Developing Ontologies for Linked Geospatial Data

Kerry Taylor[1], Laurent Lefort[1], Geoff Squire[1], Gavin Walker[1], Andrew Woolf[2], Yanfeng Shu[1], David Ratcliffe[1], Simon Cox[1], Armin Haller[1]

1. CSIRO Australia. firstname.lastname@csiro.au

2. Australian Bureau of Meteorology. A.Woolf@bom.gov.au

Abstract

We compare experiences in modelling geographic information in OWL by two very different methods. Both methods have a root in OGC standards, but one followed a semantic web development style and the other a UML-flavoured Model Driven Architecture style. We ask how much of the OGC UML modelling should be preserved in the transition from UML modelling to OWL modelling and from XML to open linked data.

Introduction

We have developed the first Australian Government Linked Data, 100 years of climate observations of the Bureau of Meteorology, and have subsequently established the Australian Government Linked Data Working Group to promote linked data for Commonwealth public sector information. This dataset, Australian Climate Observations Reference Network - Surface Air Temperature (ACORN-SAT), was modelled using several community ontologies, including one that was partly derived from the OGC O&M standard. Meanwhile, mandatory regulations require water-managing agencies in Australia to regularly provide data about measurements of water storage, water course, water quality and climate. Driven by a firm commitment to OGC models and UML modelling in particular, we built an ontology for this data through an automated model-driven tool chain anchored in ISO TC211 UML models, but the result was less than satisfying. The outcome of the hand-modelling, extending popular ontologies in the first project was unquestionably better.

Australian Climate Observations Reference Network - Surface Air Temperature Dataset

In 2012 the Australian Bureau of Meteorology publically released an important reference data set for climate science as a collection of many CSV files together with extensive PDF documentation. The Bureau proposed to additionally release it as linked data. This proceeded fairly routinely from a semantic web point of view. With an eye constantly on URI design, we borrowed from the experience of UK Bathing Water pilot (http://environment.data.gov.uk/bwq) and adopted the Epimorphics linked data API (ELDA) package and designed our ontology based on the Datacube vocabulary (now a Recommendation from the W3C Government Linked Data Working Group). We also readily adopted the SSN ontology of the W3C Semantic Sensor Network Incubator Group [1] in which we had had a hand.

The SSN ontology itself was very strongly guided by the OGC's O&M model in the relevant places, but was not bound to be a faithful representation, and so free to exploit all the modelling benefits of ontology design from scratch. In most cases the labels (and URIs) of the SSN ontology were taken directly from O&M and were annotated in the ontology accordingly (using the dc:source property). We have used the ontology in subsequent projects, including the publication of linked data related to sensing livestock, soil properties, and weather on a farm [2,3].

The SSN does not prescribe a representation for measured data; we have that measurements at a point in time are instances of both ssn:Observation and qb:Observation. In addition to the SSN, our ontology also used intervals from data.gov.uk (http://reference.data.gov.uk/def/intervals) and some custom ontology fragments to relate the external ontologies and to provide what was missing. As a last step we processed

the location identifiers (lat-longs) of weather stations with close matches to the geonames database (www.geonames.org), with a small number of manual adjustments then required for weather stations that were wrongly associated with offshore features. This enabled us to link to geonames data such as the location name and administrative district. We also built a couple of mashups for typical data navigation and presentation scenarios that fall outside the scope of the Epimorphics browser/html interface.

This linked data product became the first linked open data on Australia's data.gov.au in early 2013 and won two industry awards for innovation in Australian government. The ontology model development component of this work took about 1.5 person-months; the whole project from CSV and PDF data to linked data was complete in 6 person-months.

The Australian Climate Observations Reference Network - Surface Air Temperature Dataset published as linked data will be briefly demonstrated in the presentation [4].

Australian Government Linked Data Working Group

This linked data project provided the stimulus to establish an Australian Government Linked Data Working Group in 2012 with representatives from several Commonwealth Departments. The Group aims to develop technical guidelines and best practice advice for agencies, inform the development of data.gov.au as a platform for publishing, promote the benefits and encourage adoption of linked data and to undertake specific activities and coordinate projects in pursuit of these objectives.

The Group has prepared draft proposals for the management of subdomains of data.gov.au for linked data, and guidelines for designing URIs. It is also building an ontology for Australian Government.

Adapting OGC/XML models to OWL/RDF models

Preceding this climate linked data project, we began a two-year exercise to translate ISO-TC211 UML models to OWL (O&M ISO 19156 and its dependencies: spatial ISO 19107, temporal ISO 19108, metadata ISO 19115 and reference systems ISO 19111). The purpose here was primarily to use the ontology in a tool known as *adhoc data ingestion* [5] which exploited the notion of a simple conceptual model expressed in user-understood terms and leveraging OWL's structural permissiveness, in a context of long-term commitment to OWL modelling for water data management [6]. The adhoc tool supports user-written mappings of information from spreadsheets to a common XML format. In this case, the XML schema known as WDTF [7], built on GML, that in turn borrows informally from O&M was to be used. Later, a UML model (Application Schema) was developed for WDTF, using the ISO TC211 models mentioned above and largely following GML3.2 Annex E rules. A model-mapping tool called *full moon* (https://www.seegrid.csiro.au/wiki/Siss/FullMoon) was built to generate the WDTF XML schema from the model (and, in principle, any other model specialising that ISO TC211 family). The Bureau of Meteorology wanted the ad-hoc tool to use an OWL ontology formally derived from the WDTF UML instead of an independent custom-built ontology, following model driven architecture principles. In principle then, later updates to the WDTF model could be automatically mapped to new versions of both the WDTF XML schema and also the WDTF OWL model with some confidence of consistency. That all sounds reasonable at first glance; tools to generate XML schema from UML models that specialised those were in use (full moon and later Shapechange) and the GML standard already demonstrated how to derive an XML model from those standards, loosely speaking. Further, routine mappings from UML to OWL were well known and available in tools.

In the first attempt, it became very clear that several different mappings would be needed to do this well because the XML-based schema standards (such as GML and WDTF) were already incompatible with the UML models and special-case translation had to be employed in places. Ideally, we would need to

- generate conceptual ontologies to represent the intended content of the UML models and to expose the dependencies on the base packages;
- generate implementation ontologies to represent the implementation details of XML schemas and to expose the dependencies on base schemas (such as GML),
- derive *harmonised* ontologies from variants of the *conceptual* ontologies and *implementation* ontologies which have been altered from the normative UML-XSD mappings in the standards (e.g. GML).

So that Model Driven Architecture goals could be met and changes could be automatically propagated we needed in addition to capture as many mappings as possible between the UML models and the conceptual ontologies, the conceptual ontologies and the harmonised ontologies, the harmonised ontologies and the implementation ontologies, and the implementation ontologies and the XML schemas. However, because these mappings were not formalised on the UML/XML side it was nearly impossible to reconstruct them on the ontology side. Furthermore, the non-standard substitution groups used liberally in the XML modelling meant that ordinary UML to OWL translation rules and tools were ineffective.

Ambitions for principled model drive architecture were cut back. At the project close in 2013, an OWL version of the WDTF was completed and deployed in the ad hoc data ingestion tool. This OWL model was automatically generated from the WDTF UML by extending the solid ground tool with some generic UML to OWL mappings and reproducing the specialist handling of substitution groups as is done for mapping to GML. Vocabularies represented outside the UML as spreadsheets are also separately incorporated into the OWL model in a special-purpose but automated way. ISO TC-211 model concepts are not translated from UML directly but instead hand-authored OWL interpretations (http://def.seegrid.csiro.au/static/isotc211/) of those models are imported.

However, the resulting ontology for WDTF fails to meet the purpose for which it was needed in the *ad-hoc data ingestion*. It is far from simple, it could not be described as conceptual, it is full of the legacy of UML structural tricks and uses the language of the OGC instead of the language of the regulations that specify the information required of the user community. At least if it had been developed directly from the WDTF schema instead of the UML, it could have been cleaner and could support translation of data from the XML to RDF. As it is, it fails its primary intended purpose and is almost certainly unusable for linked data, too. It has been expensive to produce (est. 15 person-months, excluding the hand-authoring).

Conclusion

Curiously, the kind of data represented in the first case study here is similar to that in the second. Both were reporting measurements, where those measurements were made at point locations. The latter offers all the richness of the OGC geometries for descriptions of measurement points, and exploits more of the richness of the O&M. However, the former also offers summary data (aggregations of different units of time, but not space) and incorporates a more expressive description of deployment whereby observation locations move over time. In a way, both case studies were derived originally from the O&M.

In 2010 Jenni Tennison (http://www.jenitennison.com/blog/node/142) said that reusing existing modelling activity is good but you have to *work* to map from a conceptual model to a particular modelling paradigm such as RDF. You have to take advantage of what its good at---there is no point in publishing linked data if is

unusable. This may be obvious, but it means the transition from the long-held OGC dependency on UML and XML towards OWL and linked data is not a straightforward one. There are fundamental problems with the principle of treating OGC-compliant UML models as root models from which other model artifacts, such as OWL domain models, might be derived in a deterministic fashion. And this means a long term incompatibility cannot be avoided.

When the W3C SSN Incubator Group proposed a mechanism for using the ontology in the context for the OGC SWE package of standards, we took an alternative route. We recommended an approach to annotation of the standardised XML using RDFa to incorporate semantics that would have no detrimental effect on current OGC-compliant processors. This does not itself deliver linked data, but provides the missing semantic interpretation of SWE data and also a clear path to building linked data out of OGC-compliant XML data [8]. Separately, many projects have adopted the SSN ontology for linked data directly, and through that have enjoyed the benefits of the modelling investment of the OGC in O&M.

For OGC to take the linked data route for information sharing it needs to recognise that the OWL/RDF stream will need to be progressed alongside and independently of the UML/XML paths, and that the two paths should not be considered compatible in other than a weak conceptual sense. Already, general purpose UML processing tools cannot be used in the generation of XML schemas because of the heavy use of non-standard UML stereotypes and implicit properties. The second ontology of Cox [9] that is derived by hand from the UML O&M model and that treats it as a view on an unwritten external conceptual model instead of a constraining frame model is illustrative. But it does not go far enough—it should be a model designed firstly for the semantic web with only a passing reference to its UML roots.

Bibliography

[1] Compton et al, The SSN ontology of the W3C semantic sensor network incubator group, Web Semantics: Science, Services and Agents on the World Wide Web **17**(25--32) 2012.

[2] Gaire et al, Semantic Web Enabled Smart Farming, Proc 1st International Workshop on Semantic Machine Learning and Linked Open Data for Agricultural and Environmental Informatics (SML2OD)) , CEUR-Proceedings, **1035** 2013.

[3] Taylor et al, Farming the Web of Things, IEEE Intelligent Systems, **28:**6(12-19) 2014.

[4] Lefort et al, A Linked Sensor Data Cube for a 100 Year Homogenised Daily Temperature Dataset, 5th International Workshop on Semantic Sensor Networks (SSN-2012), CEUR-Proceedings, **904** 2012.

[5] Shu et al, Semantic water data translation: A knowledge-driven approach, Proc Fourteenth International Database Engineering & Applications Symposium, pp 52-60, 2010.

[6] Ackland et al, Semantic service integration for water resource management, The Semantic Web–ISWC 2005, LNCS **3729**(816-828), 2005.

[7] Walker et al, Water Data Transfer Format (WDTF): Guiding principles, technical challenges and the future." 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, 2009.

[8] Lefort et al, Semantic Sensor Network XG Final Report, W3C Incubator Group Report, http://www.w3.org/2005/Incubator/ssn/XGR-ssn-20110628/ June 2011.

[9] Cox, An explicit OWL representation of ISO/OGC Observations and Measurements, 6th International Workshop on Semantic Sensor Networks (SSN-2013), CEUR-Proceedings, **1063** 2013.