

Extending SMW+ with a Linked Data Integration Framework

Christian Becker¹, Christian Bizer², Michael Erdmann³, and Mark Greaves⁴

¹ MediaEvent Services GmbH & Co. KG, Berlin, Germany — chris@beckr.org

² Web-based Systems Group, Freie Universität Berlin, Germany — chris@bizer.de

³ ontoprise GmbH, Karlsruhe, Germany — erdmann@ontoprise.de

⁴ Vulcan Inc., Seattle WA, US — markg@vulcan.com

Abstract. In this paper, we present a project which extends a SMW+ semantic wiki with a Linked Data Integration Framework that performs Web data access, vocabulary mapping, identity resolution, and quality evaluation of Linked Data. As a result, a large collection of neurogenomics-relevant data from the Web can be flexibly transformed into a unified ontology, allowing unified querying, navigation, and visualization; as well as support for wiki-style collaboration, crowdsourcing, and commentary on chosen data sets.

Keywords: Linked Data, Data Integration, Semantic MediaWiki

1 Introduction

The Allen Brain Atlas⁵ (ABA) comprises a growing collection of interactive image databases that integrate gene expression and neuroanatomic information for a variety of different organisms. Recently, the Allen Institute has started to explore new ways to accelerate scientific progress in its area, in particular by mappings between the ABA data and other standard data resources in neuroscience. As part of this project, the Semantic MediaWiki Linked Data Extension (SMW-LDE) is being built with two broad goals: to support unified querying, navigation, and visualization through a large collection of neurogenomics-relevant data sources; and to support wiki-style collaboration, crowdsourcing, and commentary on the chosen data sets.

The software base is the SMW+ semantic wiki.⁶ SMW+ is a set of open-source extensions to the popular Semantic MediaWiki [6] software which render it more appropriate for enterprise-scale use. We added a new Linked Data Integration Framework to SMW+, building on a number of open-source components for Web data access, vocabulary mapping, identity resolution, and quality evaluation of Linked Data. In SMW-LDE, data from multiple Linked Data sources can be flexibly transformed into a unified ontology that allows researchers to pose queries over data spanning multiple domains and data sets. The first phase

⁵ <http://www.brain-map.org/>

⁶ <http://wiki.ontoprise.de/>

of the project will bring ABA, Uniprot⁷, KEGG Pathway⁸, PharmGKB⁹ and Linking Open Drug Data [5] data sets together in order to solve the challenge of finding drugs that target elements within a disease pathway, but are not yet used to treat the disease. The genes associated with the found drugs may then be compared for commonalities through an integrated analysis of the related ABA structure expression data.

2 Integrating Linked Data from the Web

In this chapter, we discuss the architecture of the SMW-LDE deployment, which is depicted in figure 1.

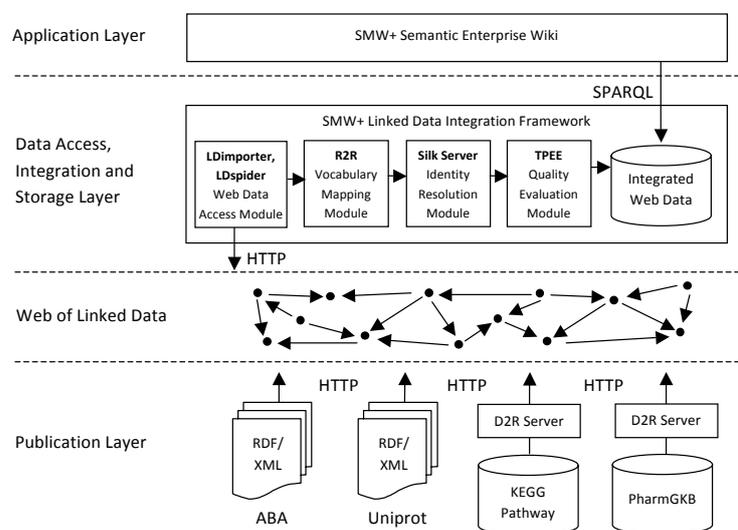


Fig. 1. Overall architecture of a Semantic MediaWiki using SMW-LDE to integrate Linked Data from the Web

All base data sets for the project, including Brain Atlas data, are published on the Web according to the Linked Data principles, thereby becoming part of a *giant global graph* – the Web of Linked Data. This logical graph is depicted in the *Web of Linked Data* layer in figure 1. Non-RDF data sources such as KEGG Pathway and PharmGKB are published using D2R Server¹⁰. The SMW+ Linked Data Integration Framework takes the role of a data access, integration and storage layer that makes Web data available to the application (in this case,

⁷ <http://www.uniprot.org/>

⁸ <http://www.genome.jp/kegg/pathway.html>

⁹ <http://www.pharmgkb.org/>

¹⁰ <http://www4.wiwiw.fu-berlin.de/bizer/d2r-server/>

the semantic wiki) in a unified manner. It consists of an integrated chain of Web data access, integration, and storage modules implemented in Java and Scala using the Jena framework. The modules and their tasks are described in the following.

1. **Web Data Access Module** The basic means to access Linked Data on the Web is to dereference HTTP URIs and to discover additional data sources by traversing RDF links, are realized using the LDspider¹¹ Linked Data crawler. In addition, our LDImporter module can also import data from various RDF dump formats or SPARQL endpoints using data set descriptions based on void descriptions [1] and Semantic Sitemaps [4].
2. **Vocabulary Mapping Module** Different Linked Data sources often use different vocabularies to represent the same type of information. For instance, both ABA and Uniprot ontologies contain the concept of a gene, but associate different relations and properties with it. In order to make Web data understandable to wiki users, the R2R Framework¹² is employed as the vocabulary mapping module in order to translate terms from different vocabularies into the wiki ontology.
3. **Identity Resolution Module** Different Linked Data sources use different URIs to identify the same entity, such as genes or proteins. The framework integrates Silk Server¹³ as its identity resolution module, which interlinks newly discovered data about entities with data about them that is already defined in the wiki or other connected data sources.
4. **Quality Evaluation Module** In order to prefer data from sources known for good quality and to resolve data conflicts [2], the framework includes a data quality evaluation module, the Trust Policy Evaluation Engine (TPEE).
5. **Integrated Web Data** The repository finally stores the Web data together with provenance information to be used by the application layer. We employ the Named Graphs data model [3] for representing Web data together with provenance information as an integrated model.

3 Using Linked Data in the Wiki

Making the integrated neurogenomics-relevant data available in a semantic wiki empowers users to access it in various ways:

- An ontology browser (cf. figure 2) of SMW+ enables the exploration of the integrated data, including representation of provenance information. In this way, correlations between data from different sources, as well as data conflicts can be identified in an interactive manner.
- Inline queries (cf. figure 2) and the interactive query interface offer means for ad hoc queries against the set of integrated Linked Data sources. The query results can be organized in wiki pages and complemented with textual descriptions or interpretations.

¹¹ <http://code.google.com/p/ldspider/>

¹² <http://www4.wiwiss.fu-berlin.de/bizer/r2r/>

¹³ <http://www4.wiwiss.fu-berlin.de/bizer/silk/server/>

