Enabling combined Software and Data engineering at Web-scale: The ALIGNED suite of Ontologies

Resource type: Set of ontologies
Permalink: https://github.com/aligned-h2020/ALIGNED_Ontologies

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

[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. 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

Recent years have seen a significant increase in the demand for data-intensive applications based on large-scale sources of data. 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. There is a need for integrated engineering approaches. The data itself must also be high-quality, which entails a curatorial process to improve and manage data over time. The expressivity of semantic models makes them useful for both addressing data quality [5] and applying model-driven approaches [3] 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 [10], 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 shows how the vocabularies have been applied to a complex collaborative software and data engineering scenario from the legal information system domain. Section 4 presents an evaluation of the ontologies in the suite. Section 5 briefly discusses related work. Finally, Section 6 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

[6] `http://www.w3.org/ns/prov-o`

[7] `https://www.jurion.de/`

[8] `https://www.poolparty.biz/`

[9] `http://wiki.dbpedia.org/`

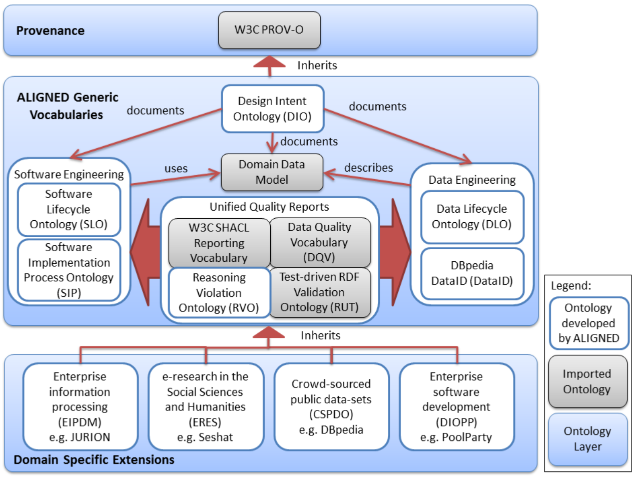
Provenance
W3C PROV-O
Inherits
ALIGNED Generic Vocabularies
documents
Design Intent Ontology (DIO)
documents
documents
uses
Domain Data Model
describes
Software Engineering
Software Lifecycle Ontology (SLO)
Software Implementation Process Ontology (SIP)
Unified Quality Reports
W3C SHACL Reporting Vocabulary
Data Quality Vocabulary (DQV)
Reasoning Violation Ontology (RVO)
Test-driven RDF Validation Ontology (RUT)
Data Engineering
Data Lifecycle Ontology (DLO)
DBpedia DataID (DataID)
Inherits
Enterprise information processing (EIPDM) e.g. JURION
e-research in the Social Sciences and Humanities (ERES) e.g. Seshat
Crowd-sourced public data-sets (CSPDO) e.g. DBpedia
Enterprise software development (DIOPP) e.g. PoolParty
Domain Specific Extensions
Legend:
Ontology developed by ALIGNED
Imported Ontology
Ontology Layer

Fig. 1. The ALIGNED Suite of Ontologies

domain. Within the project the suite of ontologies is known as the ”ALIGNED metamodel” due to the links with software engineering practices.

We briefly present here the core ontologies from the suite. Further details of the ontologies including the axiomatisations, graphical representation, serialisations in multiple formats via content negotiation, examples illustrating the usage of the ontologies, typical SPARQL queries that can be formulated using the ontologies as the data model and HTML documentation are available from the individual deployments at their persistent URIs. Due to space constraints we deliberately do not include these in this paper. The ontologies are grouped as follows:

- **Design intent**: This model is used to document the design decisions about data intensive system artefacts such as requirements, designs or datasets. It is based on the design intent ontology (DIO)[10], which allows users to express the design intent or design rationale while undertaking the design of an artefact. DIO [9] is a generic ontology that provides the conceptualisation needed to capture the knowledge generated during various phases of the overall design lifecycle. DIO provides definitions for design artefacts such as requirements, designs, design issues, solutions, justifications and evidence, and relationships between them.
- **Software engineering**: This model defines the major agents (e.g. project roles), activities (e.g. lifecycle stages), and entities (design artefacts) involved in a software engineering project and their relations with a special focus on capturing the engineering lifecycle. Two ontologies make up this model: the

[10] https://w3id.org/dio

software process ontology (SPO)[11] and the software implementation processes ontology (SIP)[12].

- **Data engineering**: As software engineering above but with a focus on data engineering and data lifecycles. Two ontologies are used: the data lifecycle ontology (DLO)[13] defined within ALIGNED and the DataID[14] ontology, defined by ALIGNED for the DBpedia association, for describing datasets. DLO provides a set of conceptual entities, agents, activities, and roles to represent the general data engineering process. Furthermore, it is the basis for deriving specific domain ontologies which represent lifecycles of concrete data engineering projects such as DBpedia or Seshat. DataID is a multi-layered meta-data system, which, in its core, describes datasets and their different manifestations, as well as relations to agents like persons or organisations, in regard to their rights and responsibilities. Depending on context, type of data and use case, this core ontology can be augmented by multiple existing extensions (e.g. Linked Data, repository descriptions etc.).
- **Unified quality reports**: Defines a unified reporting representation for data quality metrics, ontology reasoning errors, test cases, and test case results based on the W3C SHACL reporting vocabulary. It is based on four ontologies/vocabularies, three of which are externally developed: W3C SHACL[15], W3C Data Quality[16], and University of Leipzigs test-driven RDF validation ontology [5] (RUT); and one ontology developed within ALIGNED: the reasoning violation ontology (RVO)[17]. RUT is designed to capture the lifecycle of RDF validation with the test driven validation methodology. It is implemented by the RDFUnit tool. RVO describes both ABox and TBox reasoning errors for the integration of reasoners into data lifecycle tool-chains. The ontology covers violations of the OWL 2 direct semantics and syntax detected on both the schema and instance level over the full range of OWL 2 and RDFS language constructs. An overview of RVO and its design, implementation and use cases has been published in [1].
- **Domain data model**: This describes the domain of the data-intensive application being developed and is specific to that application, e.g. the Seshat ontology for historical time-series describing human societies. The lower layer includes the domain-specific extensions to the metamodels. ALIGNED has developed four domain-specific metamodels based on each of our use cases, with a focus on model elements needed for the ALIGNED phase 2 trials.
- **Enterprise information processing**: extensions and models for the JURION use case.
- **E-research in the Social Sciences and Humanities**: extensions and models for the Seshat use case.

[11] `https://w3id.org/slo`

[12] `https://w3id.org/sip`

[13] `https://w3id.org/dlo`

[14] `http://dataid.dbpedia.org/ns/core#`

[15] `https://www.w3.org/TR/shacl/`

[16] `https://www.w3.org/TR/vocab-dqv/`

[17] `https://w3id.org/rvo`

- **Crowd-sourced public datasets**: extensions and models for the DBpedia use case.
- **Enterprise software development**: extensions and models for the PoolParty use case.

3 Example deployment: the ALIGNED suite in Wolters Kluwer's JURION

JURION is an innovative legal information platform developed by Wolters Kluwer Germany that merges and interlinks over 1 million documents of content and data from diverse sources such as national and European legislation and court judgements, extensive internally authored content and local customer data, as well as social media and semantic web data (e.g from DBpedia). This data is then presented to users (such as law offices) in the form of highly customised applications for semantic search, annotation, case management and legal information retrieval.

Currently, the software development process and data life cycle are highly independent from each other. Figure 2 illustrates where ontologies from the ALIGNED suite contribute towards facilitating interoperability between the software and data engineering processes and tools used to build and maintain JURION. The two main uses are tool integration and unified governance. Tool integration includes both cases within a single domain (data or software engineering) and cross-domain tool-chain integration. Unified governance uses ALIGNED provenance records, data extraction and uplift from enterprise engineering tools and data fusion to provide end to end and cross-domain views of the JURION platform engineering processes. We elaborate on the deployment of ALIGNED ontologies for these use cases below.

RUT has been used in JURION for validating & verifying the extraction of metadata [6]. In particular, RDFUnit is used as a data validation tool integrated in JURION's continuous integration (CI) platform (Jenkins). RVO, the reasoning violations ontology, has been used to integrate advanced OWL reasoning-based data quality checks with RDFUnit's triple-query oriented tests to expand the scope of testing possible. DataID descriptors of all the JURION datasets are under evaluation and it is planned to use this to provide consistent meta-data which will be available to all tools thus facilitating further integration. The EIP, enterprise information processing, ontology has been used to describe the JURION environment, systems, artifacts and engineering processes in terms of the ALIGNED software and data lifecycle models.

An upcoming feature in JURION is the integration of search requirements with design issues/software bugs arising during their implementation. The goal is to express integrated requirements and issues as linked data, which is semantically annotated using the DIO and DIO-PP ontologies from the ALIGNED suite. This would further enable the development of customised Confluence interfaces which can be used to provide enhanced query features over the integrated data and produce bespoke reports using visual and statistical analytics.

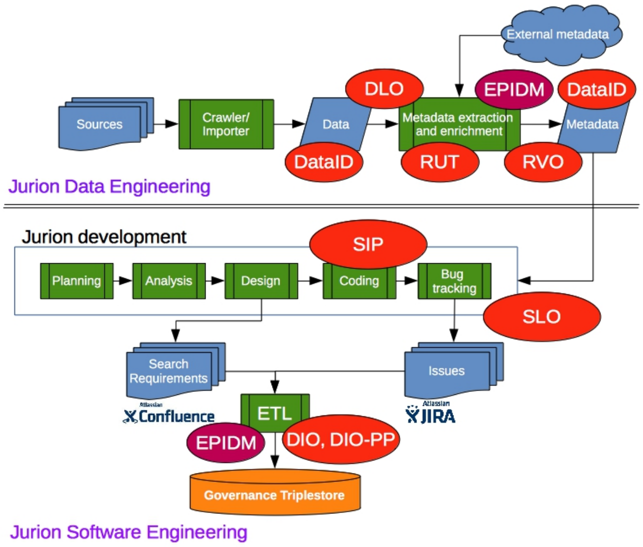

Fig. 2. Usage of the ALIGNED suite of ontologies in the JURION semantics-based legal information system

4 Evaluation

Table 1 presents the evaluation of the ALIGNED suite in accordance to the desired criteria [18].

5 Related work

SEON[20] is a family of ontologies that describe concepts in the context of software engineering, software evolution and software maintenance. SWO[21] is a resource for describing software tools, their types, tasks, versions and provenance. While they cover some general aspects of software engineering, they do not address the description of design intents and software lifecycles. Representing design intents or design rationales as ontologies have been captured for various specialised domains such as software engineering [2] however there is no generic, domain-independent design intent capture model available as a design pattern. OOPS! [8] is a tool with a catalogue for validating ontologies by spotting common pitfalls, however it detects design flaws rather than logical errors and does not use an ontology for error reporting. The DCAT vocabulary includes the special class Distribution for the representation of the available materialisations of

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

[20] http://se-on.org/#publications

[21] http://theswo.sourceforge.net/

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 governance reports by integrating requirements 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 engineering principles [7]. Axiomatisations in the ontologies have been defined based on the competency questions identified during requirements scoping.
Availability	Ontologies have been made publically available at `http://aligned-project.eu/data-and-models/`. Further, they have been given persistent w3id URIs, deployed on public facing servers and are content negotiable. DIO has been cited in [9] and RUT in [6]. All ontologies have been licensed under a Creative Commons Attribution License. DIO has also been registered[19] in LOV.
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 [4] 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 [6, 9] 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

a dataset. These distributions cannot be described further within DCAT. The Asset Description Metadata Schema[22] (ADMS) is a profile of DCAT, which only describes a specialised class of datasets: so-called Semantic Assets.

6 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

[22] `https://www.w3.org/TR/vocab-adms/`

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.

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. B. Bozic, R. Brennan, K. Feeney, and G. Mendel-Gleason. Describing reasoning results with rvo, the reasoning violations ontology. In *ESWC 2016 (to appear)*, 2016.
2. A. P. de Medeiros, D. Schwabe, and B. Feijó. Kuaba ontology: Design rationale representation and reuse in model-based designs. In *Proceedings of the 24th International Conference on Conceptual Modeling*, ER'05, Berlin, Heidelberg, 2005. Springer-Verlag.
3. D. Gasevic, D. Djuric, and V. Devedzic. *Model Driven Engineering and Ontology Development.* Springer Publishing Company, Incorporated, 2nd edition, 2009.
4. 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.
5. D. Kontokostas, M. Brümmer, S. Hellmann, J. Lehmann, and L. Ioannidis. Nlp data cleansing based on linguistic ontology constraints. In *ESWC 2014*, 2014.
6. 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.
7. N. F. Noy and D. L. Mcguinness. Ontology development 101: A guide to creating your first ontology. Technical report, Stanford Center for Biomedical Informatics Research (BMIR), 2001.
8. M. Poveda-Villalón, M. C. Suárez-Figueroa, and A. Gómez-Pérez. Validating ontologies with oops! In *Knowledge Engineering and Knowledge Management*, pages 267–281. Springer, 2012.
9. 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.
10. P. Turchin, R. Brennan, T. Currie, K. Feeney, P. Francois, D. Hoyer, J. Manning, A. Marciniak, D. Mullins, A. Palmisano, P. Peregrine, E. A. Turner, and H. Whitehouse. Seshat: The global history databank. *Cliodynamics*, 6(1):77–107, 2015.