Enabling combined software and data engineering: the ALIGNED suite of ontologies

Monika Solanki[1], Bojan Bozic[2], Markus Freudenberg[3], Dimitris Kontokostas[3], Rob Brennan[2], and Christian Dirschl[4]

[1] Department of Computer Science, University of Oxford, UK
[2] KDEG, School of Computer Science and Statistics, Trinity College Dublin, Ireland
[3] AKSW/KILT, University of Leipzig, Germany
[4] Wolters Kluwer, Germany

Abstract. Effective, collaborative integration of software and big data engineering for Web-scale systems, is now a crucial technical and economic challenge. This requires new combined data and software engineering processes and tools. Semantic metadata standards and linked data principles, provide a technical grounding for such integrated systems given an appropriate model of the domain. In this paper we introduce the ALIGNED suite of ontologies specifically designed to model the information exchange needs of combined software and data engineering. The models have been deployed to enable: tool-chain integration, such as the exchange of data quality reports; cross-domain communication, such as interlinked data and software unit testing; mediation of the system design process through the capture of design intents and as a source of context for model-driven software engineering processes. These ontologies are deployed in web-scale, data-intensive, system development environments in both the commercial and academic domains. We exemplify the usage of the suite on a complex collaborative software and data engineering scenario from the legal information system domain.

1 Introduction

This paper has been accepted in the ISWC 2016 Resources track. Recent years have seen a significant increase in the demand for data-intensive applications. However our engineering techniques for building data-intensive systems are both immature and often partitioned into software engineering and data engineering processes, tasks or teams. The expressivity of semantic models makes them useful for both addressing data quality [2] and applying model-driven approaches to software engineering. Semantic data, in the form of enterprise linked data is also useful for describing, fusing and managing the combined data and software engineering lifecycles to increase productivity, agility and system quality.

In this paper, we present a suite of ontologies developed within the ALIGNED[5] project, that aim to align the divergent processes encapsulating data and software engineering. The key aim of the ALIGNED ontology suite is to support

[5] `http://aligned-project.eu`

the generation of combined software and data engineering processes and tools for improved productivity, agility and quality. The suite contains linked data ontologies/vocabularies designed to: (1) support semantics-based model driven software engineering, by documenting additional system context and constraints for RDF-based data or knowledge models in the form of design intents, software lifecycle specifications and data lifecycle specifications; (2) support data quality engineering techniques, by documenting data curation tasks, roles, datasets, workflows and data quality reports at each data lifecycle stage in a data intensive system; and (3) support the development of tools for unified views of software and data engineering processes and software/data test case interlinking, by providing the basis for enterprise linked data describing software and data engineering activities (tasks), agents (actors) and entities (artefacts) based on the W3C provenance ontology[6].

This ontology suite has been deployed for validation and incremental improvement in the ALIGNED project on four, large-scale data-intensive systems engineering use cases: the Seshat Global History Databank which is compiling linked data time series relating to all human societies over the past 12,000 years; JURION[7], a legal information platform developed by Wolters Kluwer Germany; PoolParty[8], a semantic technology middleware developed by the Semantic Web Company; and the DBpedia+[9] data quality and release processes.

The paper is structured as follows: Section 2 presents an overview of the ALIGNED suite. It provides a brief description of the core ontologies in the suite.Section 3 presents an evaluation of the ontologies in the suite. Finally, Section 4 presents conclusions.

2 Overview of the ALIGNED suite

Figure 1 illustrates the ALIGNED suite of ontologies split into the provenance, generic, and domain-specific layers. As can be seen from the figure, a high emphasis has been placed on reusing existing, well known and standardised specifications where available. At the top layer, the W3C provenance standard forms the baseline for all our specifications and all our models extend it in some way. The split of the ALIGNED ontology suite between a generic layer and a domain specific extensions layer allows rapid evolution of domain-specific extensions for the ALIGNED use cases/trial environments (JURION, Seshat, DBpedia, PoolParty) based on a stable set of core concepts modelled in the generic layer. As the project progresses these extensions will be evaluated and incorporated into the generic layer if they prove valuable or more widely applicable than a single domain. Within the project the suite of ontologies is known as the "ALIGNED metamodel" due to the links with software engineering practices.

[6] `http://www.w3.org/ns/prov-o`
[7] `https://www.jurion.de/`
[8] `https://www.poolparty.biz/`
[9] `http://wiki.dbpedia.org/`

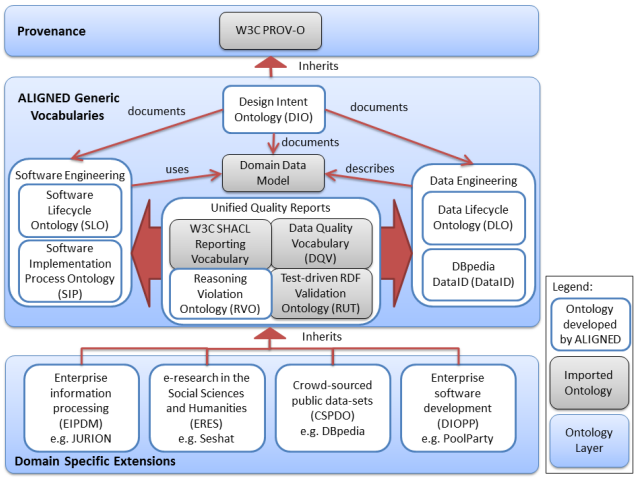

Fig. 1. The ALIGNED Suite of Ontologies

3 Evaluation

Table 1 presents the evaluation of the ALIGNED suite in accordance to the desired criteria [10].

4 Conclusions

Combining data and software engineering processes to increase productivity and agility, is a challenge being faced by several organisations aiming to exploit the benefits of big data. Ontologies and vocabularies developed in accordance to competency questions, objective criteria and ontology engineering principles can provide useful support to data scientists and software engineers undertaking the challenge. In this paper we have proposed the ALIGNED suite of ontologies that provide semantic models of design intents, domain specific datasets, software engineering processes, quality heuristics and error handling mechanisms. The suite contributes immensely towards enabling interoperability and alleviating some of the complexities involved. We have exemplified the usage of the suite on a real-world use case from the legal domain and evaluated it against the desired criteria. As ontologies from the suite are now in various stages of adoption by the ALIGNED use cases, the next steps would incorporate their empirical evaluation.

[10] `https://figshare.com/articles/ISWC2016_Resources_Track_Review_Instructions/2016852`

Generic criteria	Evaluation
Value Addition	(1) The ontologies add data and software engineering specific metadata to the process and enrich information about process specific procedures within data and software engineering for a tool, which in return can use this context dependent information for automation and automatic generation purposes. (2) DLO is used to provide details about the data engineering process and SLO details about the software engineering process. (3) RVO helps producing information about reasoning errors in the knowledge base, while DIO enables the mining of design intents from requirements specification as well as the generation of unified governnce reports by integrating requiremebnts and design issues.
Reuse	(1) Potential reuse across a wider community of content producers, owners of large amounts of data, data managers, ontology engineers of new related ontologies and vocabularies (2) Software development model designers, and developers of human societies datasets (e.g. Seshat Global History Databank). (3) The metamodels are easy to reuse and published on the Web together with detailed documentation. Top level models are general and can be applied for all data and software engineering models. Furthermore, the models are extendable and can be inherited by specialised domain ontologies for specific software and data engineering platforms.
Design and Technical quality	All ontologies have been designed as OWL DL ontologies, in accordance to ontology engineeirng principles. Axiomatisations in the ontologies have been defined based on the competency questions identified during requirements scoping.
Sustainability	All ontologies are deployed on a public Github repositories. Long term sustainability has been assured by the ontology engineers involved in the design.
Specific criteria	
Design suitability	Individual ontologies in the suite have been developed in close association with the requirements emerging from corresponding, potential exploiting application.Thus they closely conform to the suitability of the tasks for which they have been designed.
Design elegance and quality	Axiomatisation in the ontologies have been developed following Gruber's principles [1] of clarity, coherence, extendability, minimum encoding bias and minimum ontological commitment.
Logical correctness	The ontologies have been verified using DL reasoners for satisfiability, incoherency and inconsistencies. Specifically, inconsistencies for DIO has been checked against the instance data in the governance triple store.
External resources reuse	External ontologies such as PROV-O, SKOS have been extensively used.
Documentation	The ALIGNED public deliverables and publications [3, 4] include detailed descriptions of the models. The ontologies have been well documented using rdfs:label and rdfs:comment. HTML documentation via the LODE service has also been enabled. All ontologies have been graphically illustrated.

Table 1. Evaluating the ALIGNED suite of Ontologies

Acknowledgement

This research has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 644055, the ALIGNED project (www.aligned-project.eu).

References

1. T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, Dec. 1995.
2. D. Kontokostas, M. Brümmer, S. Hellmann, J. Lehmann, and L. Ioannidis. Nlp data cleansing based on linguistic ontology constraints. In *ESWC 2014*, 2014.
3. D. Kontokostas, C. Mader, C. Dirschl, K. Eck, M. Leuthold, J. Lehmann, and S. Hellmann. Semantically enhanced quality assurance in the jurion business use case. In *ESWC 2016 (to appear)*, 2016.
4. M. Solanki. DIO: A pattern for capturing the intents underlying designs. In *Proceedings of the 6th Workshop on Ontology and Semantic Web Patterns (WOP 2015)*, volume Vol-1461. CEUR-WS.org, 2015.