

# Extending SMW+ with a Linked Data Integration Framework

Christian Becker<sup>1</sup>, Christian Bizer<sup>2</sup>, Michael Erdmann<sup>3</sup>, and Mark Greaves<sup>4</sup>

<sup>1</sup> MediaEvent Services GmbH & Co. KG, Berlin, Germany — [chris@bechr.org](mailto:chris@bechr.org)

<sup>2</sup> Web-based Systems Group, Freie Universität Berlin, Germany — [chris@bizer.de](mailto:chris@bizer.de)

<sup>3</sup> ontoprise GmbH, Karlsruhe, Germany — [erdmann@ontoprise.de](mailto:erdmann@ontoprise.de)

<sup>4</sup> Vulcan Inc., Seattle WA, US — [markg@vulcan.com](mailto:markg@vulcan.com)

**Abstract.** In this paper, we present a project that 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 support as for wiki-style collaboration, crowdsourcing, and commentary on chosen data sets.

**Keywords:** Linked Data, Data Integration, Semantic MediaWiki

## 1 Introduction

The Allen Brain Atlas<sup>5</sup> (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 Environment (SMW-LDE) is being built with two broad goals: to support unified querying, navigation, and visualization through a large collection of neurogenomics-relevant data; and to support wiki-style collaboration, crowdsourcing, and commentary on the chosen data sets.

The software base is the SMW+ semantic wiki.<sup>6</sup> SMW+ is a set of open-source extensions to the popular Semantic MediaWiki [7] 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 the Web can be flexibly transformed into a unified ontology that allows researchers to pose queries over

---

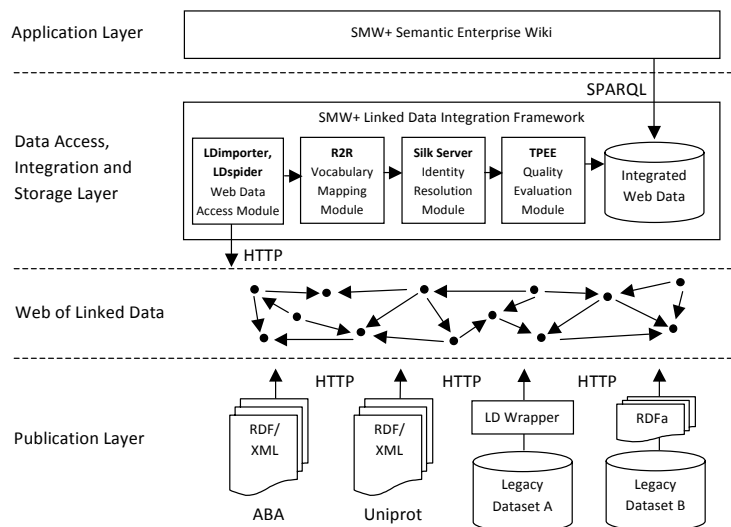
<sup>5</sup> <http://www.brain-map.org/>

<sup>6</sup> <http://wiki.ontoprise.de/>

data spanning multiple domains and data sets, such as how gene expression correlates with drug or protein pathways. The first phase of the SMW-LDE project will initially use Uniprot<sup>7</sup> and Linking Open Drug Data [6] data sets. Additional neuroscience data sets that do not yet have RDF serializations will also be published as Linked Data as part of the project.

## 2 Integrating Linked Data from the Web

In this chapter, we discuss the architecture, which is depicted in figure 1.



**Fig. 1.** The SMW+ Linked Data Integration Framework within the architecture of an SMW+ wiki

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. Legacy data sources are published using Linked Data Wrappers such as D2R Server<sup>8</sup>. 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, the semantic wiki) in a unified manner. It consists of an integrated chain of *Data Access, Data Processing and Storage* modules implemented in Java and Scala using the Jena framework<sup>9</sup>. The modules and their tasks are described in the following.

<sup>7</sup> <http://www.uniprot.org/>

<sup>8</sup> <http://www4.wiwiw.fu-berlin.de/bizer/d2r-server/>

<sup>9</sup> <http://jena.sourceforge.net/>

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 LDspider<sup>10</sup> 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 Framework<sup>11</sup> 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 Server [5] 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. The framework hereby employs 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 missing or incorrect mappings 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.

The most prominent benefit of using a semantic wiki environment in this project is the collaborative creation of semantic meta-data via annotations. Rather than just *querying* integrated Web data, users of the wiki can make

<sup>10</sup> <http://code.google.com/p/ldspider/>

<sup>11</sup> <http://www4.wiwiss.fu-berlin.de/bizer/r2r/>

The screenshot displays the SMW+ Ontology Browser interface. On the left, a 'Category Tree' shows a hierarchy of categories, with 'Gene' selected. The middle pane shows 'Instances' for the 'Structure' class, listing various IDs (P10, S95, S90, etc.) and their corresponding structure names. The right pane shows a 'Properties' table for 'Structure', listing properties like 'Structurename' and 'Values'. The top right pane shows a 'Wiki' article with a 'Sparql Query for Genes and their associated Structures' section. The bottom right pane shows the 'Sparql Query for Genes and their associated Structures' result table, which lists genes and their associated structures.

Gene	Genename	Structure	Structurename
uniprot:P10470	rdm_IF	Ace	<a href="http://purl.uniprot.org/tissues/01">http://purl.uniprot.org/tissues/01</a>   Pallium
uniprot:P10470	rdm_IF	Ace	<a href="http://purl.uniprot.org/tissues/01">http://purl.uniprot.org/tissues/01</a>   Brain cortex
uniprot:P10470	rdm_IF	Ace	<a href="http://purl.uniprot.org/tissues/01">http://purl.uniprot.org/tissues/01</a>   Cerebral cortex
uniprot:P10470	rdm_IF	Ace	<a href="http://purl.uniprot.org/tissues/01">http://purl.uniprot.org/tissues/01</a>   Cerebrum cortex
uniprot:P10470	rdm_IF	Ace	<a href="http://purl.uniprot.org/tissues/01">http://purl.uniprot.org/tissues/01</a>   Cerebellar cortex
uniprot:P10470	rdm_IF	Ace	<a href="http://purl.uniprot.org/tissues/01">http://purl.uniprot.org/tissues/01</a>   Cerebellum cortex
uniprot:P10470	rdm_IF	Ace	<a href="http://purl.uniprot.org/tissues/01">http://purl.uniprot.org/tissues/01</a>   Cortex
uniprot:P10470	rdm_IF	Ace	<a href="http://purl.uniprot.org/tissues/05">http://purl.uniprot.org/tissues/05</a>   Breast
uniprot:P10470	rdm_IF	Ace	<a href="http://purl.uniprot.org/tissues/05">http://purl.uniprot.org/tissues/05</a>   Mammary
uniprot:P10470	rdm_IF	Ace	<a href="http://purl.uniprot.org/tissues/05">http://purl.uniprot.org/tissues/05</a>   Mammary gland

**Fig. 2.** SMW+ Ontology Browser showing mapped Uniprot data (left), a related query within a wiki article (top right) and rendered query result in that article (bottom right)

use of the data by referring to imported genes or proteins in their own articles. By doing so, further statements about Web data resources are created that are in turn exported as Linked Data. This includes cross data source connections – for instance, the Allen Institute can publish their own data in the wiki and cross-reference it to Uniprot proteins.

## Acknowledgement

This work was supported in part by Vulcan Inc. as part of its Project Halo ([www.projecthalo.com](http://www.projecthalo.com)).

## References

- Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J. Describing linked datasets. In *Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009)*, 2009.
- Bleiholder, J., Naumann, F. Data fusion. *ACM Computing Surveys*, 41(1):1–41, 2008.
- Carroll, J., Bizer, C., Hayes, P., Stickler, P. Named graphs. *Journal of Web Semantics*, 3(4):247–267, 2005.
- Cyganiak, R., Delbru, R., Stenzhorn, H., Tummarello, G., Decker, S. Semantic sitemaps: Efficient and flexible access to datasets on the semantic web. In *Proceedings of the 5th European Semantic Web Conference (ESWC2008)*, 2008.
- Isele, R., Jentzsch, A., Bizer, C. Silk server - adding missing links while consuming linked data. Submitted to *1st Int. Workshop on Consuming Linked Data (COLD 2010)*.
- Jentzsch, A., Hassanzadeh, O., Bizer, C., Andersson, B., Stephens, S. Enabling tailored therapeutics with linked data. In *Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009)*, 2009.
- Krötzsch, M., Vrandečić, D., Völkel, M., Haller, H., Studer, R. *Semantic wikipedia*. *Journal of Web Semantics*, 5:251–261, September 2007.