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

Making information models work harder

W. Francis [a], R. Atkinson[b] , P. Box[c], S.J.D. Cox[d], and J. Yu[d]

[a] CSIRO Land and Water, Black Mountain, Australian Capital Territory, Email: will.francis@csiro.au
[b] CSIRO Land and Water, Lucas Heights, New South Wales , [c] CSIRO Land and Water, North Ryde, New South Wales, [d] CSIRO Land and Water, Highett, Victoria,

Abstract: Information models are useful for representing concepts and relationships in a domain of discourse. These models are typically used to guide system design and implementation and are also used as documentation. Models are also used to assist with system integration enabling multiple stakeholders to agree on a common structure and semantics for sharing data. For example, information models developed in Unified Modelling Language (UML) developed according to ISO 19100 series standards may be used to develop Geography Markup Language schema which specify how geographic data may be encoded for exchange as XML.

In this paper we propose additional uses for UML information models, enabling them to be connected to additional information specifying the semantic content of data, and delivered using Linked Data approaches. We also describe the role of models in enabling transformation and integration of heterogeneous data to a common model.

We present two case studies (i) publishing a suite of related information models using the Water Data Transfer Format (WDTF) schema at the Australian Bureau of Meteorology and (ii) the use of information models for harvesting content in the Spatial Identifier Reference Framework (SIRF).

In the WDTF case study, model publication is underpinned by transformation of UML models to Web Ontology Language (OWL) ontologies based on a (draft) ISO standard. The models are published in a Feature Type Catalog (FTC) delivered through a RESTful interface. The FTC is implemented as a Linked Data application in which model elements are identified using URIs, with content negotiation to access HTML or OWL/RDF forms.

In the SIRF case study we describe how models and mapping between models are used to support transformation and integration of data, and are then published together with the integrated data using Linked Data approaches.

Keywords: *Information models, metadata, UML, OWL, Semantic Web, Linked Data*

1. INTRODUCTION

An information model (hereinafter "model") is used to define and represent concepts and relationships of a 'domain discourse' i.e. the things that a particular community cares about and for which shared definitions are sought. Through use of a common language stakeholders working in the domain are able to reach a shared understanding a domain and enable the development of information systems that meet the needs of the community.

Various formalizations have been used for information models. In the contemporary era, Entity-Relationship diagrams provide a direct link to relational database schemas (Chen, 1976) The Unified Modeling Language (UML) (Booch et al., 2000) is another frame-based approach which includes a number of features associated with object-orientation. More recently the Web Ontology Language (OWL) (W3C, 2004, 2012) has its roots in descriptions logics, in which inferences may be drawn from the declared semantics using logical reasoning tools. OWL is based on Resource Description Framework (RDF) which is compatible with the Semantic Web. RDF is used in Linked Data implementations.

A model-driven approach seeks to derive system artefacts such as schemas, interfaces and vocabularies automatically from the conceptual representation of the system. Systems may be Modelled in different ways at different levels of abstraction. A level of abstraction refers to the amount of detail captured in a model and to a particular implementation. The level of abstraction of a model may range from: (i) definitions of the underlying patterns in modelling (meta-models); (ii) definitions of concepts; and (iii) platform-specific implementation specifications. The overall challenge is to find the balance of efficiencies in automating production of system artefacts against the cost of effort in modelling and model maintenance.

In this paper we address some key concerns of information modelling and its application in some case studies. In Section 2, we outline how we can use information models for data exchange, reuse and publishing. In Section 3, we present the advantages that can be gained from leveraging the ISO/TC211 standard information models in the context of two projects: (i) Spatial Information Reference Framework (SIRF); and (ii) the Tools and Documentation (TnD) and Sustainable Water Information Models (SWIM) projects which were undertaken under the Water Information Research and Development Alliance (WIRADA) between the Bureau of Meteorology ("the Bureau") and CSIRO. Section 4 provides a discussion of a key modelling design issue related to the appropriate level of abstraction to be used in information modelling weighed against the cost of implementation. Section 5 provides some conclusions.

2. PRINCIPAL CONCERNS

2.1. Information modelling for data exchange

The emergence of the Web as a data exchange platform has led to the development of various proprietary and open standards based "Web services" for the exchange of data. In the case of geospatial data, ISO Technical Committee 211 and the Open Geospatial Consortium (OGC) have developed a suite of standards to support data exchange and services. Use of these standards is often mandated as part of government policy through Spatial Data Infrastructure (SDI) activities, such as INSPIRE (European Commission, 2007). Hence geographic information modelling is of interest to stakeholders in many domains.

The underlying meta-model for geographic information modelling is described in *ISO 19101 Reference Model* and *ISO 19109 Rules for Application Schema*. These documents define the General Feature Model (GFM**)** for data structures, which defines the semantics of key structures and relationships for geographic information exchange. Geography Markup Language (GML) provides an XML schema to represent (or describe) geographical features defined using the meta-model (Portele, 2007).

UML was adopted to describe the models in ISO 19100 standards, and this innovation allowed the models to be used as machine readable artefacts to directly generate GML schema (the FullMoon software package performs this function - `http://projects.arcs.org.au/trac/fullmoon/wiki/FullMoon`). Typically, using this method, once the GML schema has been generated, the UML model is cataloged and plays no further role in operational systems. Although the catalogued model could be revised to create new versions access, version management and the reality of distributed governance of models, pose some challenges.

The Web Ontology Language (OWL) uses RDF to express semantics of data. OWL/RDF may express both the semantics of the structural elements and potentially also semantics of the content and terminology that is used to populate such structures. Since OWL is also based on classes and properties it may be used as an alternative representation for models designed originally in UML. OWL/RDF uses URIs to denote all classes, properties and individuals, so has advantages for global referencing through the web, in contrast to UML

which uses local identifiers (see Section 4). Note, however, that RDF-based technologies are not generally recommended for exchange of data that primarily contains numbers and arrays, such as geometries and grids.

2.2. Modularity and re-use

A concept may be referenced (referred to) and represented in multiple datasets. For example, a water body may be represented by a representative point, or by multiple polygons at different scales or a *stream* is referred to in the context of descriptions of stream flow measurements, and how a given stream is hydrologically connected with another. From both a data governance and usage perspective, there is a requirement to be able to determine if and when the same concept is being used. This can be addressed by either: (i) Linking related models (a "mapping model"), or (ii) Re-using common concepts (importing shared common domain models)

Either approach implies the reuse of models in the context of a larger system. In this context, the degree of modularity of a model (i.e. the extent to which complex domain models are constructed of smaller modules that may be reused in different domain models) has several implications for model use in practice. First, tools can be built to support common patterns such as: Simple Features (typical GIS systems) [Herring, 2011]; Observations and Measurements [Cox, 2011]; metadata records. Second, relationships between data can be discovered by looking at relationships between models: mapping or inheritance can be used to determine that different data products implement views of the same real world concepts. From this, Linked Data implementations (approaches to publishing and connecting structured data on the Web) can be generated directly from the model space. This forms the basis of the SIRF project described in Section 3.2.

2.3. Registration

There are a number of challenges related to modularity and consistency in domain modelling. In order to manage this, we have developed a toolset and registry application, called Solid Ground (Atkinson et al., 2010). Key principles that underpin the Solid Ground modelling methods are: (i) modular design of Application Schemas based on the GFM; (ii) re-use of common concepts through importing packages; (iii) management and access to models; and (iv) processing information models to generate artefacts useful in other systems.

2.4. Provenance as Living Metadata

Provenance is a key concern for any information generated by observation or modelling processes. In particular, transformation as part of the process of configuring analytical models relies on the interpretation of the semantics of the data. The meaning of data can be captured using information models, and standard information models allow such provenance metadata to be processed, for example, to discover which data may need to be re-analysed in the light of significant updates to source data.

2.5. Identity and Linked Data

At the heart of provenance and analytical processing is an understanding of what exactly is being processed. This often boils down to an issue of identity: what dataset, what monitoring site, what parameter, what interpolation method, what quality control status is being used? Interpreting data or repeating analysis is dependent on being able to get information describing a particular object. The ability to "de-reference" an identifier (retrieve a representation of an object) is therefore a critical requirement. 'Linked Data' refers to a set of best practices for publishing and connecting structured data on the Web which meets this requirement.

3. CASE STUDIES

In this section we present several applications of Linked Data to support access to model-derived artefacts from identifiers of modelled concepts. Linked Data relies on the HTTP protocol, which in turn uses MIME media types to describe content. However, this does not appear to provide the level of detail required for many applications so this approach will be limited unless various extensions are provided (Jain et al., 2009). In the following examples information models are provided along with the data they describe, within the same environment, as definitions published by the Feature Type Catalog are also a form of Linked Data.

3.1. "Tools and Documentation" for the Water Data Transfer Format standard

Information Model Derived artefacts for WDTF - The Water Data Transfer Format (WDTF) (Walker et al., 2009) is the standard XML format for data ingest into the Bureau's water data system. Implicit in the WDTF XML format is the WDTF information model which captures a set of concepts (including feature

definitions), relationships, controlled vocabularies and rules regarding combinations of data. A WDTF information model has been formalized and has been subsequently registered and then published to deliver model elements including feature definitions, relationships, and controlled vocabularies to the respective system components. Thus, formalizing the WDTF information model allows it to be more flexibly reused as a reference artifact and for deriving artefacts for WDTF compared to the standalone WDTF XML format.

The WDTF information model has been used for the WDTF documentation system to manage the connection to other documentation products. This includes PDF documents. It is also used in product generation systems, including a system that generates an OWL encoding of the information model. In addition, the WDTF information model is used to specify data products by creating a "profile model". A profile model restricts the semantics of the WDTF information model without affecting the WDTF XML schema by specifying a set of applicable controlled vocabularies and business rule constraints than the original information model. This demonstrates modular model reuse as we are able to reuse common concepts, specialize certain components of the WDTF information model and specify a set of applicable vocabularies which constrains the content contained within a schema such as the WDTF XML format for delivering a particular data product.

Feature Type Catalog - The Feature Type Catalog (FTC) is an Application Programming Interface (API) implemented on top of the OWL encoding of an information model. Each Feature Type definition and each definition of attributes and relationships is given a unique Uniform Resource Identifier (URI). A Linked Data application provides a Web interface to this content. In addition, a convenience API is provided to access information needed for specific applications, for example, a list of all the properties of a Feature Type. An example use of the FTC is in the WDTF documentation system, which uses it to access the models for WDTF-based data products and allows for a unified view of the product specification via a Web interface.

3.2. Information model driven harvest for SIRF

Spatial Identifier Referencing Framework - The Spatial Identifier Reference Framework (SIRF) enables discovery and cross-referencing of geospatial references used in application domains (Atkinson et al., 2013). SIRF provides a suite of web service interfaces that enable applications to reference spatial data from multiple heterogeneous sources.

Different representations of a feature are aggregated using a URI. This requires provenance information for features from their capture to publication. In order to do this, this provenance metadata is not only maintained by the system but it is promoted as a first class resource. The metadata captured acts as configurations for run-time system processing of the data. A key feature of the SIRF data processing workflow is that all artefacts that define the data and their relationships are explicitly managed in the system. This workflow from Francis et al. (2013) is depicted in Figure 1.

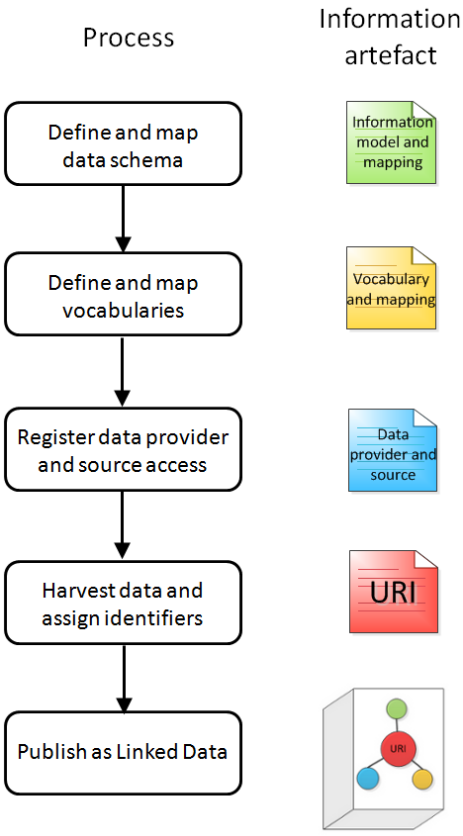

Figure 1 : SIRF Data Processing Workflow from Francis and Atkinson et al., (2013)

Model mappings - In order to index spatial references, the harvesting component captures spatial metadata from data provider services prior to publishing. A Web Feature Service presents the data structure as a GML schema, in compliance with the GFM (discussed in Section 2). An additional information model is generated in UML to describe the relationships that map the data provider schema to the common SIRF data schema. This mapping information model is then used as system metadata for the Extract-Transform-Load (ETL) process to index the data. The respective models and mappings are registered and managed,. The "Solid Ground" registry implementation exposes these interrelationships in a way in which UML data interchange standards do not support, and provides version control and distributed governance over the information models. These functions facilitate modularity and re-use.

Vocabularies used in the harvested dataset are also mapped. The SIRF vocabularies are delivered through the SISSVoc (`https://www.seegrid.csiro.au/wiki/Siss/SISSVoc`).

Linked data API extensions - Provenance metadata is denoted using persistent URIs in accordance with Linked Data approaches. SIRF supports content negotiation, discovery and cross-linking with other data sources using URIs. SIRF standard API extensions includes predictable patterns for finding not just other formats but other views on a referenced object, standardising basic object descriptions and listing of available resources.

4. CONCEPTUAL ABSTRACTION VERSUS AUTOMATION – A TRADEOFF?

In the above sections, we have demonstrated how information models can be used to address specific issues in two case studies using model-driven approaches. In these projects, model-driven artefacts have been produced to provide efficiencies, and also ensure that the relevant metadata are treated as first class citizens. Using the Linked Data paradigm, we have presented a framework for identity and access for content via the FTC implementation. In doing so, this paradigm simultaneously addresses a number of provenance challenges.

In designing information models for system use we may consider the work that information models do, and how can they work harder.. That is, how we can further leverage information models to allow for more consistency and efficiencies in information systems. However, there is a tradeoff in the effort spent on developing and refining information models and tooling for the management of its governance, maintenance, versioning, and use..What is the right balance of effort in generating and managing these models?

On the one hand we can aim for a high level of abstraction, perhaps aligning with fundamental models or 'upper ontologies' to ensure greater interoperability added rigour, alignment, consistency and completeness of the respective models and any other derived models such as the profile models discussed in this paper. This may better serve the various applications of information models in information system similar to ones presented. But the cost in doing so is in managing deep hierarchies as the starting point is likely to be a long way detached from implementation.

Alternatively, we could start with artefacts that are more concrete and closer to the implementation level, containing elements and statements of relationships that are ready to process by machines. These limit the need to traverse levels of abstraction upward and are typically less costly to implement. However, it is virtually impossible to compare similar or related models to determine semantic commonality, unless more effort is made in mapping back to the conceptual information models and how these are implemented in different systems. Using a more abstract model means the system implementer declares the semantic relationships between similar implementations, rather than leaving it to an end-user to locate and interpret heterogenous forms of documentation of multiple systems. Choosing the appropriate level of abstraction is part of the 'art of modelling'. The choice will be determined by a number of factors including: degree of interoperability being targeted, the scope of the modelling effort, the number of stakeholders, the complexity of the universe of discourse and expectations around reuse,

Two key features of the application of information models that allow both abstraction and automation, as presented in the SIRF and TnD project case studies, are: (i) transformation pathways - the ability to derive simple artefacts for which systems can easily be tailored for their use and (ii) aggregation through registration and Linked Data API extension for multiple views of the same conceptual identity.

In the TnD project the artefacts used for documentation were transformed from the original UML artefacts into RDF/OWL versions of the application schema and then published through the FTC. The FTC provides multiple views including the more direct conceptual view and the simplified property view. The simplified property view has the role of listing all inherited attributes of a class through all available levels of abstractions to provide a low abstraction information artefact. This is a specific example of a transformation pathway which negotiates the abstraction/automation trade-off to derive an artefact that can be more easily used by the respective systems.

In the SIRF project, an important feature of publishing references was the ability to publish provenance information for the point of harvest. This was achieved by specifying information models which describe data structures as an application schema (a native part of Web Feature Services), then building on to provide machine readable artefacts for the harvest. The Solid Ground mapping models captured the relevant data semantics from the point of harvest and mapped them to a common model. These mapping models were used to generate the ETL statement used by the harvest sub-component of SIRF, thereby providing the ability to aggregate the respective provenance information against a common model.

In both the TnD and SIRF project, model-driven artefacts are derived from abstractions captured in information models with the relevant metadata treated as first class citizens to facilitate automation at the system level. The key artefacts and their derivation pathways from the TnD and SIRF projects are presented against the level of machine readability and the level of abstraction in **Error! Reference source not found.**. The colours indicate the more conceptually abstract artefacts (green), directly machine readable artefacts (red) and the intermediate (yellow). This contrasts with traditional metadata statements to describe data structures that are conceptually flat and are not easily used in system processes. Therefore, we have demonstrated how we can increase the usefulness of metadata, specifically, the use of information models for deriving a number artefacts at the systems level using model-driven approaches to enable consistent use of semantics across those artefacts, reuse of system components, increase efficiencies in system processes and the aggregation of otherwise disparate information resources.

Figure 2 Level of conceptual abstraction against machine readability for artefacts hand-generated and derived in the General Feature Model context

5. CONCLUSIONS

In this paper, we have discussed the principal concerns of information modelling and its use in case studies of the WIRADA TnD and SIRF projects case studies. Information models may contain a rich representation of data semantics and concepts. Different levels of abstraction may be used to specify system behavior and to link different aspects of systems together. In changing the focus to capture and use of data concepts and relationships for system use, a more advantageous "provenance as living metadata" approach has been implemented. Using the case studies, we have shown the benefits of this approach and how we can further utilize information models to allow for more consistency and efficiencies in information systems.

In acknowledging the tradeoff in the effort spent on developing and refining information models and tooling for the management of its governance, maintenance, versioning, and use, we highlight two key features of our proposed application of information models in this paper that allow both abstraction and automation. First, we presented tooling and transformation pathways for providing the ability to derive simple artefacts from the information model, such as the use of a FTC to provide simplified property views for a given feature definition. Second, the ability to aggregate multiple views of a defined information resource with approaches the respective provenance information against a common model through registration and using Linked Data APIs.

This approach overlaps with important developments in machine readable metadata in the world of linked data, such as standard models for common patterns such as provenance or dimensional data. With the native use of RDF there is the opportunity of combining information modelling using ontology languages such as OWL with management of controlled vocabularies, thesauri and richer ontologies.The use of modelling in the WIRADA Tools and Documentation and SIRF projects provide practical examples of how further value can be gained from information models. This paper has briefly described how this can be achieved by

efficient transformation from conceptual information models to simple artefacts for system use, coupled with the use of URIs and Linked Data extensions to aggregate references to the conceptual identity.

ACKNOWLEDGMENTS

The cases studies projects described in this paper have been funded jointly by CSIRO and the Bureau of Meteorology (TnD project) under WIRADA and CSIRO and AusAID (for the SIRF project). The authors would like to acknowledge contributions from the reviewers and express their gratitude to project team members in CSIRO, Bureau of Meteorology for the TnD project and SIRF project partners, Badan Informasi Geospasial (the Indonesian Geospatial Information Agency – BIG.

REFERENCES

Atkinson R, B. P., Kostanski L. Spatial Identifier Reference Framework (SIRF)- A Case Study on How Spatial identifier data structures can be reoriented to suit present and future technology needs. In *Proceedings of the Pole to Pole - 26th International Cartographic Conference* (Dresden, Germany, 25-30 August, 2013)

Atkinson, R., Francis, W, Meng, R, Lemon D. *Tools and Techniques for Creation, Use and Management of Information Models as Metadata for System-of-Systems Interoperability*. CSIRO, 2010.

Chen P, 1976. The Entity-Relationship Model--Toward A Unified View Of Data. In: *Acm Transactions On Database Systems* 1/1/1976 Acm-Press Issn 0362-5915, S. 9–36 http://csc.lsu.edu/news/erd.pdf

Cox, S.J.D., 2011 OGC Observations and Measurements v2.0. Open Geospatial Consortium Abstract Specification Topic 20, http://www.opengeospatial.org/standards/om (also published as ISO 19156:2011).

European Commission, *2/EC of the European Parliament and of the Council of 14 March 2007, establishing an Infrastructure for Spatial information in the European Community (INSPIRE).* Official Journal of European Union, 2007

Francis W., Atkinson R., Box P., Rankine T., Woodman S., Kostanski L., 2013 K-Cap '13, June 23-26, 2013, Banff, Canada

Grady Booch, Ivar Jacobson & James Rumbaugh (2000) OMG Unified Modelling Language Specification, Version 1.3 First Edition: March 2000. http://www.omg.org/docs/formal/00-03-01.pdf Retrieved 12 August 2008.

Hartig O, Z. J., 2010 Publishing and consuming provenance metadata on the web of linked data. *Lecture Notes in Computer Science: Provenance and Annotation of Data and Processes*, (v63782010), 78-90.

Herring, J.R., 2011 OGC Simple Feature Access – Open Geospatial Consortium Implementation Standard, http://www.opengeospatial.org/standards/sfa

Jain, P., Hitzler, P., Yeh, P., Verma, K. and Sheth, 2009. A. Linked Data is Merely More Data. *Association for the Advancement of Artificial Intelligence*, (http://knoesis.wright.edu/library/publications/linkedai2010_submission_13.pdf).

Portele, C. (2007). ISO 19136:2007 Geographic information - Geography Markup Language (GML). (also published as OGC 07-036 http://www.opengeospatial.org/standards/gml).

Walker, G., Taylor, P., Cox, S., Sheahan, P., Anderssen, R., Braddock, R., & Newham, L. (2009, July). Water Data Transfer Format (WDTF): Guiding principles, technical challenges and the future. In *18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers, in Simulation* (pp. 2377-2383).

World Wide Web Consortium, OWL Web Ontology Language Overview, February 2004, available at http://www.w3.org/TR/owl-features/.

World Wide Web Consortium, OWL 2 Web Ontology Language Document Overview, December 2012, available at http://www.w3.org/TR/owl2-overview/.

W. Zhao , J.K. Liu OWL/SWRL representation methodology for EXPRESS-driven product information model: Part II: Practice Computers in Industry, Volume 59, Issue 6, August 2008, Pages 590–600