Semantic rule processing for real-time data integration

Pedro Lopes*, José Luís Oliveira

DETI/IEETA, University of Aveiro, Aveiro, Portugal
{pedrolopes,jlo}@ua.pt

Abstract. Integration-as-a-service platforms arise as a modern strategy to integrate data from distributed environments. Nowadays, service interoperability strategies, such as workflows and static service-oriented architectures, are giving place to more dynamic environments, where the path from the original resource to the integrative destination is triggered autonomously and in real-time. However, these concepts still have not been applied to bioinformatics, in great part due to the complexity underlying the data validation and transformation tasks. In this manuscript we introduce a component to enhance these activities by enabling the execution of complex pre- and post-integration semantic algorithms. By leveraging on comprehensive Semantic Web constructs, these activities are better suited to the life sciences domain.

Keywords: Semantic Web, data integration, real-time, application framework.

1 Introduction

Real-time data integration continues to be a challenge [1], particularly in the life sciences domain [2]. Whereas in other scenarios, such as mechanical engineering or embedded software, several solutions are in place, the automated integration of biomedical data has plenty room for innovation.

With the evolution of cloud-based technologies, integration-as-a-service ideals are now being used to describe new platforms that enable real-time automated integration of data and services [3].

More importantly, the Semantic Web's underlying flexibility and dynamics can be combined with integration-as-a-service strategies to reach a whole that is more than the sum of its parts. The complex Extract-Transform-Load (ETL) data warehousing workflow can be improved through the surgical inclusion of semantic-based components [4].

This work introduces such component, enhancing the integration workflow with new methods to pre- and post-process data in the integration pipeline [5, 6]. Data can be validated or transformed, according to a predefined set of customizable rules. Simpler validation examples include regular expression to match text content or Boolean arithmetic expressions to evaluate numeric values. At a more complex level, inference and reasoning can be used to transform content before integration.

2 Methods

The semantic rule-processing component will interact directly with the ETL engine in the integration pipeline - Figure 1. The basic integration workflow is comprised by three tasks, moving the data from the original source to the desired destination: (1) data extraction, (2) data transformation, and (3) data loading. Our rule processing algorithms will divide the second step, data transformation, in two complementary activities: validation and transformation.

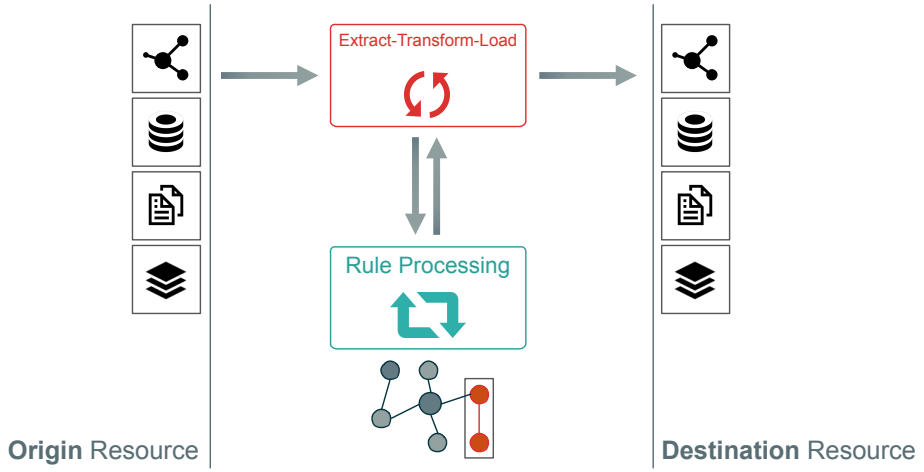

Fig. 1. Semantic rule-processing engine inclusion the integration pipeline. The new module interacts directly with the Extract-Transform-Load engine to validate and transform data before it finalizes the integration.

Applying these semantic rules is an interactive workflow, requiring the communication between the main application engine, the rule processing engine, and a knowledge base containing the rule configuration and translated content. This iterative process is described next and highlighted in Figure 2, using COEUS [7] as the system knowledge base.

1. Get content: load the content for processing from COEUS;
2. Return content: the application engine receives the content graph for semantic rule processing;
3. Get rules: the application engine requests the semantic processing rules from COEUS;
4. Return rules: COEUS returns the matching rules, if existing;
5. Process content: Apply matched validation and transformation rules to content, sending it to the final destination for integration;
6. Log: log all performed activities in the system.

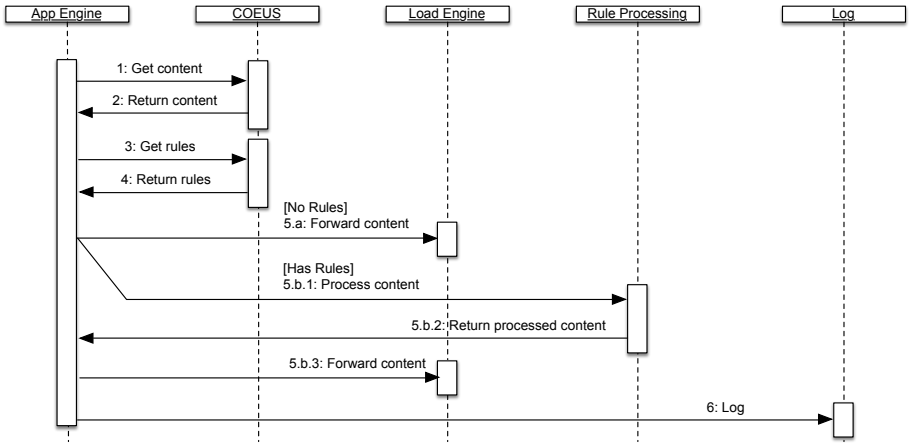

Fig. 2. Semantic rule processing sequence diagram, highlighting interactions amongst the various integration pipeline components.

3 Results

Leveraging on Semantic Web's capabilities, this modular engine can perform complex processing algorithms. These are divided in two categories: validation and transformation.

As the name implies, validation rules evaluate data against predefined conditions. Validation algorithms output a Boolean value: if true (content is valid) the integration proceeds, if false (content is not valid) the integration workflow stops.

There are two types of validation rules, based on conditional statements and regular expressions. Conditional validation rules assess if the content obeys to a predefined condition. These mimic simple arithmetic comparisons: less than (**<**), less or equal than (**<=**), more than (**>**), more or equal than (**>=**), equal (**=**), different (**!=**). While these are suitable for numerical values, string-based content requires more complex validation. For instance, emails, URLs, UniProt or reference sequence identifiers require specific matches for validation. Hence, the inclusion of regular expressions in the validation process is imperative.

Transformation rules are used to generate new content from existing data. Like in the validation rules, there are two types of possible transformations, basic operations and semantic. Basic operations cover number and string manipulation tasks, including mathematical equations for numbers and string concatenation or replacement for text.

Semantic transformations rely on inference and reasoning to generate new knowledge for integration, using complex ontology rules. These extremely powerful features are of growing importance, especially in the life sciences context, as they allow automating intelligent knowledge generation [8].

4 Discussion

Despite the ever-growing number of frameworks in this domain, there are several untapped challenges regarding the integration of information from distributed resources. Within these, improving the execution of semantic processing and validation rules arises as a key opportunity to greatly improve data integration platforms.

In this manuscript we propose a modular engine that adds a layer of semantics to traditional data integration workflows. With this strategy, the engine enables executing multiple pre- and post-integration processing algorithms, including data validation and transformation.

Acknowledgments. The research leading to these results has received funding from the European Community (FP7/2007-2013) under ref. no. 305444 – the RD-Connect project, and from the QREN "MaisCentro" program, ref. CENTRO-07-ST24-FEDER-00203 – the CloudThinking project

References

1. Bruckner, R.M., List, B., Schiefer, J.: Striving towards near real-time data integration for data warehouses. Springer (2002)
2. Mouttham, A., Peyton, L., Eze, B., Saddik, A.E.: Event-Driven Data Integration for Personal Health Monitoring. Journal of Emerging Technologies in Web Intelligence; Vol 1, No 2 (2009): Special Issue: E-health Interoperability (2009)
3. Naeem, M.A., Dobbie, G., Webber, G.: An Event-Based Near Real-Time Data Integration Architecture. In: Enterprise Distributed Object Computing Conference Workshops, 2008 12th, pp. 401-404. (Year)
4. Teymourian, K., Paschke, A.: Semantic Rule-Based Complex Event Processing. In: Governatori, G., Hall, J., Paschke, A. (eds.) Rule Interchange and Applications, vol. 5858, pp. 82-92. Springer Berlin Heidelberg (2009)
5. Anicic, D., Fodor, P., Rudolph, S., Stühmer, R., Stojanovic, N., Studer, R.: A Rule-Based Language for Complex Event Processing and Reasoning. In: Hitzler, P., Lukasiewicz, T. (eds.) Web Reasoning and Rule Systems, vol. 6333, pp. 42-57. Springer Berlin Heidelberg (2010)
6. Paschke, A., Kozlenkov, A.: Rule-Based Event Processing and Reaction Rules. In: Governatori, G., Hall, J., Paschke, A. (eds.) Rule Interchange and Applications, vol. 5858, pp. 53-66. Springer Berlin Heidelberg (2009)
7. Lopes, P., Oliveira, J.L.: COEUS:"semantic web in a box" for biomedical applications. Journal of Biomedical Semantics 3, 1-19 (2012)
8. Vandervalk, B.P., McCarthy, E.L., Wilkinson, M.D.: SHARE: A Web Service Based Framework for Distributed Querying and Reasoning on the Semantic Web. arXiv preprint arXiv:1305.4455 (2013)