19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

A Knowledge Model for Bridging Semantic Gaps between Multiple Water Information Sources

Y. Shu[a], M. Compton[b], G. Squire[b], D. Ratcliffe[b] and G. Walker[b]

[a]*Tasmanian ICT Centre, CSIRO, GPO BOX 1538, Hobart, TAS 7001, Australia*
[b]*ICT Centre, CSIRO, GPO BOX 664, Canberra, ACT 2601, Australia*
Email: yanfeng.shu@csiro.au

Abstract:

In response to the increasing demand in Australia for improving the efficiency of water management practices, the Bureau of Meteorology (BOM) has been given a mandate to build and maintain an integrated national water information system. Over 240 water organisations are required to provide data. These organisations use a wide range of systems and data formats. To ensure robust and reliable data delivery, the Bureau established the Water Data Transfer Format (WDTF) as a standard format for data transfer. Meanwhile, the Water Regulations 2008 were enacted to specify the water information required from organisations. As a result, data from organisations have to be translated into WDTF; also, the translated data have to comply with WDTF and the Regulations.

To facilitate these two tasks, in previous work, we proposed using knowledge models to capture semantic gaps between data, WDTF, and the Regulations. In this paper, we describe in detail how to construct such a model. In particular, we focus on what concepts and constraints to be covered in the model, and how to represent them.

We first summarise what we perceive as semantic gaps between data provided by organisations, and information requirements as expressed in WDTF or the Regulations:

- Various terminologies and units may be used by organisations to describe data.
- Data can be interpreted in various ways without adequate contextual information.
- Information requirements are expressed at different levels of granularity and detail.

Based on this, we then define model concepts and constraints, and represent them using the Web Ontology Language (OWL) with a modular approach. We choose OWL for its expressivity and reasoning support, which enables us not only to provide a precise translation of data, but also to perform data validation automatically.

Finally, we describe an application scenario, in which the model is used for translating spreadsheets into WDTF. We have tested the application on the data from several organisations. The initial experience with the application is encouraging: with a little help from users, data can be easily translated into WDTF. Our next step is to make the application operational.

Keywords: Water Regulations, Water Data Transfer Format (WDTF), knowledge model, OWL

1 INTRODUCTION

Under the Commonwealth Water Act 2007, the Bureau of Meteorology (BOM) of Australia is tasked with a range of functions which require it to collect, hold, manage, interpret and disseminate national water information. The Water Act provides for the establishment of the Water Regulations[1] to support these functions, which came into effect on 30th June 2008. The Regulations define the requirements for the collection of water information by the Bureau in 10 categories, with each category further defined by subcategories. The Regulations also specify the organisations which must provide specified water information to the Bureau and the time in which they must provide it.

Over 240 organisations are involved. These organisations use a wide range of systems and data formats, as revealed by an online survey conducted by the Bureau in 2008. An examination of data further indicates that various structural and semantic heterogeneities are present. To ensure robust and reliable data delivery, the Bureau established WDTF (the Water Data Transfer Format)[2] (Walker et al., 2009) as the standard for data transfer. WDTF is an XML schema built on the International Standard Organisation's General Feature Model, ISO 19109, and the Open Geospatial Consortium (OGC)'s Observations and Measurements (O&M) model (Cox, 2007a,b).

As a result, data from organisations have to be translated into WDTF; also, the translated data have to comply with WDTF and the Regulations. The current practice is that each organisation is responsible for translating its data, and the Bureau for providing a validation service. The translation work has proven to be nontrivial, especially for small organisations which typically do not have the capability for doing this. To do the translation, one needs to be familiar with WDTF, its structure, elements and related semantics. Further, it is time-consuming and error-prone if the translation is done manually, considering the number of organisations and the volume of data involved.

To address this, in previous work (Shu and et al., 2010a,b), we proposed using knowledge models to capture semantic gaps between data, WDTF, and the Regulations, in a way that facilitates data translation and validation. In this paper, we describe in detail how to construct such a model. In particular, we focus on what concepts and constraints to be covered, and how to represent them with a modular approach. In the rest of the paper, Section 2 gives a review of information sources and semantic gaps, Section 3 describes the model construction process, Section 4 presents an application scenario in which spreadsheets are translated into WDTF via the use of the model, and finally, Section 5 concludes the paper.

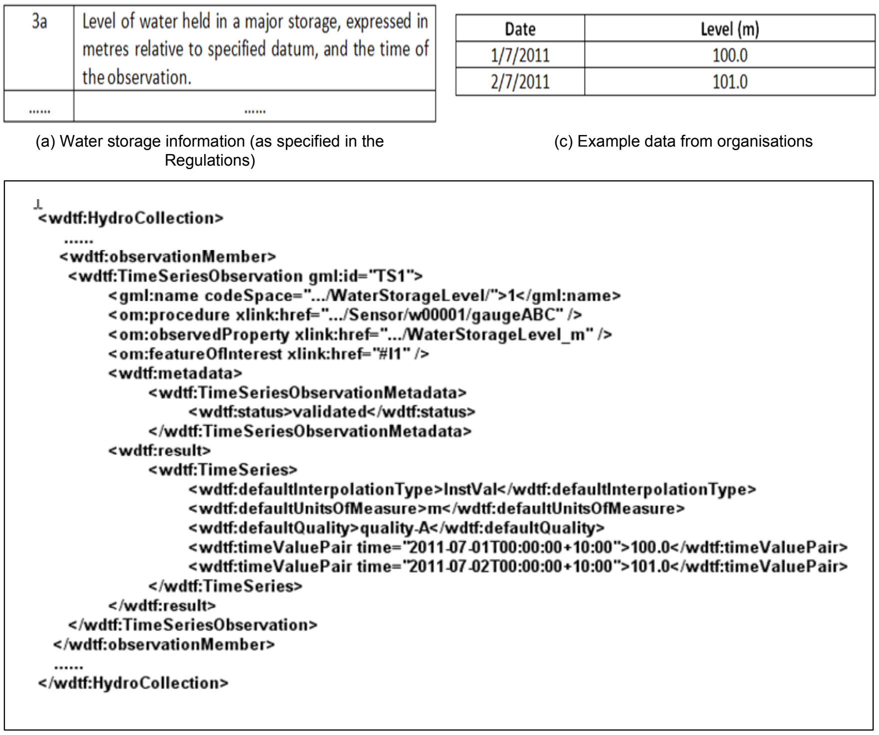
2 INFORMATION SOURCES AND SEMANTIC GAPS

The information sources we consider for the construction of the model include the Water Regulations 2008, WDTF, and data from organisations. The Regulations specify, among others, the water information to be provided by organisations. There are 10 categories of such information, including water course and water storage; each category is further defined by subcategories. A subcategory typically describes a particular type of observation. Figure 1(a) shows a subcategory of water storage information, $3a$, which specifies that water levels of a storage have to be measured in metres; when a measurement is taken, the time of the observation, and the datum used, have to be recorded.

WDTF also specifies the water information to be provided, but at a different level of granularity and detail. It is an XML schema, and its root element is `HydroCollection`. Observations in WDTF are classified into four types: `Measurement` (observations about water specimens), `ComplexObservation` (observations of gaugings for conversions), `TimeSeriesObservation` (Observations of single properties over time), and `GeometryObservation` (surveys of storages and water courses). For each of these observation types, WDTF defines the information to be provided, as well as the constraints on the information. Figure 1(b) shows part of a WDTF instance document, where observations of water level of a storage are encoded as a time-series observation which contains a set of time-value pairs, and the contextual information associated with the set, e.g. the observed property. Compared to the Regulations, information requirements specified by WDTF are in general more detailed and explicit. On the other hand, WDTF only loosely constrains relationships between observations, features, properties and units,

[1] http://www.bom.gov.au/water/regulations/.

[2] http://www.bom.gov.au/water/regulations/wdtf/. There are several versions of WDTF. In this paper, we refer to WDTF v1.0.



3a	Level of water held in a major storage, expressed in metres relative to specified datum, and the time of the observation.
......	......

(a) Water storage information (as specified in the Regulations)

Date	Level (m)
1/7/2011	100.0
2/7/2011	101.0

(c) Example data from organisations

```
<wdtf:HydroCollection>
   ......
   <wdtf:observationMember>
    <wdtf:TimeSeriesObservation gml:id="TS1">
        <gml:name codeSpace=".../WaterStorageLevel/">1</gml:name>
        <om:procedure xlink:href=".../Sensor/w00001/gaugeABC" />
        <om:observedProperty xlink:href=".../WaterStorageLevel_m" />
        <om:featureOfInterest xlink:href="#I1" />
        <wdtf:metadata>
            <wdtf:TimeSeriesObservationMetadata>
                <wdtf:status>validated</wdtf:status>
            </wdtf:TimeSeriesObservationMetadata>
        <wdtf:result>
            <wdtf:TimeSeries>
                <wdtf:defaultInterpolationType>InstVal</wdtf:defaultInterpolationType>
                <wdtf:defaultUnitsOfMeasure>m</wdtf:defaultUnitsOfMeasure>
                <wdtf:defaultQuality>quality-A</wdtf:defaultQuality>
                <wdtf:timeValuePair time="2011-07-01T00:00:00+10:00">100.0</wdtf:timeValuePair>
                <wdtf:timeValuePair time="2011-07-02T00:00:00+10:00">101.0</wdtf:timeValuePair>
            </wdtf:TimeSeries>
        </wdtf:result>
    </wdtf:TimeSeriesObservation>
  </wdtf:observationMember>
  ......
</wdtf:HydroCollection>
```

(b) A WDTF instance document

Figure 1. Information sources.

due to its coarse classification of observations. For example, there are no constraints in WDTF to fully express the ones of 3a in Figure 1(a).

Data from Organisations are provided by typically following the Regulations' classification of water information. There are many different data formats used, including spreadsheets, CSV (Comma-Separated Values) files, and relational tables. One major issue with data is the lack of contextual information. As such, the meaning of data is often not clear, and accordingly data can be interpreted in various ways. For example, Figure 1(c) can be interpreted as observations of water level of a water course or of a water storage. Also, various terminologies may be used to describe data, and various units may be used to express values. See (Shu and et al., 2010a,b) for more details.

To sum up, semantic gaps between data provided by organisations, and information requirements as expressed in WDTF or the Regulations are reflected in the following major aspects:

- Various terminologies and units may be used by organisations to describe data.
- Data can be interpreted in various ways due to lack of contextual information.
- Information requirements are expressed at different levels of granularity and detail.

3 KNOWLEDGE MODEL

To bridge the above gaps, the model needs to have a certain concept and constraint coverage. Also, the model needs to be constructed in a way that facilitates data translation and validation. In this section, we describe the model construction process, which involves three major steps:

Define the concepts to be covered. One primary use of the model in data translation is to act as an intermediate schema between data from organisations and WDTF. As such, the model should be easy for users to understand and map their data to, and on the other hand, easy for the system to map to WDTF. Accordingly, the concepts to be covered should be those being able to facilitate the mapping between data and the model, and the mapping between the model and WDTF. More specifically,

- the concepts that describe the Regulations' classification of water information;
- the concepts of WDTF, including those describing observations and measurements in general, such as `Feature`, `Observation`, `Property` and `Unit`, and those specific to the water domain, such as `WaterCourse`, `WaterStorage`, and `RatingTable`;
- the concepts that describe various units; and
- the concept synonyms and abbreviations that are commonly used in practice by organisations.

Define the constraints to be covered. Besides the concepts, the model also needs to capture the constraints in WDTF and the Regulations, so that the translated data can be ensured to comply with them. Constraints in WDTF are expressed through XML schema restrictions or occurrence indicators, e.g. *each time-series has exactly one observed property*. Constraints in the Regulations, on the other hand, are expressed through textual descriptions, and can only be interpreted by human. For example, from 3a, we know that *water storage levels have to be expressed in metres; also, datum information and observation times have to be provided.* The model should cover all the WDTF and Regulations constraints which are related to the concepts decided earlier. When multiple constraints are defined on the same concepts, only the most restrictive one is considered. For example, for the constraints on the relationship between units and properties, we consider the ones defined in the Regulations, which specify one particular unit for one particular property, instead of the ones in WDTF, which basically allow a set of units to be used for a property.

Model concepts and constraints using OWL. With the concepts and constraints defined, next we model them using the Web Ontology Language (OWL)[3]. We choose OWL because of its expressivity and reasoning support. OWL is expressive enough to represent the concepts and constraints which we need to cover in the model (with the support of SWRL rules[4]). Also, its reasoning support enables us to perform data validation automatically.

We employ a modular approach to the model design. The model consists of two parts, an upper ontology part and a domain extension part (Figure 2). The upper ontology includes a set of general concepts describing observations and measurements. It is built on existing modelling efforts (Henson et al., 2009; Madin et al., 2007; Probst, 2006). A key concept of the upper ontology is `Observation`. Associated with it are concepts such as `Property`, `Procedure`, and `Feature`. For each observation, there are exactly one observed property, one estimated value of the property, one procedure used to generate the value, one sampling feature and one observation time. We represent this in OWL2 with qualified cardinality restrictions:

$$\begin{array}{lll} \mathit{Observation} \sqsubseteq & (=1 \mathit{hasProperty.Property}) & \sqcap \\ & (=1 \mathit{hasProcedure.Procedure}) & \sqcap \\ & (=1 \mathit{hasSamplingFeature.Feature}) & \sqcap \\ & (=1 \mathit{hasTime.Time}) & \sqcap \\ & (=1 \mathit{hasValue.xsd{:}anyType}) & \end{array}$$

Each `Property` is measured in one type of `Unit`[5]; there is no restriction on which units are used, as long as they belong to the same type. For example, the length of an object can be measured in any length units (e.g. metres or feet). This facilitates the translation of data in different systems of measurement.

[3] http://www.w3.org/TR/owl-features/.

[4] http://www.w3.org/Submission/SWRL/.

[5] We model `Property` and `Unit` based on POSC (Petroleum Open Standards Consortium) definitions, http://www.energistics.org/.

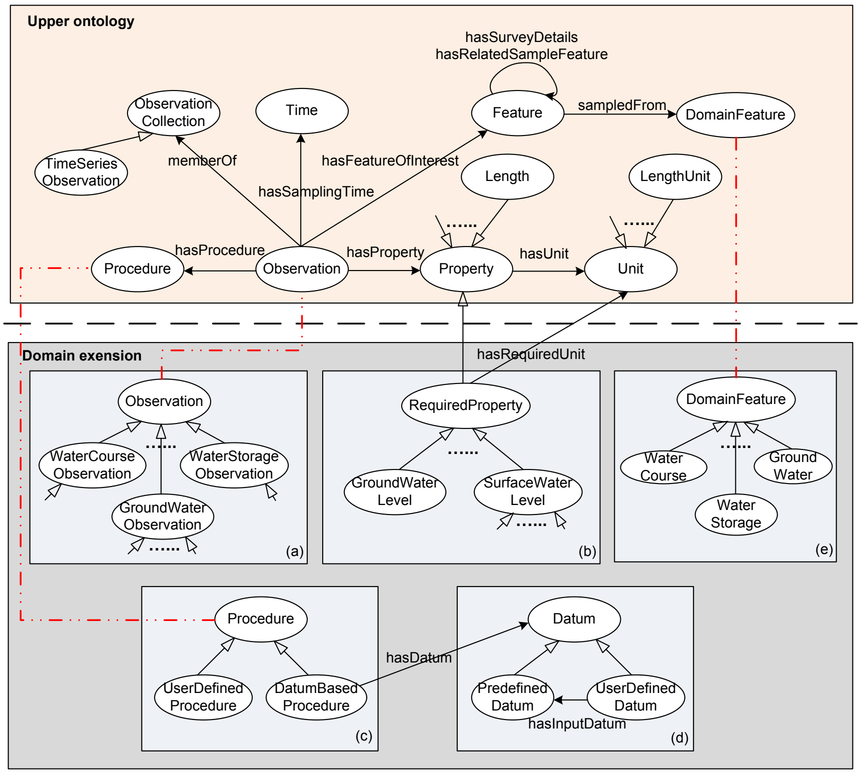


Figure 2. Part of the knowledge model.

A set of observations constitute `ObservationCollection`. We model `TimeSeriesObservation` as a subclass of `ObservationCollection`. Each time-series consists of observations which share the same property, the same procedure, the same feature but have different observation times. We express this using Semantic Web Rule Language (SWRL) rules. For example, to express members of a time-series having the same observed property, we define:

$$TimeSeriesObservation(?s) \wedge hasMember(?s, ?x) \wedge hasProperty(?s, ?p1) \\ \wedge hasProperty(?x, ?p2) \rightarrow sameAs(?p1, ?p2)$$

The domain extension part basically covers the concepts specific to the water domain, which we represent by extending the upper ontology. Figure 2 (the lower part) gives a few examples:

(a) shows part of the observation class hierarchy. Each observation class corresponds to a category or subcategory of water information as defined in the Regulations.

(b) depicts part of the required property class hierarchy. Each `RequiredProperty` has exactly one `Unit`. This is to ensure that the translated data are expressed in the units as specified in the Regulations. For level properties, datum information needs to be provided.

(c) extends `Procedure` with two subclasses, `UserDefinedProcedure` and `DatumBasedProcedure`. For a procedure based on datum, datum information needs to be provided.

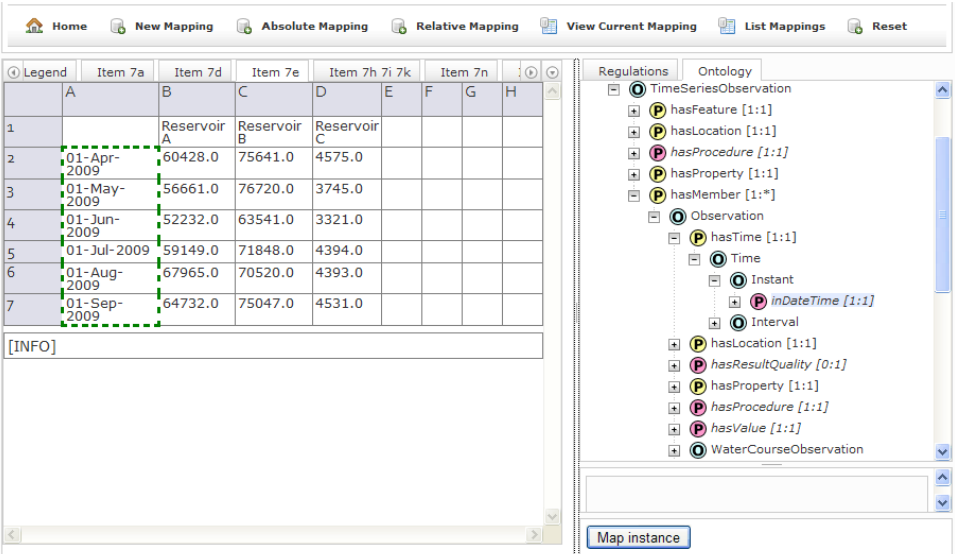
(d) shows two subclasses of `Datum`, `PreDefinedDatum` and `UserDefinedDatum`. Examples of predefined datums include WaterSurface and AHD (mainland Australian Height Datum). For a user-defined datum, its vertical coordinate reference system, value, and input datum (a predefined datum) need to be specified.

(e) extends `DomainFeature` with subclasses such as `GroundWater`, `WaterCourse`, and `WaterStorage`. This extension will be replaced by definitions from the Australian Hydrological GeoFabric (AHGF)[6] in the future.

4 APPLICATION SCENARIO

Driven by the fact that 32% of data from organisations are represented in spreadsheets, we have developed an application in which the model is used for translating spreadsheet data into WDTF. The application includes three major parts:

- A GUI frontend (Figure 3) for users to load spreadsheets into, view the model, and compose mapping expressions between them.
- A mapping evaluator which evaluates mapping expressions and generates model instances using Jena[7] and Pellet[8]. See Shu and et al. (2010a) for mapping details and evaluation.
- A back-end to retrieve instance data from the model using SPARQL[9] and generate WDTF instance documents through XQuery scripts.



	A	B	C	D	E	F	G	H
1		Reservoir A	Reservoir B	Reservoir C				
2	01-Apr-2009	60428.0	75641.0	4575.0				
3	01-May-2009	56661.0	76720.0	3745.0				
4	01-Jun-2009	52232.0	63541.0	3321.0				
5	01-Jul-2009	59149.0	71848.0	4394.0				
6	01-Aug-2009	67965.0	70520.0	4393.0				
7	01-Sep-2009	64732.0	75047.0	4531.0				

[INFO]

Figure 3. Using the model for spreadsheet data translation.

We have tested the application on the data from several organisations. The initial experience with the application is encouraging: with a little help from users, data can be easily translated into WDTF. Our next step is to make the application operational.

5 CONCLUSION

To address the Bureau's data ingestion problem, knowledge models have been proposed to capture semantic gaps between data from organisations, WDTF, and the Regulations. In this paper, we focus on

[6] http://www.bom.gov.au/water/geofabric/.
[7] http://jena.sourceforge.net/.
[8] http://clarkparsia.com/pellet/.
[9] http://www.w3.org/TR/rdf-sparql-query/.

the construction of such a model. We first summarise our findings on semantic gaps. Based on this, we then define the concepts and constraints to be covered, and represent them using OWL with a modular approach. Finally, we describe an application using the model for spreadsheet data translation. We have tested the application on the data from several organisations. The initial experience with the application is encouraging. Our next step is to make the application operational.

ACKNOWLEDGEMENT

This work is part of the Water Information Research and Development Alliance (WIRADA) between CSIRO's Water for a Healthy Country Flagship and the Bureau of Meteorology. The work is supported by the Tasmanian ICT Centre. The Tasmanian ICT Centre is jointly funded by the Australian Government through the Intelligent Island Program and CSIRO. The Intelligent Island Program is administered by the Tasmanian Department of Economic Development, Tourism and the Arts.

REFERENCES

Cox, S. (2007a). Observations and Measurements - Part 1 - Observation schema Version 1.0 OGC. In *OGC document 07-022r1.*

Cox, S. (2007b). Observations and Measurements - Part 2 - Sampling Features Version 1.0 OGC. In *OGC document 06-188r1.*

Henson, C., H. Neuhaus, A. Sheth, K. Thirunarayan, and R. Buyya (2009). An Ontological Representation of Time Series Observations on the Semantic Sensor Web. In *Proceedings of the 1st International Workshop on the Semantic Sensor Web (SemSensWeb 2009).*

Madin, J., S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa (2007). An Ontology for Describing and Synthesizing Ecological Observation Data. In *Ecological Informatics 2 (2007). pp.279-296.*

Probst, F. (2006). An Ontological Analysis of Observations and Measurements. In *Proceedings of the 4th International Conference on Geographic Information Science (GIScience).*

Shu, Y. and et al. (2010a). Making Sense of Spreadsheet Data: A Case of Semantic Water Data Translation. In *Proceedings of the 6th Australasian Ontology Workshop (AOW2010).*

Shu, Y. and et al. (2010b). Semantic Water Data Translation: A Knowledge-driven Approach. In *Proceedings of the 14th International Database Engineering and Applications Symposium (IDEAS10).*

Walker, G., P. Taylor, S. Cox, and P. Sheahan (2009). Interim-water data transfer format(iwdtf): Guiding principles, technical challenges and the future. In *Proceedings of the 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation*.