

Data Ingestion for Water e-Research

Kerry Taylor¹, Yanfeng Shu², Li Li¹

¹CSIRO, ICT Centre, Canberra, Australia, {kerry.taylor, lily.li}@csiro.au

²CSIRO, Tasmania ICT Centre, Hobart, Australia, {yanfeng.shu}@csiro.au

INTRODUCTION

It is very frequently recognised that many of the future scientific breakthroughs arise from cross-disciplinary research. One of the major hindrances to cross-disciplinary research is the effort required to collect and standardise data for study. In order for the study conclusions to be verifiable, it is desirable that the data processing is transparent and repeatable. Further, when the study involves repeated periodic data collection and processing, it is important that previous processing techniques are reusable and evolvable. In this paper we propose a suitable technique and supporting tools, making extensive use of semantic web technology for contextual knowledge representation and processing.

The Bureau of Meteorology is required to manage and hold Australia's Water Information under the Water Act 2007. This information is necessary for policy development, planning and enforcement to manage our scarce water resources, and will provide a resource for water e-research. It requires the Bureau to collect data from over 240 independent organisations across Australia as listed in the Water Regulations 2008. CSIRO and the Bureau are developing a uniform water data transfer standard (WDTF) [1] to be used for transfer from data providers. There will be an ongoing need for tools to manage transformations to and from WDTF relative to existing tools and the currently 240+ file structures.

The Regulations specify the types of water information required by the Bureau. There are ten categories, with each category further defined by subcategories. Each subcategory corresponds to an observation related to a feature (e.g., watercourse, reservoir); each observation is related to a measurement, and each measurement is related to a unit. These relationships among features, observations, measurements, and units are not captured in WDTF/XML, although a schematron validator is proposed. Data can be encoded in WDTF/XML that does not comply with the Regulations.

Traditionally, data translation involves schema matching (data fields are matched to elements of WDTF) and mapping (rules are specified to translate data instances into WDTF). Although many existing Extract-Transform-Load (ETL) tools include these two steps, they are not effective where source data contains only part of the information required by WDTF, and where the data we have is described quite differently in each source. Manual programming steps are required to insert the missing information into the processing tools: this is difficult for non-programmers and opaque to scrutiny. Instead, we develop a tool for translating data into WDTF by means of domain knowledge (represented as a thesaurus and a domain ontology), ensuring the translated data complies with the Regulations (represented in the domain ontology).

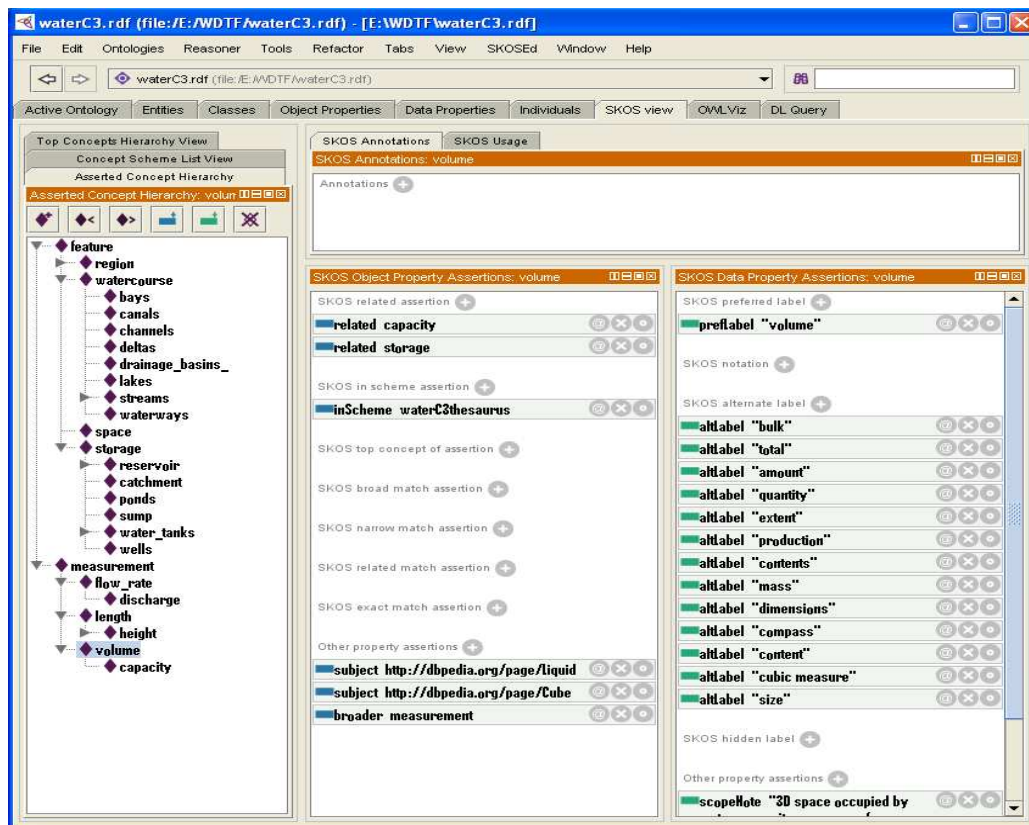


Figure 1: A Domain Thesaurus

DEVELOPING A DOMAIN THESAURUS

The thesaurus covers of two aspects of water storage vocabularies: (1) Measurement, including volume, level, and capacity; and (2) Feature, including watercourse, region, and storage structure. Synonyms and relations between these terms are specified using the emerging SKOS standard. The thesaurus is constructed by review of a collection of expert documents from hydrology, water resources, water management and geography domains: we produce separate SKOS files for each resource and then combine them with the SKOS Editor [2] within Protégé (Figure 1). We retain the reference to the source of each term in SKOS for provenance.

SCHEMA MATCHING AND MAPPING USING A DOMAIN ONTOLOGY AND A THESAURUS

We develop an ontology represented in OWL 2.0 to act as a conceptual model of the information in WDTF and the Regulations, and use its structure, together with the thesaurus, to assist in matching data fields with WDTF elements. During the matching, data providers may be asked to provide more information required by WDTF that cannot be gleaned from the ontology, matching, or provided data. As the ontology indicates which information is required, it can be used to guide an interactive information elicitation process.

After matching, rules are specified regarding how data instances are translated into the ontology concepts of WDTF. Again, the ontology acts as a medium between data from organisations and WDTF. Schema mapping involves the conversion of different units, the conversion of different measurements, and creation of new values for those WDTF elements which are keys or foreign keys or are both not nullable and not optional. More detail is given in [3].

SEMANTIC SERVICE ARCHITECTURE FOR MATCHING, MAPPING AND TRANSLATION

We propose to embed the mappings in a mature tool for Semantic Service Architectures (SSA) that has been used for semantic service composition [4]. In its usual application, the tool offers a domain ontology for users to specify service compositions in the language of the ontology. Pre-configured logical mappings [5] between the ontology and the information services (including databases and web services) are interpreted according to the user specification to generate an executable workflow over the information services.

For the data ingestion problem, we propose to enhance the technology by integrating the matching and mapping tools, and to configure the SSA with the original data files, domain ontology and mappings as they are developed. This provides an execution environment for the mappings and data translation process. The ontology may be used as an interactive query mechanism over the source data for inspecting aspects of the data/ontology relationships; the representation of the mappings is available both visually and in a format aligned with the Rule Interchange Format (RIF). With the addition of a script to generate the WDTF XML format from the ontology, the SSA will also provide translation to WDTF that is easily maintainable as the WDTF evolves, independently of the mappings between source data and the conceptual model.

CONCLUSION

Our method exploits domain knowledge in the form of a thesaurus and a domain ontology to translate data between messy file formats and a conceptual model of the required information. The thesaurus and domain ontology are independent models of the water regulations and water terminology respectively and are reusable in other applications. By virtue of the ontology structure, the method ensures that the translated WDTF data complies with the Regulations. The rule-language mappings between files and the domain ontology serve as convenient documentation of the relationships between source files and the WDTF artefacts. This may be useful when studying the impact of proposed changes to WDTF, or in explaining the intention of WDTF elements to those responsible for producing WDTF natively over time. Throughout the data processing lifecycle, the sources of knowledge and data are maintained explicitly as provenance to support both verification and evolution.

Acknowledgement

Thank you to Laurent Lefort for help with the thesaurus development and to the SSA research team, especially Mark Cameron, David Ratcliffe, Geoffrey Squire and Bella Robinson. Thank you to Gavin Walker for assisting with WDTF interpretation, and Bryn Kingsford and Paul Sheahan for providing datasets.

REFERENCES

1. G. Walker, P. Taylor, S. Cox, and P. Sheahan, *Interim-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. July 2009. http://www.mssanz.org.au/modsim09/J4/walker_g.pdf. pp. 4381–4387.
2. *The SKOS Editor*. Available from: http://protegewiki.stanford.edu/index.php/SKOS_Editor, accessed 29 June 2009.
3. Y. Shu, J. Wu, K. Taylor, R. Ackland, and A. Terhorst, *Towards Semantic Water Data Translation through a Knowledge-driven Approach*. Poster submitted to ISWC 2009.
4. M. Cameron, J. Wu, K. Taylor, D. Ratcliffe, G. Squire and J. Colton, *Semantic Solutions for Integration of Ocean Observations*. Submitted to 2nd International Semantic Sensor Networks Workshop 2009.
5. M. Cameron and K. Taylor, *First-order patterns for information integration*. In *Proceedings of ICWE 2005*, http://dx.doi.org/10.1007/11531371_25 pp. 173-184.