18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Water Data Transfer Format (WDTF): Guiding principles, technical challenges and the future

Walker, G. [1], P. Taylor [1], S. Cox [2] and P. Sheahan [3]

[1] CSIRO ICT Centre, [2] CSIRO Exploration and Mining, [3] Australian Bureau of Meteorology
Email: Gavin.Walker@csiro.au

Abstract: In response to the growing water crisis in Australia, the Bureau of Meteorology has been given a mandate to develop and maintain an integrated national water information system. Section 126 of the Commonwealth Water Act 2007 provides for the making of Regulations to support these functions, which came into effect on 30th June 2008. These regulations (BoM, 2008) define the requirements for the collection of water data by the Bureau of Meteorology in 10 data type categories including flows, groundwater levels, reservoir storage, water quality, water use, water entitlements and water trades. Over 200 data providers are involved, grouped in 8 categories.

This paper describes the joint effort by CSIRO and the Bureau of Meteorology to design the Water Data Transfer Format (WDTF), an eXtensible Markup Language (XML) (W3C, 2008) format for the ingestion of water data by the Bureau. The paper focuses on defining principles, technical challenges and the future of this work.

WDTF development was guided by the following principles:

- A solid modelling foundation: The format must be underpinned by a conceptual model expressed in a modelling language such as the Unified Modelling Language (UML) and supported by a model driven approach used to derive products such as XML Schema, documentation and validation tools.
- Adopt, adapt, invent: Where best practice exists models and encoding should be adopted, adapted as required and finally invented if no other options exist.
- Validation: All derived documents must be testable for compliance to the conceptual model and the Bureau of Meteorology's business rules.
- XML for current tools: The format should be suitable to use with relevant standards-based commercial off the shelf tools.
- Normalisation: Similar to database normalisation, the main entities or concepts should be encoded as distinct well-contained XMLSchema fragments, and identifiers used to define relationships between fragments.

In following these principles WDTF has built on the International Standard Organisation's (ISO) General Feature Model, ISO 19109 (ISO/TC-211, 2005), and the Open Geospatial Consortium's (OGC) Observations and Measurements (O&M) model (Cox, 2007a & b). Encodings were based on an adaptation of O&M to a simple profile of OGC's Geography Markup Language (GML), to make it more compliant with current tools and ease the validation process.

The paper also presents a summary of some of the challenges faced during development. These include:

- Semantics: Coming to a common understanding of concepts in the format.
- Procedure metadata: Describing the context of the observation to use the data appropriately.

Finally there is a brief discussion of plans for WaterML2.0, a partial successor of WDTF. The Consortium of Universities for the Advancement of Hydrological Science Incorporated (CUAHSI[1]), CSIRO and the Bureau of Meteorology aims to have WaterML2.0 accepted by the World Meteorological Organisation.

Keywords: *XML, Bureau of Meteorology, Water Data Transfer Format (WDTF), WaterML, OGC*

[1] http://www.cuahsi.org/

1. INTRODUCTION

Standards development for information systems are changing: the mostly textual reference documents are progressively being replaced by machine-readable specifications in languages like the eXtensible Markup Language (XML) (W3C, 2008) and based on a modelling approach in a language like the Unified Modelling Language (UML) (OMG, 2009a).

XML is a significant change from the current comma-separated values approach used across the water industry for the last decade. It is well supported by technologies (to validate it, to transform it, import and export it from databases) which facilitate the creation of software components for data management. Use of XML was driven by the Bureau of Meteorology's need for a robust data ingestion process, which XML aids through its ability to identify and incrementally improve data quality through validation tools. The Bureau also chose to produce WDTF to drive modernisation and standardisation within the water industry.

In June 2008 the Bureau of Meteorology was faced with the need to start collecting data by October 2008 under the national water information system regulations. With no format prescribed by the regulations, the Bureau of Meteorology requested CSIRO help develop a format for water data transfer. The first version of WDTF (version 0.1) was produced in August 2008, with version 0.3 released in February 2009.

This paper looks at the principles underlying the format development, the key conceptual models underpinning the encoding, technical challenges in defining the format and lastly it discusses the future, primarily in the context of developing an international water data transfer standard.

2. PRINCIPLES

Development of the format was guided by the following principles: a solid modelling foundation, "adopt, adapt, invent", validation, XML suitable for current tools and normalisation.

2.1. A solid modelling foundation

The transformation of a model to exchange formats, documentation and validation tools are the outcomes of a model-driven approach (OMG, 2009b). Automation of these transformations makes maintenance easier and promotes consistency in derived products. A model drive approach encourages re-use of previous modelling investments and takes a long-term perspective of encodings by allowing evolution of encodings in new technologies from the same sound modelling background. WDTF must be underpinned by a conceptual model expressed in a modelling language such as the Unified Modelling Language (UML). It has two foundational models, the ISO General Feature Model (GFM) (ISO/TC-211, 2005) and the OGC's Observation and Measurement Model (O&M) (Cox, 2007a & b), which will be detailed in section 3.

2.2. Adopt, adapt, invent

Development of WDTF followed the Water Resource Observation Network Reference Model (WRON-RM) (O'Hagan et al., 2007) "Adopt, adapt, invent" principle: adopting existing standards, protocols and procedures where possible, adapting where necessary and inventing as a last resort. Adopting models and encodings from other bodies relieves the burden of maintenance of those aspects of the format as well as encouraging reuse of tools based on the same aspects.

The OGC Observations and Measurements (O&M) model is the point of convergence of a range of ISO TC 211 and OGC activities (Bacharach, 2007) and has been adopted by a growing list of communities including: GeoSciML(Simons et al 2006), Climate Sciences Modelling Language (CSML) (Woolf & Lowe, 2007), Groundwater Markup Language (GWML) (Boisvert & Brodaric, 2007) the Integrated Ocean Observing System (IOOS) (Alexander, 2008) and Eurocontrol's WXXM 1.1 (Hart, 2009). The O&M model is the basis of all observations in WDTF.

2.3. Validation

It was imperative to supply effective validation tools to ensure provider compliance with the format. The initial validation suite consisted of XMLSchema validation, including some enumerations and patterns, however this was not sufficient to enforce the business rules - also known as a Data Product Specification (DPS) (ISO/TC211, 2007), of the format developer. A DPS includes the schema but also specifies a complete vocabulary definition for all elements and the correct values and occurrences of elements in the context of the values of other elements in the document.

Current work using Schematron (Schematron, 2009) allows validation of XML fragments with attention to both the content and structure of the context. Complex validation rule base maintenance is at present a challenge. In future versions of WDTF, model driven approaches offer the capacity to generate complex validation suites and documentation based on the DPS, improving consistency in updates and therefore maintenance.

2.4. XML for current tools

The Bureau of Meteorology has a business requirement to process and ingest WDTF. In order to be useable, the format had to be supported by current tools, including the ability to turn schema objects into code. WDTF has introduced XML to many agencies; therefore, it was essential the standards based commercial off-the-shelf tools could be used to ensure an easy transition for adopting agencies. This impacted the design of the XML schemas, adding a requirement not to use overly complex XMLSchema constructs. The two main guidelines were:

1. No wildcard fragments: That is, no Xml Schema "anyType"
2. No recursion or cyclic inclusion.

The most significant implementation of O&M is in the context of OGC's Sensor Web Enablement (SWE) suite (Botts et al., 2007) using the Geography Markup Language (GML) (Portele, 2007). The GML used in OGC SWE contains many wildcards and recursion points and so is unsuitable based on the criteria above.

The approach taken was to adapt key classes from OGC SWE to comply with a simple profile of GML (GMLSF) (Vretanos, 2006) which meets the rules above. The GMLSF profile was created to support common WFS (Web Feature Service) implementations by restricting the XML schema, and thereby subsetting GML.

2.5. Normalisation

The term "normalisation" comes from the database world (Codd, 1990) where a database schema is either un-normalised or normalised in varying degrees (first normal form through to sixth normal form). An un-normalised schema has everything in the database schema included in one big table, and allows entries to be lists. High levels of normalisation split the schema into multiple tables where the tables are designed to support a specific set of data. The goal of normalisation is to reduce redundancy in the database and therefore alleviate update problems. Entries in tables are connected using common keys or identifiers.

For example, a lake has properties including: name, location, type of water (fresh or salt) and depth. If depth readings are taken at different times then there is a list of depth and time observations. In an un-normalised schema (figure 1a) the lake information is required when specifying the observation data. In a normalised schema (figure 1b) there is a separate observation table and a lake table. The lake information is stated only once and the observation table contains all the depth observations and the name of the lake as an extra property.

Everything

Name	Location	Water	Observation
Eagle	45,141	fresh	1 Mar 2.0m 2 Mar 2.1m 3 Mar 1.9m

Figure 1a. (above) Un-normalised table

Lake

Name	Location	Water
Eagle	45,141	fresh

Observation

Time	Depth	Lake Name
1 Mar	2.0m	Eagle
2 Mar	2.1m	Eagle
3 Mar	1.9m	Eagle

Figure 1b. (right) Normalised tables

Similarly, in WDTF, the model and the XMLSchema encoding must represent significant specialisation of data as distinct entities. An XMLSchema built as a single tree structure is equivalent to an un-normalised database schema. One clear advantage of a normalised exchange format is to allow separate transmission of significant entities. For example, the depth observations in the table above can be transmitted without the lake information. Only the identity of the lake needs to be transmitted. This reduces transmission volume and, more importantly, avoids the observation producing system needing to know about lake properties.

Entities are identified with Universal Resource Identifiers (URI). Enabling references and providing persistence requires almost everything in WDTF to be given a name or identifier. These are constructed following simple rules. WDTF follows the standard pattern to support normalisation in GML based schemas, using the xlink:href attribute (W3C, 2001) to refer to entities provided once but used many times.

3. DATA MODELLING

WDTF has adopted the O&M model as the basis of its observation model. As indicated in section 2.2 there is significant international support behind this model. The most significant encodings of O&M have been carried out using OGC standards (see section 2.4). All OGC standards are based on the ISO TC 211's General Feature Model (GFM), so that too has been adopted, by way of inheritance.

3.1. The General feature model (GFM)

The GFM, ISO 19109 (ISO/TC211, 2005), is the cornerstone of ISO TC 211's spatial information standards. Features are mostly representations of physical entities (e.g. roads, rivers, bridges, sensors) but may also include more abstract entities (e.g. water entitlements, water restrictions, observations). Features have properties, both simple and complex. For example, a monitoring station is a feature having properties such as manufacturer, sensors on board, maintenance schedule and location. The location is itself an identified entity (say a gml:Point) with properties srsName and pos, which are the coordinate system and coordinates respectively. A property might also represent a feature by reference using the xlink:href syntax in WDTF. For example, the featureOfInterest property of the TimeSeriesObservation feature uses xlink:href to reference the feature being observed.

WDTF documents encode the HydroCollection feature (i.e. the outer element is a HydroCollection) whose properties are the features (such as observations and sampling features) to be transmitted in the document For example, if a document is to contain a TimeSeriesObservation feature, then the HydroCollection feature will contain an observationMember property which in turn contains the TimeSeriesObservation feature. The properties of HydroCollection and the features they contain include

- siteMember: SamplingPoint, SamplingGroup.
- specimenMember: Specimen (both samples and bottles taken from samples).
- observationMember: Measurement (or Specimen observations), ComplexObservation (for gaugings), TimeSeriesObservation and GeometryObservation (for river cross-sections).
- featureMember: Storage, WaterCourse. A full set of these are yet to be defined.
- conversionMember: Conversion (rating) tables and DurationGroups being their scheduling.

3.2. Observations and measurements (O&M)

An important type of feature in WDTF is the concept of the observation. These are derived from OGC's O&M standard (Cox, 2007a). O&M observations (Figure 2) can be summarised as follows: (Cox, 2007a)

An Observation is an action whose result is an estimate of the value of some property of the feature-of-interest, obtained using a specified procedure

For example, observing salinity in a lake will produce salinity in mg/L (result) which is an estimate of the salinity (observed property) in Eagle Lake (feature of interest) using salinity meter 00435 (procedure). While this breakup seems simple in theory, each aspect (feature, property, procedure and result) is complex and an efficient way to capture that information is challenging.

In many practical cases, observations are not made on the feature of ultimate interest, or sampled feature, but it is necessary to choose a sampling strategy for the feature. The Sampling Features information model is described in Part 2 of O&M (Cox, 2007b). For example, a submerged water sensor is measuring salinity in a lake. It is not measuring the salinity of the lake but the small amount of water around the sensor. The sampling feature can be represented as the volume around the sensor, or more commonly, the location of the platform or station on which the sensor is mounted, depending on the amount of accuracy required. In this case the sampled feature is the lake.

This relationship between a property of the sampling feature and the equivalent property in the sampled feature might be explicitly described, computed dynamically, undefined or defined as having the same value. Specification of this relationship is not currently within the scope of WDTF.

Sampling features in WDTF consist of SamplingGroup and SamplingPoint, imported from O&M Simple Features sampling schema and adapted to the WDTF application. It is useful to think of a SamplingGroup as

a place one can drive to and then walk to a SamplingPoint. The SamplingPoint is typically the physical location of the sensor. The SamplingGroup is a spatially cohesive set of SamplingPoints. For example, a Sampling Group may be a dam site within which there are Sampling Points such as: An intake tower measuring water drawn down a dam wall crest level gauge, sensor tower including salinity, turbidity and level sensors. Every SamplingPoint is required to be a member of a SamplingGroup.

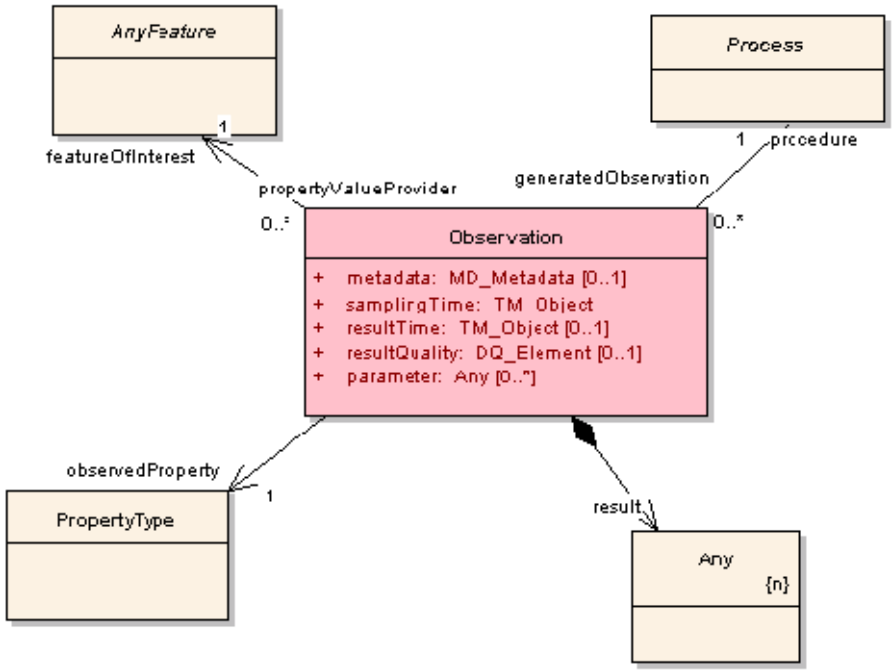

Figure 2. Observations and Measurements UML model

4. TECHNICAL CHALLENGES

With the data provision process driven by legislation, there has been widespread acceptance of the need for common formats. Adoption of WDTF, however, has not been without challenges. Semantics and procedure metadata are discussed briefly to give a flavour of the challenges faced by the standard developers.

4.1. Semantics

Semantics is the allocation of meaning to terms or words used in communication. Without clearly assigned meaning parties can be using the same words but meaning different things. Without a way to translate the agreed meaning onto the data held by a party, the party cannot provide data that matches the meaning. Development of WDTF was assisted by two content reviews, comprising members from the water software industry in Australia. The reviews agreed on the semantics of various terms. One controversial term was "site". Site is a feature containing sensors, such as a dam, lake or river. To group (a) a site contains sensors and the site is the only given location for all sensors, for example, the Derwent River. To group (b) the sensors have individual locations within the site, for example, specific points on the Derwent River. After much discussion it was agreed that a site would represent a group of sensors and all sensors would be given locations even if all that was known was that they were approximately at the site. Likewise, other semantic issues were resolved, other examples include: standardisation of quality codes and units.

4.2. Procedure metadata

Procedures in O&M describe the process used to obtain a measurement of the property of a feature. Typically this is the name of a sensor or well known measurement method. Feature metadata provides information to interpret the result correctly such as: the measurement method, interpolation type or if the result was derived from other data. In OGC implementations of O&M, such as the Sensor Observation Service (SOS) (Na & Priest, 2007), procedure metadata is a reference to a document that can be later retrieved. SOS uses the SensorML encoding which is quite complex (Botts & Robin, 2007). There was little

appetite for full SensorML documents because the procedure data is not readily available from the data providers. On the other hand the content reviews emphasized that the water data has little meaning unless sufficient metadata is provided to understand the context in which the data was captured. While WDTF supports some procedure metadata, most providers omit it or specify it as unknown. There is a need to establish the minimum procedure metadata required and to develop a strategy to obtain it from the data providers.

5. DIRECTIONS FOR WATERML 2.0

An international summit in September 2007 (Cox & Brodaric, 2007) agreed that a conceptual model of water information and one or more encodings of it were needed. Such an information model would facilitate better interoperability and promote better dissemination, understanding and collaboration between water data users. This would also make possible access to hydrological data sets by other interested domains such as climate change research, meteorology and oceanography.

There are other, somewhat similar models in existence; the most closely aligned being the Consortium of Universities for the Advancement of Hydrological Science Incorporated (CUAHSI) WaterML (Zaslavsky et al., 2007). CUAHSI have effectively leveraged their WaterML specification to allow integration of national water systems in the US; this is further evidence that common data and information exchange models facilitate and encourage data sharing and interoperability. The Bureau of Meteorology's use cases, to meet the regulations, are captured in WDTF. WaterML, in a similar way, was developed to address a specific requirement set, both capturing similar concepts in a slightly different fashion.

CUAHSI and CSIRO are now working together to develop a common information model, currently being termed WaterML2.0, which addresses common needs in the water industry internationally. This, or a future version of it, will be put forward to the OGC Hydrology working group and to the World Meteorological Organization (WMO), as the beginnings of an international standard. The project is accounting for current best practices for the development of international standards of relevance, such as CSML (Woolf & Lowe, 2007), GWML (Boisvert & Brodaric, 2007) and GeoSciML (Simons et al., 2006).

The model will be one that is extensible for specific needs, while being flexible enough to include any future requirements that may arise. It is envisaged that a future WDTF will contain at its core the WaterML2.0 model, with further extension containing the specific requirements of the Bureau of Meteorology. The current work on the model development is in the early stages, but looks set to:

- Focus on time series observations and sampling features; WDTF also covers features and rating tables.
- More soft-typing in metadata. Metadata in WDTF currently use strongly typed tags in XML. This was done to allow XML Schema-based validation. The alternative is to use weakly typed tags (string) with attributes defining the types. This allows the same schema to be used with a different set of metadata or vocabulary. This is the preferred approach when developing a standard that is intended to be used across organisational boundaries, as organisations often maintain or utilise particular sets of vocabularies.
- There will be a common vocabulary of core concepts to allow some interoperability across jurisdictions. Coming to agreement on these core concepts will be of key concern in the development of the model.

6. CONCLUSION

WDTF was developed, in a short time frame, to meet the Bureau of Meteorology's needs for a water transfer format. It was guided by the principles of: A solid modelling foundation, "Adopt, adapt, invent", Validation, XML for current tools and normalisation. Significant support for the O&M model made it the obvious choice to underpin development. Its implementation in the OGC's standard required the adoption of ISO/TC211's General Feature Model (GFM) also.

Implementation of WDTF was not without its challenges, with difficulties in semantics and metadata. An international collaboration with the CUAHSI will see WaterML2.0 developed and proposed as an international standard.

Opportunities to see widespread adoption of standards are rare. The requirement to deliver consistent data to the Bureau of Meteorology has the potential to improve water data management nationally, with further funding available to support adoption. With data already being received, significant work is still required to help data owners translate their own data models into the WDTF data model and ensure quality data is provided. As this initiative matures, with better tool support, more knowledge of such approaches at the operational level, it is hoped that other natural resource management communities will benefit from our experiences.

ACKNOWLEDGMENTS

This work is part of the water information research and development alliance between CSIRO's Water for a Healthy Country Flagship and the Bureau of Meteorology. The authors would like to thank Laurent Lefort (CSIRO ICT Centre) and Spenser Kao (Bureau of Meteorology) for their contributions to the format, Jonathan Yu (CSIRO Land and Water) for his work on validation and Amit Parashar and Peter Thew for reviewing the paper.

REFERENCES

Alexander, C. (2008), NOAA Integrate Ocean Observing System (IOOS) Program: Data Integration Framework Design Document version 1.0. Accessible at http://ioos.noaa.gov/library/difdmacdocs.html

Bacharach, S. (2007), New Implementations of OGC Sensor Web Enablement. *Standards Sensors Magazine,* Dec. 2007

Boisvert, E. and Brodaric, B. (2007), GroundWater Markup Language (GWML): Extending GeoSciML for Groundwater. American Geophysical Union, Fall meeting 2007.

Bom (2008), Water Regulations 2008: Select legislative instrument 2008 No. 106 as amended. Retrieved on 31 March 2009 from http://www.bom.gov.au/water/regulations/regulations.php

Botts, M. and Robin, A. (eds) (2007), OpenGIS Sensor Model Language (SensorML) Version 1.0 OGC Document 007-000 17 Jul. 2007 Accessible from http://www.opengeospatial.org/standards/sensorml

Botts, M., Robin, A., Davidson, J. and Simonis, I. (2007), OGC Sensor Web Enablement: Overview and High Level Architecture, OGC document 07-165. Accessible from http://www.opengeospatial.org/pressroom/papers

Codd, E.F. (1990), *The Relational Model for Database Management: Version 2*. Addison-Wesley p. 271

Cox, S. (ed.) (2007a), Observations and Measurements – Part 1 - Observation schema Version 1.0 OGC document 07-022r1. Accessible from http://www.opengeospatial.org/standards/om

Cox, S. (ed.) (2007b), Observations and Measurements – Part 2 - Sampling Features Version 1.0 OGC document 07-002r3. Accessible from http://www.opengeospatial.org/standards/om

Cox S. and Brodaric, B. (2007), Water Resources Information Model Workshop Canberra, 25-27 September, 2007. Accessible from http://wron.net.au/documents/WaterML_workshop_report.pdf

Hart, D. (2009), MET *matters* in a Single European Sky. Presentation to OGC Meteorology Domain Working Group. Accessible from http://portal.opengeospatial.org/files/?artifact_id=33012

ISO/TC-211 (2005), 19109 Geographic information - Rules for application schema. International Organization for Standardization.

ISO/TC-211 (2007), 19131 Geographic information – Data product specification. International Organization for Standardization.

Na, A. and Priest, M. (eds.) (2007), Sensor Observation Service Version 1.0 OGC document 06-009r6. Accessible from http://www.opengeospatial.org/standards/sos

O'Hagan, R.G., Atkinson, R., Cox, S., Fitch P., Lemon, D. and Walker, G. (2007), A Reference Model for a Water Resources Observation Network. Proceedings of the International Congress on Modelling and Simulation (MODSIM 07), Christchurch, New Zealand, 10-12 December, 2007.

OMG (2009a), Unified Modeling Language (UML). Accessible from http://www.uml.org/

OMG (2009b), Model Driven Architecture (MDA). Accessible from http://www.omg.org/mda

Portele, C. (ed.) (2008), OpenGIS Geography Markup Language (GML) Encoding Standard version 3.2.1 OGC Document 07-036. Accessible from http://www.opengeospatial.org/standards/gml

Schematron (2009), Schematron. Accessible from http://www.schematron.com/

Simons, B., Boisvert, E., Brodaric, B., Cox, S., Duffy, T.R., Johnson, B.R., Laxton, J.L. and Richard, S. (2006), GeoSciML: Enabling the Exchange of Geological Map Data. Australian Earth Sciences Convention 2006

Vretanos, P. (2006), Geography Markup Language (GML) Simple Features Profile. OGC document 06-049r1. Accessible from http://www.opengeospatial.org/standards/profile

Woolf, A. and Lowe, D. (2007), CSML User's Manual, v2. Accessible from http://ndg.nerc.ac.uk/csml/

W3C (2001), XML Linking Language (XLink) version 1.0. Accessible from http://www.w3.org/TR/xlink/

W3C (2004), XML Schema part 1: Structures 2nd ed. Accessible from http://www.w3.org/TR/xmlschema-1

W3C (2004), XML Schema part 2: Datatypes 2nd ed. Accessible from http://www.w3.org/TR/xmlschema-2

W3C (2008), Extensible Markup Language (XML) 1.0 (5th ed.). Accessible from http://www.w3.org/TR/xml

Zaslavsky, I., Valentine, D. and Whiteaker, T. (2007), CUAHSI WaterML 0.3.0 07-041r1 OGC Discussion paper. Accessible from http://www.opengeospatial.org/standards/dp