The ALIGNED Project – Aligned, Quality-centric Software and Data Engineering Driven by Semantics

Odhrán Gavin[1], Dimitris Kontokostas[2], Christian Dirschl[3], Andreas Koller [4], Jim Davies [5], Pieter Francois [5/7], Arkadiusz Marciniak [6], Bojan Bozic[1], Gavin Mendel-Gleason[1], Kevin Feeney[1] and Rob Brennan[1]

[1] Trinity College Dublin (coordinator), [2] University of Leipzig, [3] Wolters Kluwer Germany, [4]Semantic Web Company, [5] University of Oxford, [6]Adam Mickiewicz University, [7]University of Hertfordshire
`info@aligned-project.eu`

Abstract. This paper describes the H2020 ALIGNED project (#644055) which investigates RDF-based data quality, semantic model-driven software engineering, enterprise linked data for software and data engineering process management and engineering of data intensive systems based on linked data.

1 The ALIGNED Project

The Horizon 2020 project ALIGNED[1], Aligned Quality-centric Software and Data Engineering, which started in February 2015 brings together computer science researchers (Trinity College Dublin, University of Oxford, University of Leipzig), software companies specialised in data-intensive systems (Semantic Web Company), information companies (Wolters Kluwer) and academic curators of the Seshat Global History Databank, large datasets describing world history and archaeology (University of Oxford, Adam Mickiewicz University in Poznań).

The last few 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 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. ALIGNED will tackle these challenges with five objectives:

1. A methodology for combined software and data engineering, based on a metamodel which describes the software and data lifecycles.

---

[1] http://www.aligned-project.eu

2. Tools to produce software development models from the metamodel, including transformations that generate or configure software applications.
3. Tools to produce data development models from the metamodel, incorporating data quality and integrity constraints, data curation, and data transformations.
4. Methods to use the metamodel and tools as part of a unified software and data engineering process, emphasising techniques which ensure data quality and integrity, as well as software security and reliability.
5. Evidence that the ALIGNED methodology and tools lead to greater development productivity and agility in enterprise and web scale data-intensive systems.

ALIGNED has four 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 [1]; JURION, a legal information platform developed by Wolters Kluwer Germany [2]; PoolParty, a semantic technology middleware developed by the Semantic Web Company; and DBpedia.

ALIGNED has already made significant progress towards fulfilling these goals. The underlying metamodel has progressed through two releases, and the ontologies which have been developed are available online[2]. The first ALIGNED tools[3] have already been deployed in the live use case environments: We have provided data validation which has significantly reduced the error rate of the Seshat data. External dataset quality checks have been integrated into JURION to improve ease of schema maintenance. Import validation and constraint violation checks features have been added to PoolParty. DBpedia have deployed our tools as part of the release process.

ALIGNED is looking for opportunities to engage with other projects in the linked data and software engineering fields. In particular, we would hope to collaborate with other projects which share our focus on data-intensive systems and the processes which they use to manage and maintain their data. *We could demonstrate some of our data quality and data curation tools (RDFUnit, PoolParty, Dacura) to the other projects at the session. We could also show videos of ALIGNED tools in action.*

References

1. Brennan R., Feeney K., Mendel-Gleason G., Bozic B., Turchin P., Whitehouse H., Francois P., Currie T. and Gohmann S. Building the Seshat Ontology for a Global History Databank. (Accepted) ESWC 2016.
2. Kontokostas D., Mader C., Dirschl C., Eck K., Leuthold M., Lehmann J., and Hellmann S. 2016. Semantically Enhanced Quality Assurance in the JURION Business Use Case. (Accepted) ESWC 2016.

---

[2] http://aligned-project.eu/data-and-models/
[3] http://aligned-project.eu/open-source-tools/