Bridging Semantic Heterogeneities in Open Source Software Development Projects with Semantic Web Technologies

Wikan Danar Sunindyo and Stefan Biffl (Faculty Mentor)
Institute of Software Technology and Interactive Systems
Vienna University of Technology
Vienna, Austria
Email: {wikan,biffl}@ifs.tuwien.ac.at

Abstract — *The semantic heterogeneity of Open Source Software (OSS) development projects comes from the using of different tools and models by the various stakeholders. These differences make the process of integration become difficult, since the project managers should recognize the different structure of the tools and models for analyzing the state of the projects. This manual analysis is costly and error prone. In this work we propose a semantic web technology approach to bridge these semantic heterogeneities, by using engineering knowledge base (EKB). The EKB enables mapping between local and domain ontology layers to allow querying the local tool knowledge using the domain-level knowledge and syntax. We empirically evaluate the feasibility of an EKB-based project monitoring system based on real-world data.*

I. INTRODUCTION

Stakeholders in open source software (OSS) usually come from heterogeneous backgrounds and use a range of tools and models for developing and managing their common software. For example, the developers may use SVN for managing source code versions and mailing lists for communication between developers that become the use case for this paper.

In the OSS development projects, the stakeholders often work globally distributed and in different time zones. Therefore, the project managers need to be able to manage and monitor the status of the project and how the people work, e.g., by monitoring the level of communication between the stakeholders. To address this goal, project managers often use a quantitative measurement approach for managing and monitoring the status of the projects, i.e., finding related issues in SVN and mailing list for a certain projects. Further analysis to identify related information across data sources gets conducted manually, which is costly, error-prone, and often takes too much time for the analysis results to be useful for decision making.

In this paper we propose the use of a semantic web technology approach called engineering knowledge base (EKB). The principle of EKB is using ontology to model the project/domain knowledge that are used and understood by all stakeholders and tools. Concepts in this model are then mapped to local tool data models, which model common domain concepts in different ways. Based on this mapping queries to the domain knowledge can be resolved via the mappings by local tool data queries.

We will check the feasibility of querying the ontologies in EKB written in SPARQL[1] format with the case from an open source project. The major result is the easier definition of queries on project data originating from heterogeneous background.

II. RESEARCH ISSUES

The core challenge for OSS development project managers is to get stable domain concepts but varying encoding of a concept in local tools. The current approach is by working on variety of local tool data model. The domain model stays "virtual", i.e., is not explicitly modeled. Our approach is by explicitly model the domain data model and map to tool models (similar to Global-as-View) by using EKB [1]. EKB is a repository that holds all knowledge on project data models and user information that is relevant for OSS development. EKB consists of two ontology layers. The first layer contains the common domain knowledge, while the second layer describes the local tool knowledge. Mappings between local and domain ontologies allow querying the local knowledge by using the domain ontology syntax.

The research issue is how to evaluate the feasibility of using EKB for bridging semantic heterogeneities comparing to the traditional database approach. We will use *Apache Tomcat*[2] project as a test bed.

III. EVALUATION

To give better understanding on the approach proposed, we will illustrate a project manager use case to analyze relationships across heterogeneous data

[1] http://www.w3.org/TR/rdf-sparql-query/
[2] http://tomcat.apache.org/

sources, for example between mailing list and SVN. The domain concept concepts consist of all concepts used in OSS development projects, the local tool concepts consist of the concepts used in mailing list and SVN tools. The mappings are done between interrelated local tool concepts and domain concepts, to integrate the different concepts across the tools/models. The queries from the project managers will be done to the domain concepts, which will be continue by resolution with the local tool concepts, in SVN and mailing lists ontologies.

Database approach. The using of database approach demands the use of a common homogeneous database. Hence there should be some efforts done to change the different models and formats of data from heterogeneous sources and stored in homogeneous database. Then the project manager can use the database and query the repository. There are several weaknesses with this approach: (a) the different structures from heterogeneous data sources need to be reconciled for using the common repository; (b) updates and synchronization between the data sources and the common repository have to be repeated each time new data entries update the local data source; and (c) reconfiguration of the data sources, e.g., adding new data sources to the monitoring system, breaks the automation of this approach.

EKB approach. The EKB approach demands the local tool concepts stored in the separate ontologies from the domain concepts ontology. The mapping of related concepts is done from the local tool ontology to the domain concept ontology, makes the translation between different local ontology possible. This approach is more suitable than the implicit transformation with the previous approach, because the project managers don't have to transform each different concept into homogeneous model and they only have to deal with the domain concept ontology to know further about the local tools ontology.

The example of query given by the project managers to find the relationships between mailing list and SVN ontologies can be seen in listing 1 that uses a simplified OWL[3] syntax. The data from different stakeholders are stored the SVN and mailing list ontologies, which are mapped to the domain concept ontology. The project managers make query to the domain concept ontology which then continues to the local ontologies. The result from the local ontologies is returned to the project manager via domain concept ontology.

From this evaluation, we can learn that EKB approach benefits on flexible and efficient model for transformation of data between tools. The limitations are the cost to design the EKB and performance of ontology technology which is depend on the size of the ontology.

```
SELECT (?a) WHERE {domain:build.xml
domain:relatedWith ?a}
mailinglist:build.xml owl:equalTo
domain:build.xml

SELECT (?b) WHERE
{mailinglist:build.xml mailing-
list:hasAffectedArtifact ?b}
Result: b = mailing-
list:build.xml_890256

mailinglist:build.xml_890256
owl:equalTo domain:build.xml_890256
SVN:build.xml_890256 owl:equalTo
domain:build.xml_890256

SELECT (?c) WHERE
{?c SVN:hasAffectedArtifact
SVN:build.xml_890256}
Result: c = SVN:SVN_890256_build.xml

SVN:SVN_890256_build.xml owl:equalTo
domain:SVN_890256_build.xml

Result: a = domain:SVN_890256.xml
```

Listing 1. EKB query example to find a related SVN entry from the mailing list issue.

IV. Conclusion and Future Work

In this paper, we have explained the EKB approach as a semantic web technology approach to address important aspects of the semantic heterogeneity challenge in the OSS development projects domain. By separating the level of ontologies in EKB, the complexity of querying can be reduced, while still makes the approach feasible to be done.

Future work includes make empirical measurement on the efforts needed, especially for adding new data sources to the project monitoring system. This could be useful for better decision making support on the status of the project, since a wider range of development knowledge sources in the project can be included for analysis.

References

[1] T. Moser, S. Biffl, W. D. Sunindyo, and D. Winkler. Integrating Production Automation Expert Knowledge Across Engineering Stakeholder Domains. In Proceedings of the 2010 International Conference on Complex, Intelligent and Software Intensive Systems (to be published), pages 8, Krakow, Poland, 2010.

[3] http://www.w3.org/TR/owl-semantics/