SOS) [3]. For eReefs, delivery of water quality data is crucial, and both NetCDF and WaterML 2.0 formats are being used for publishing.

Although there are standardized data formats and delivery services, there is still the issue of resolving semantic heterogeneity. Resolving semantic heterogeneity is a significant barrier to seamlessly integrating multiple data sources [1]. Semantic technologies and Linked Data approaches specify best practices for publishing, linking and sharing structured data using existing web technologies. Linked Data approaches are based on HTTP URIs and the use of web standards such as Resource Description Framework (RDF) to define semantics and allowing machine-readable representations of knowledge and data [2]. The use of HTTP URIs provides a web-resolvable identifier to descriptions of relevant information, such as data. Languages such as the Web Ontology Language (OWL) [18] and the Simple Knowledge Organization Scheme (SKOS) [19] capture domain semantics as machine readable descriptions.

In this paper, we consider how existing data services can be enhanced with consistent semantics to allow data to be easily discovered, accessed, integrated and analysed. We propose approaches for utilising standard vocabularies to embed semantics identified with web-resolvable URIs to link vocabulary definitions of observable properties, objects of interest, unit of measure, feature of interest terms for a given dataset. We show how existing data services publishing publishing NetCDF and WaterML 2.0 can be enhanced with links to the water quality vocabulary.

2. THE WATER QUALITY VOCABULARY

A water quality vocabulary that harmonises a number of water quality terms and aligns with existing ontologies, is formalized using RDF, and has been published as Linked Data [23]. The vocabulary used terms that were extracted from existing water quality vocabularies in the Australian context. It aligns both the chemical substances with an existing ontology (CHEBI [13]), as well as the NASA/TopQuadrant ontology for Quantities, Units, Dimensions and Types (QUDT) [14]. It contains vocabulary definitions of observable properties, objects of interest, unit of measure, and feature of interest terms for the water quality domain. Publishing the vocabulary as Linked Data provides semantic interoperability using unique and resolvable HTTP URIs as identifiers for the vocabulary terms. This allows any web client to resolve a term to its vocabulary definition via its HTTP URI. It also allows an ontology to be defined which precisely captures the semantics of domain and the domain-independent classes and relationships between them in a machine readable format. The semantic technologies also allow import statements between ontologies, thus promoting reuse of existing definitions.

The Observable Property ontology[1] is the basis for which the water quality vocabulary has been defined (Figure 1). The Observable Property ontology extends the QUDT ontology [14] for Unit and Quantity Kind classes, and is aligned with the CHEBI ontology [13] for definitions of objects related to chemical substances. Figure 2 illustrates an example from the water quality vocabulary, where '*dissolved nitrogen concentration'* has been defined with the objectOfInterest relationship to the appropriate nitrogen definition from the CHEBI ontology; the general quantity kind, '*nitrogen concentration*'; and the appropriate units of measure which are related, '*MilligramsPerLitre*' or '*MolePercent*'. Thus, we are able to capture the semantics of the specific domain and refer to each definition via a resolvable HTTP URI rather than just a string label. The vocabulary terms are identified with the http://environment.data.gov.au/water/quality/def/ namespace (base URI).

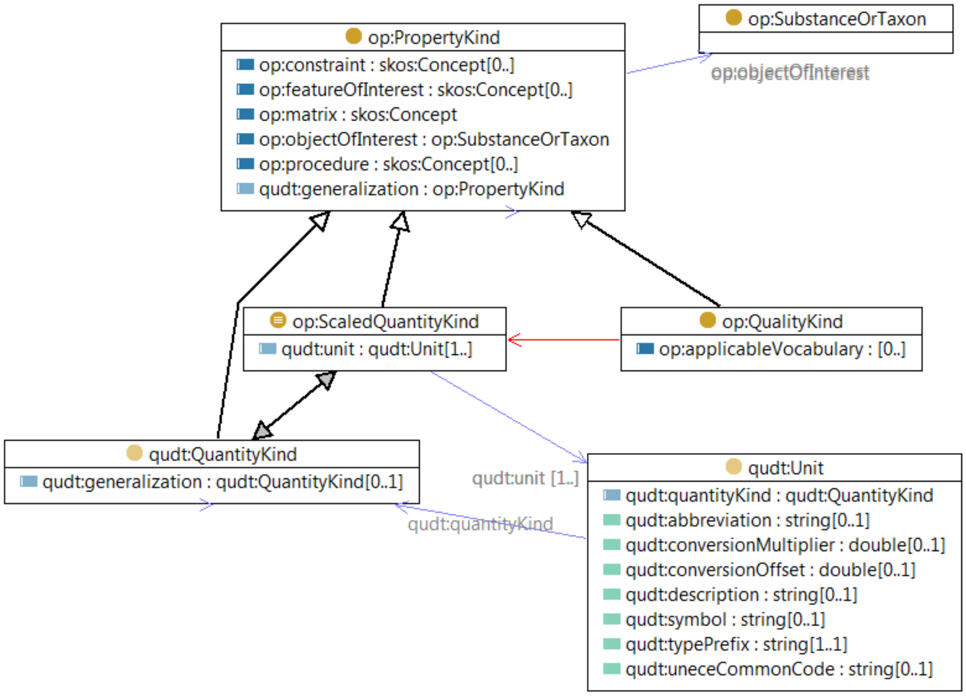
op:PropertyKind
op:constraint : skos:Concept[0..]
op:featureOfInterest : skos:Concept[0..]
op:matrix : skos:Concept
op:objectOfInterest : op:SubstanceOrTaxon
op:procedure : skos:Concept[0..]
qudt:generalization : op:PropertyKind
op:SubstanceOrTaxon
op:objectOfInterest
op:ScaledQuantityKind
qudt:unit : qudt:Unit[1..]
op:QualityKind
op:applicableVocabulary : [0..]
qudt:QuantityKind
qudt:generalization : qudt:QuantityKind[0..1]
qudt:unit [1..]
qudt:quantityKind
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]

Figure 1. Core classes in the Observable Property Ontology

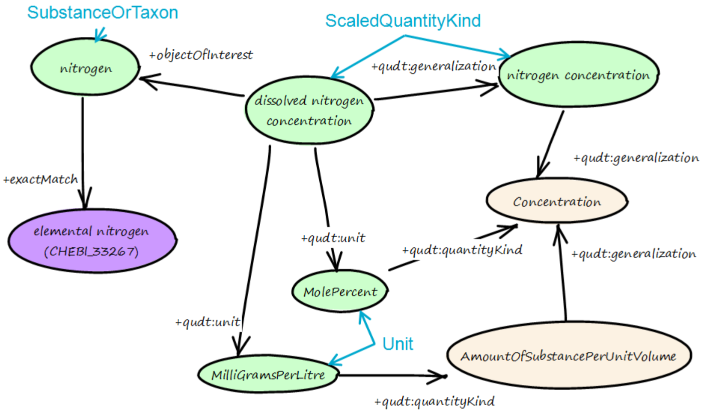

Figure 2. Example water quality vocabulary definitions

[1] Available: http://environment.data.gov.au/water/quality/def/op

As the key elements are defined to be sub-classes or sub-properties of SKOS elements, a SKOS view can be published through standard vocabulary APIs, alongside the full view. The ontology is published via the using the SISS Vocabulary service (SISSVoc) [12] and can be accessed via the SPARQL endpoint [2] or the SISSVoc search client[3].

3. LINKING VOCABULARIES

Although NetCDF is an efficient and well-used data exchange format with established client software, it has a number of limitations. One that severely restricts interoperability is the use of 'tokens' for standard terms. A number of conventions, such as CF-climate conventions [11], specify what tokens to use to denote units of measure and observed parameters. Figure 3 gives an example of a NetCDF metadata header variable "Nap_MIM" defined with the term 'total_suspended_solids', with units of measure "mg/L". What is lacking is the ability to resolve relevant NetCDF tokens into meaningful vocabulary terms.

```
float Nap_MIM(time, latitude, longitude) ;
   Nap_MIM:_FillValue = -999.f ;
   Nap_MIM:long_name = "TSS, MIM SVDC on Rrs" ;
   Nap_MIM:units = "mg/L" ;

       Nap_MIM:valid_min = 0.01209607f ;

       Nap_MIM:valid_max = 226.9626f ;

   Nap_MIM:standard_name = "total_suspended_solids";
```

Figure 3. NetCDF metadata example

In contrast, an example vocabulary standard term in the Water Quality Vocabulary has the URI, http://environment.data.gov.au/water/quality/def/property/solids-total_suspended, which resolves to a semantic description of a total suspended solids concept with the label, "*total suspended solids*". The use of HTTP URIs as values within the metadata headers allows for resolution using common web technologies to richer information, such as definitions, foreign language equivalent labels, and other semantic relationships, via vocabulary services.

We propose an approach whereby the binding from a dataset will be achieved by including appropriate metadata tags to embed the URI for the vocabulary term. For applications in the water quality domain, values will come from the instances of the respective water quality classes. Additional values, such as those derived from the Procedure class, will be included to indicate the method used to determine the result, and Matrix to indicate the conducting medium of the parameter.

The values included in each of these will be URIs from appropriate Water Quality vocabularies. Where the parameter being delivered is a composite parameter (for example "Kd_490_MIM = Attenuation coefficient with depth at 490nm wavelength using the MIM SVDC method", "EpiTN = Total N in epibenthos"), all appropriate URIs are to be included. The list of URIs for the respective classes is listed below, for which the instances of each of the classes are used in the metadata fields linking to the vocabularies:

- SubstanceOrTaxon: http://environment.data.gov.au/water/quality/def/op#SubstanceOrTaxon
- PropertyKind: http://environment.data.gov.au/water/quality/def/op#PropertyKind
  - ScaledQuantityKind: http://environment.data.gov.au/water/quality/def/op#ScaledQuantityKind (where the quantity kind has a unit of measure associated with it) QuantityKind: http://qudt.org/schema/qudt#QuantityKind is a QUDT equivalent class to ScaledQuantityKind without the qudt:unit property.

[2] http://sissvoc.ereefs.info/ereefs/sparql
[3] http://sissvoc.ereefs.info/search with a service URI of http://sissvoc.ereefs.info/sissvoc/ereefs

   - PropertyKind where the quantity kind does not necessarily have a unit of measure associated with it, (e.g. categories of things, categoricals)
- Unit:http://qudt.org/schema/qudt#Unit or http://environment.data.gov.au/water/quality/def/unit/
- Procedure: http://data.ereefs.org.au/def/Procedure

4. ANNOTATING DATASET METADATA WITH VOCABULARIES

In this section, we propose conventions for NetCDF and WaterML 2.0 data services which allow the appropriate metadata to be annotated with HTTP URIs linking to the water quality vocabulary.

4.1. THREDDS/NetCDF

For NetCDF datasets, typically scaled values are encoded, that is, each value has a unit of measure or some scale associated with it. Thus, the ScaledQuantityKind concept is used for the PropertyKind metadata. The parameter terms scaledQuantityKind_id, unit_id, substanceOrTaxon_id, procedure_id and medium_id should also be HTTP URIs to appropriate vocabulary terms. However, NetCDF-4 does not currently allow URIs as parameter terms.

- <parameter>:scaledQuantityKind_id = <URI for an appropriate ScaledQuantityKind Concept>
- <parameter>:unit_id= <URI for Unit concept>
- <parameter>:substanceOrTaxon_id= <URI for SubstanceOrTaxon concept>
- <parameter>:procedure_id= <URI for Procedure concept>
- <parameter>:medium_id=<URI for the Medium concept>

An example NetCDF response with the vocabulary terms encoded is shown in Figure 4.

```
float Nap_MIM(time, latitude, longitude) ;
  Nap_MIM:_FillValue = -999.f ;
  Nap_MIM:long_name = "TSS, MIM SVDC on Rrs" ;

        Nap_MIM:units = "mg/L" ;

        Nap_MIM:valid_min = 0.01209607f ;

        Nap_MIM:valid_max = 226.9626f ;

        Nap_MIM:scaledQuantityKind_id
               = "http://environment.data.gov.au/water/quality/def/property/solids-total_suspended" ;

  Nap_MIM:unit_id = "http://environment.data.gov.au/water/quality/def/unit/MilliGramsPerLitre" ;
  Nap_MIM:substanceOrTaxon_id = "http://environment.data.gov.au/water/quality/def/object/solids";
  Nap_MIM:medium_id = "http://environment.data.gov.au/water/quality/def/object/ocean"
  Nap_MIM:procedure_id = "http://data.ereefs.org.au/ocean-colour/MIM_SVDC_RRS" ;
```

Figure 4. Example annotated NetCDF metadata header using Water Quality Vocabulary URIs

4.2. SOS / WaterML 2.0

WaterML 2.0 datasets can contain either scaled values or categories. Thus, the PropertyKind concept is used. The metadata fields that are populated are:

- om:procedure - <URI for Procedure concept>
- om:observedProperty - <URI for an appropriate PropertyKind Concept>
- om:featureOfInterest - <URI for SubstanceOrTaxon concept>
- wml2:interpolationType - <URI for WML2 InterpolationType concept>
- wml2:uom - <URI for Unit concept>

Figure 5 is an example excerpt of a WaterML 2.0 encoded SOS response.

```
<wml2:Collection xsi:schemaLocation="http://www.opengis.net/waterml/2.0
http://www.opengis.net/waterml/2.0/waterml2.xsd" gml:id="sample.1">
...
    <wml2:observationMember>
        <om:OM_Observation gml:id="sample.obs.1">
...
            <om:procedure
xlink:href="http://data.ereefs.org.au/procedure/insitu-wq-sensing"
xlink:title="example procedure"/>
            <om:observedProperty
xlink:href="http://environment.data.gov.au/water/quality/def/property/solids-
total_suspended" xlink:title="TSS"/>
            <om:featureOfInterest
xlink:href="http://environment.data.gov.au/water/quality/def/object/water"
xlink:title="water body"/>

<om:result>
    <wml2:MeasurementTimeseries gml:id="sample.Ts.1">
...
    <wml2:uom
xlink:href="http://environment.data.gov.au/water/quality/def/unit/MilliGramsPerL
itre"/>
...
      <!-- the data -->
      <wml2:point>
         <wml2:MeasurementTVP>
            <wml2:time>1990-09-01T00:00:00.000+01:00</wml2:time>
            <wml2:value>193.0</wml2:value>
         </wml2:MeasurementTVP>
      </wml2:point>
      <wml2:point>...</wml2:point>
    </wml2:MeasurementTimeseries>
 </om:result> </om:OM_Observation> </wml2:observationMember> </wml2:Collection>
```

Figure 5. Example annotated WaterML 2.0 using Water Quality Vocabulary URIs

5. DISCUSSION AND RELATED WORK

Using the water quality vocabulary and configuring data services as proposed in this paper allows for datasets to be discovered across heterogeneous services, in the case shown in the previous section, across NetCDF and WaterML2.0 data services. Using the URI for the vocabulary definition of the Total Suspended Solids property kind, http://environment.data.gov.au/water/quality/def/property/solids-total_suspended, we can configure NetCDF and WaterML 2.0 clients to filter results based on the respective fields, e.g. <parameter>:scaledQuantityKind_id for NetCDF data services and om:observedProperty for WaterML 2.0 data services. Other queries such as query datasets by feature of interest, procedure and other attributes are also possible with our proposed approach.

Vocabulary services provide the identity and resolution point for definitions. They allow the use of validation to complement traditional XML schema validation with content validation [27, 28]. The AuScope Discovery portal[4] uses vocabulary services to populate user interface elements to capture queries for matching any preferred and alternative labels for earth resource concepts, which in turn allows for aggregated queries to be made on earth resources datasets collected by individual Australian jurisdictions [26]. Ma et al. present the use of ontologies to assist in annotation and visualisation of geological time-scale information in mapping applications [16]. The NETMAR semantic web service provides client applications with discovery and service chaining capabilities using vocabulary and ontology definitions [15].

[4] http://portal.auscope.org

An alternative approach is to create annotation documents or descriptions external from the scientific dataset or documents themselves. Ciccarese et al. presents an annotation ontology that can be used for creating annotation documents to link scientific documents and other web resources with semantic descriptions stored as RDF/OWL [8]. Cao et al. propose approaches for annotating semantics using an OWL ontology for describing observational data [4]. The advantage of these approaches is that queries can then be issued about the datasets from the annotations, e.g. using SPARQL. However, these approaches do not enhance any dataset metadata to be more self-describing as our approach proposes, that is, the metadata themselves do not provide a means for linking to vocabulary terms.

6. CONCLUSION

In this paper, we have proposed methods for improving access and discovery of growing scientific research datasets in earth system science and earth observations. We have considered how existing datasets can be enhanced with consistent semantics to allow data to be easily discovered, accessed, integrated and analysed. Our approach uses semantic web technologies for defining and publishing standard vocabularies, particularly in the water quality domain. We have shown how existing data services publishing NetCDF and WaterML 2.0 can be enhanced with links to the water quality vocabulary using a Linked-data approach and standard vocabularies. Our methodology for embedding web-resolvable URIs which links to vocabulary definitions of observable properties, objects of interest, unit of measure, feature of interest terms for a given dataset.

Acknowledgments

The authors would like to acknowledge Edward King, Vittorio Brando and Farhan Rizwi at CSIRO for their domain expertise and contributions to this paper.

REFERENCES

[1] B. Beran and M. Piasecki. Engineering new paths to water data. *Computers & Geosciences*, 35(4):753–760, 2009.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5:1–22, 2009.

[3] A. Bröring, C. Stasch, & J. Echterhoff. OGC Sensor observation service interface standard. OGC 2012.

[4] H. Cao, S. Bowers, and M.P. Schildhauer. Approaches for semantically annotating and discovering scientific observational data. In Proc. *Database and Expert Systems Applications*, *LNCS 6860*, pp526–541. Springer Berlin Heidelberg, 2011.

[5] N.J. Car. The eReefs information architecture. In *Proc 20th International Congress on Modelling and Simulation*, pages 831–837, Adelaide, Australia, December 2013. MSSANZ.

[6] N.J. Car, P.G. Fitch, and D. Lemon. Scoping study: ereefs work package 2 – interoperable data and information systems. Technical report, CSIRO, 2012. ISBN 978-1-922173-07-2 .

[7] N.J. Car and J. Hodge. eReefs : distributed data, a unified picture. In *7th eResearch Australasia Conference*, pages 1–4, Brisbane, Australia, October 2013. eResearch 2013 Conference Secretariat.

[8] P. Ciccarese, M. Ocana, L.J. Garcia Castro, S. Das, and T. Clark. An open annotation ontology for science on web 3.0. *J Biomed Semantics*, 2(Suppl 2):S4, 2011.

[9] P. Cornillon and T. Gallagher, J.and Sgouros. Opendap: Accessing data in a distributed, heterogeneous environment. *Data Science Journal*, 2(0):164–174, 2003.

[10] S. Cox. Observations and measurements. *OGC Best Practices Document. OGC*, 2006.

[11] B. Eaton, J. Gregory, B. Drach, K. Taylor, and S. Hankin. Netcdf climate and forecast (cf) metadata conventions, Dec 2011. http://cf-pcmdi.llnl.gov/documents/cf-conventions/1.6/cf-conventions.html (Accessed Jan 2014).

[12] J. Githaiga, G. Duclaux, S. Cox, and J. Yu. Spatial information services stack (SISS) vocabulary service. In *Proc. eResearch Australasia*, 2010.

[13] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck. The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, 41(D1):D456–D463, 2013.

[14] R. Hodgson and P.J. Keller. Qudt: Quantities, units, dimensions and data types in owl and xml. http://www.qudt.org (Accessed Jan 2014).
[15] A.M. Leadbetter and D.O. Lowry, R.K.and Clements. Putting meaning into netmar–the open service network for marine environmental data. *International Journal of Digital Earth*, pp. 1–18, 2013.
[16] X. Ma, E.J.M. Carranza, C. Wu, and F.D. van der Meer. Ontology-aided annotation, visualization, and generalization of geological time-scale information from online geological map services. *Computers & Geosciences*, 40:107–119, 2012.
[17] L.H Marcial and B.M. Hemminger. Scientific data repositories on the web: An initial survey. *Journal of the American Society for Information Science and Technology*, 61(10):2029–2048, 2010.
[18] D.L. Mcguinness and F. van Harmelen. OWL web ontology language overview. W3C recommendation, W3C, February 2004.
[19] A. Miles and J.R. Pérez-Agüera. Skos: Simple knowledge organisation for the web. *Cataloging & Classification Quarterly*, 43(3):69–83, 2007. Accessed July 2011.
[20] S. Nativi and L. Bigagli. Discovery, mediation, and access services for earth observation data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE* 2(4):233–240, 2009.
[21] R Rew, E Hartnett, J Caron, et al. Netcdf-4: Software implementing an enhanced data model for the geosciences. In *22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanograph, and Hydrology*, 2006.
[22] R. Rew and Glenn Davis. Netcdf: an interface for scientific data access. *Computer Graphics and Applications, IEEE*, 10(4):76–82, 1990.
[23] B.A Simons, J. Yu, and S.J.D. Cox. Defining a water quality vocabulary using qudt and chebi. In *Proc. 20th Intl. Congress on Modelling and Simulation.*, pp 2548– 2554. MSSANZ, Dec 2013.
[24] P. Taylor. OGC WaterML 2.0: Part 1 - timeseries,. OGC Implemented Standard 10-126r3, 2012.
[25] P. Taylor, S. Cox, G. Walker, D. Valentine, and P. Sheahan. WaterML2. 0: development of an open standard for hydrological time-series data exchange. 2013.
[26] R. Woodcock, B. Simons, G. Duclaux, and S. Cox. Auscope's use of standards to deliver earth resource data. In *EGU General Assembly Conference Abstracts*, volume 12, page 1556, 2010.
[27] J. Yu, S. Cox, G. Walker, P.J. Box, and P.A. Sheahan. Use of standard vocabulary services in validation of water resources data described in xml. *Earth Science Informatics*, 4(3):125–137, 2011.
[28] J. Yu, P. Taylor, S. Cox, and G. Walker. Towards validating observational data in WaterML 2.0. In *Proc. 10th Intl. Conf. Hydroinformatics*, Hamburg, Germany, 2012. IWA Publishing.