Exploiting Linked Open Data for Enhancing MediaWiki-based Semantic Organizational Knowledge Bases

Matthias Frank¹ and Stefan Zander²

¹FZI Research Center for Information Technology, Haid-und-Neu-Str. 10-14, Karlsruhe, Germany
²Fachbereich Informatik, Hochschule Darmstadt, Darmstadt, Germany

Keywords: Linked Open Data, Semantic Web, Wiki Systems, Knowledge Engineering.

Abstract: One of the main driving forces for the integration of Semantic Media Wiki systems in corporate contexts is their query construction capabilities on top of organization-specific vocabularies together with the possibility to directly embed query results in wiki pages. However, exploiting knowledge from external sources like other organizational knowledge bases or Linked Open Data as well as sharing knowledge in a meaningful way is difficult due to the lack of a common and shared schema definition. In this paper, we introduce Linked Data Wiki (LD-Wiki), an approach that combines the power of Linked Open Vocabularies and Data with established organizational semantic wiki systems for knowledge management. It supports suggestions for annotations from Linked Open Data sources for organizational knowledge bases in order to enrich them with background information from Linked Open Data. The inclusion of potentially uncertain, incomplete, inconsistent or redundant Linked Open Data within an organization’s knowledge base poses the challenge of interpreting such data correctly within the respective context. In our approach, we evaluate data provenance information in order to handle data from heterogeneous internal and external sources adequately and provide data consumers with the latest and best evaluated information according to a ranking system.

1 INTRODUCTION

The adoption of semantic wiki approaches in organizational contexts and corporate environments has recently begun and is continuously growing (Ghidini et al., 2008; Kleiner and Abecker, 2010; Aveiro and Pinto, 2013). This is particularly the case for Semantic MediaWiki (SMW) (Krötzsch et al., 2006), an extension for the popular MediaWiki engine of Wikipedia, which introduces elements of the W3C’s semantic technology stack¹ (W3C, 2007) such as the Resource Description Framework’s triple model (Klyne and Carroll, 2004), semantic properties (so-called roles in Description Logic terms) as well as Concepts, i.e., dynamic categories that resemble the notion of domains in the RDF Schema language (Brickley and Guha, 2004). Those semantic features in conjunction with its collaborative knowledge engineering capabilities make semantic MediaWiki systems even more attractive for a deployment in professional environments (cf. listing “Wiki of the Month”²). SMW provides enhanced query construction capabilities with respect to organization-specific vocabularies and their specific contexts and allows to treat query results as first-class citizens and present them dynamically within wiki pages. Organizations like enterprises, NGOs or civil services can benefit from such features, which enable query construction, query expansion, and filtering using a lightweight set of ontological semantics (Vrandecic and Krötzsch, 2006; Zander et al., 2014).

However, although existing semantic wiki approaches like SMW, Ontowiki or Wikibase are built upon established semantic Web technologies, their utilization in wiki-based representation frameworks is primarily bound to a syntactic level. Moreover, those systems focus on building organization-specific lightweight ontologies and do not incorporate a common schema knowledge (cf. (Janowicz et al., 2014)) and its semantics per default. As a consequence, current semantic wiki systems are not able to exploit and benefit from the growing availability of Linked

¹A newer version of the semantic Web technology can be accessed at: https://smiy.wordpress.com/2011/01/10/the-common-layered-semantic-web-technology-stack/

²https://www.semantic-mediawiki.org/wiki/Wiki_of_the_Month
Open Data (LOD) (see (Bizer et al., 2009; Hausenblas, 2009; Heath and Bizer, 2011)). Moreover, the exploitation of additional knowledge from external sources hosted by other organizations or Linked Open Data sources as well as sharing knowledge in a meaningful way across organizational boundaries is difficult due to the lack of a common vocabulary among these approaches. Figure 1 illustrates the different levels of open data according to Berners-Lee (Berners-Lee, 2009):

The proposed approach overcomes the limitation of lacking schema knowledge in organizational knowledge bases by supporting the annotation of organization-specific schema knowledge with the common terminology of Linked Open Vocabularies (Janowicz et al., 2014) and extend the schema knowledge by interlinking modelled entities with entities represented as Linked Open Data in the LOD Cloud. Based on the resulting extended and interlinked schema knowledge, the so-called TBox in description logics (Baader et al., 2003), we provide the users with additional information for local entities, so that their correctness and validity can be evaluated on the basis of acquired externally hosted data where a common and shared agreement is prevalent. This leads to the following research questions:

- RQ1: How can users of organizational wikis be supported in establishing new links to Linked Open Data entities?
- RQ2: How can provenance information related to entities in an organizational wiki be represented, especially if these statements are inferred or gathered from Linked Open Data?
- RQ3: How can potentially uncertain, incomplete, inconsistent or redundant Linked Open Data be identified and tracked in order to increase the informative value of an organizational knowledge base?

We hypothesize that the information value of organizational knowledge bases will increase with the integration of LOD. For the evaluation of this hypothesis, we test our approach with the existing organizational wiki of our research group and compare the information derived from LOD with verified information wherever possible and calculate the rate of correctly derived information in relation to the false derivations. If this relation is better than the relation of our existing wiki, we regard our hypothesis as confirmed for this specific use case.

The remainder of this work is organized as follows: In Section 2, we discuss current semantic wiki approaches wrt. the implementation of semantic Web technology both on a syntactic and semantic level. In Section 3, we detail our approach of interlinking organizational knowledge bases with LOD. The implementation of the approach is described in Section 4. We conclude the current state of our work in Section 6 and discuss future work.

2 RELATED WORK

The review of related works is separated in two parts: In Section 2.1 existing semantic wiki software is reviewed where special emphasis is given to their openness towards a semantic technology stack. The findings are then summarized in Section 2.2 and close this section.

2.1 Semantic Wiki Software

Some software applications for creating semantic wikis do already exist. One of the best know application is SMW, see (Krötzsch et al., 2006). As many other wiki approaches, SMW is based on the MediaWiki engine, which is famous for providing the technical base for Wikipedia. The latest release of Semantic MediaWiki does support the development of an organization-specific knowledge base and the querying of this data within the wiki. It is also possible to export the semantically described facts to an external RDF store which does also allow to use SPARQL for querying the data. More extensions for

---

3 We use the terms ‘knowledge base’ and ‘semantic wiki’ interchangeably throughout this work.
4 The Linking Open Data cloud diagram: http://lod-cloud.net/
5 I.e., data or facts hosted internally in a local knowledge base.
6 https://github.com/SemanticMediaWiki/SemanticMediaWiki/releases
### Table 1: Characteristics of Semantic Wiki Applications.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Underlying Engine</th>
<th>Internal Data Storage</th>
<th>Data Export Format</th>
<th>Query construction</th>
<th>Integration of LOV/LOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic MW</td>
<td>MediaWiki</td>
<td>Relational (RDF mirror possible)(^6)</td>
<td>RDF (OWL only)</td>
<td>#ask: (SPARQL)</td>
<td>manual import of single terms</td>
</tr>
<tr>
<td>OntoWiki</td>
<td>-</td>
<td>Relational or RDF</td>
<td>RDF</td>
<td>SPARQL</td>
<td>publish ontology with LOV</td>
</tr>
<tr>
<td>WikiBase</td>
<td>MediaWiki</td>
<td>Relational</td>
<td>JSON or RDF</td>
<td>WB-Client (SPARQL)</td>
<td>WikiData-scheme</td>
</tr>
<tr>
<td>Cargo</td>
<td>MediaWiki</td>
<td>Relational</td>
<td>CSV</td>
<td>#cargo_query (SQL-like)</td>
<td>-</td>
</tr>
</tbody>
</table>

MediaWiki exist that provide better syntactical linkage of data modelled in SMW and RDF data like the Triple Store Connector (ontoprise GmbH, discontinued), the SparqlExtension\(^8\), or the RDFIO\(^9\) extension. All these approaches have in common that they provide semantic web technology on a syntactical layer, rather than a semantic web integration on a semantical layer. Only the LinkedWiki\(^10\) extension focuses on exploiting LOD for organizational knowledge bases which are built on top of the MediaWiki engine. Due to the fact that most semantic wiki applications are just used to structure and query data within an organizational wiki, rather than integrating data on a semantical level, Koren (Koren, 2015) presents the MediaWiki extension Cargo. The Cargo extension does also provide functionality for structuring on querying data, but without employing any semantic web technology. Rather than using semantic web technology, Cargo provides a wrapper around relational databases and exploits the well-established functionality of SQL. One example for a non-MediaWiki based semantic wiki applications is OntoWiki, see (Auer et al., 2007). OntoWiki has its focus on modelling a plain knowledge base, without providing a knowledge presentation for human readers like free text and natural language. A summarization of the previously discussed characteristics of semantic wiki applications is presented in Table 1.

Although the semantic wiki software applications introduced in Section 2.1 support semantic web technology like the Resource Description Framework (RDF) or even the SPARQL Protocol and RDF Query Language (SPARQL) on a syntactical level, the data integration across multiple data sources is still hard due to a common data scheme on a semantical level. Vrandecic and Krötzsch (Vrandecic and Krötzsch, 2014) describe the collaborative data scheme in Wiki-

---

\(^8\)https://www.mediawiki.org/wiki/Extension:SparqlExtension
\(^9\)https://www.mediawiki.org/wiki/Extension:RDFIO
\(^10\)https://www.mediawiki.org/wiki/Extension:LinkedWiki

Data as one possible solution for a common data scheme in order to extend schema knowledge in other wikis, especially Wikipedia. However, this approach does also define a data schema which is independent from Linked Open Vocabulary (LOV). In contrast to the WikiData approach, the OpenAnno approach by Frank and Zander (Frank and Zander, 2016) is focused on mapping proprietary ontologies to LOV in order to support the interlinkage of local knowledge bases with existing LOD.

### 2.2 Discussion

In Section 2.1 we have shown that current semantic wiki applications provide technical integration of semantic Web technology on a syntactic level. However, the introduced semantic wiki approaches do not support the annotation and interlinkage of organizational knowledge with LOD on a semantic level while considering the formal, model-theoretic semantics of the underlying ontology language, i.e., a vocabulary’s formal semantics. Such a recommendation system is provided by Open-Anno, but it is not integrated in any of the introduced semantic wiki applications. The statements maintained by one of these semantic wiki applications cannot be updated by external services as the statements contained within a wiki are always considered as master data. When importing statements from external sources into an organizational wiki, none of the introduced semantic wiki applications consider the context or the linkage of the data. Both is important in order to evaluate given statements, especially when they are inconsistent, redundant or ambiguous. To overcome these limitations, we introduce our Linked Wiki approach in Section 3.

### 3 APPROACH

In this Section, we introduce the Linked Data Wiki (LD-Wiki), our approach to combine the power of
Linked Open Vocabularies and Data with established organizational semantic wiki approaches.

3.1 Architecture

Our approach is technically based on MediaWiki in combination with an RDF-store. The main contribution is thus not on a technical layer, but aims at supporting the schema integration on a semantical layer. We provide a set of established LOV to encourage the reuse of these vocabularies in organizational wikis. The resulting organizational knowledge base using LOV is the foundation for suggestions of annotations from LOD. These annotations allow to enrich the organizational knowledge base with additional information from LOD. In order to distinguish organization-specific statements from statements gathered from LOD, we track the provenance information of each statement. The provenance information is stored using named graphs in the RDF-store, which extends the default triple model consisting of subject, predicate and object to quadruples, containing an ID for each statement. This ID allows us to attach provenance information to each statement. Using the provenance information, we can also handle uncertain or inconsistent data and provide the data consumers with the latest and most suitable information. One characteristic of our approach is the strict separation of statements in the triple store maintained by our extension and the non-semantic part of the wiki, like free text, MediaWiki syntax and place-holders for the semantic statements, which are still maintained by the MediaWiki engine. By separating the semantic statements from the non-semantic part of the wiki, we avoid the issue of syncing statements between the wiki and the knowledge base. Additionally, we are able to maintain and curate the semantic statements outside of the wiki without causing inconsistent data. This separation of semantic and non-semantic data is therefore a prerequisite for the transparent integration of statements from the wiki itself and external statements from LOD.

Figure 2 shows the architecture of our approach including the two layers for knowledge management and human friendly presentation. Our Linked Data Management Module (LDaMM) is the stand-alone business logic module for the knowledge management layer which queries LOD on demand, updates the local RDF knowledge graph and serves the MediaWiki engine for a human friendly presentation of the knowledge graph. Other useful features of LDaMM, which are not implemented yet, are reasoning and rule execution which help to curate the local knowledge graph. For the business logic module of the human friendly presentation layer we employ the MediaWiki engine which relies on the knowledge provided by LDaMM, rather than relying on an own knowledge serialization as it is done by e.g. SMW. Avoiding redundant management of knowledge, we ensure that the organizational knowledge management is always in a consistent state. However, the MediaWiki engine has to provide additional information for a human friendly presentation like free text and markup information which is stored in a separate relational database provided by a MySQL instance for local data management on the presentation layer.

3.2 Challenges

In contrast to providing only wiki-based statements within an organizational wiki, our approach does also include external statements from multiple LOD sources. The inclusion of external statements causes issues when the same entity is described in multiple sources. One of these issues is the fact of potential redundant or inconsistent data. We address this issue by exploiting the gathered provenance information and evaluate the statements bases on a ranking derived from contained provenance statements. The ranking is influenced by the inter-linkage of the source as an indicator of reference and by the evaluation of statements by users of the wiki. Another issue is the potential amount of provenance information. Although this provenance information is necessary in order to evaluate the trustworthiness of statements, it would be confusing for the user to show all available provenance information for each statement. We address this challenge by evaluation the provenance information in the background and just showing the resulting statement to the user with an option to expand the underlying provenance-based derivation of the statement.

4 IMPLEMENTATION

For the implementation of the LD-Wiki approach, we build on the open source framework of MediaWiki. In contrast to other MediaWiki-based approaches, we implement the knowledge management as a stand-alone module that controls storing, querying, updating, reasoning and rule execution of RDF-statements, rather than integrating the knowledge management in MediaWiki itself. This allows for a lightweight MediaWiki extension that triggers the knowledge management module for rendering wiki pages (Section 4.1) and also if a user creates new wiki pages for Terminological Box (TBox) (Section 4.2) or Assertional Box (ABox) (Section 4.3) of our knowledge
Figure 2: Architecture of the Linked Open Data Wiki Approach: Knowledge layer for enriching the local RDF-graph with Web knowledge and curating it, presentation layer for a user friendly presentation of RDF-data enriched with free text and formatted by MediaWiki-Syntax stored in local MySQL-DB.

4.1 Rendering Wiki Pages

The Wiki pages in our approach consist of free text for a human readable presentation, placeholder for data from the knowledge management module and MediaWiki syntax to format the style of the page. When a page is requested, the according parser function\textsuperscript{11} of the LD-Wiki Extension requests the necessary data from the knowledge management module and replaces each placeholder with the according value from the knowledge base.

4.2 New TBox Pages

The key factor to let the LD-Wiki approach work well and build a TBox which can be interpreted in the context of LOD, it is necessary to interlink new concepts with concepts from LOV. Concepts are represented as categories in MediaWiki. Therefore, whenever a new category is created, LDaMM is triggered to query for existing concepts in LOV with the same label as the label for the new category. If one or more classes are found, the user of the wiki can select the concepts that represent the intended meaning.

Figure 3 shows how this looks like in the LD-Wiki. For creating a new concept within the local knowledge management, the user opens the special page for creating a new category in MediaWiki and provides the string that labels that new concept. When submitting this string, MediaWiki sends it to LDaMM in the knowledge management layer. LDaMM invokes SPARQL queries to search for concepts in LOD that are labeled with the same string. If, for example, the user would like to create a new concept for cities for a German-language terminology, he would probably enter the string "Stadt" for this concept. To find concepts related to that string in LOD, LDaMM produces the query string 'SELECT * WHERE ?category rdf:type rdf:Class; rdfs:label "Stadt"@de.' to discover any concept that has the label "Stadt" with a German language tag. This query string is then executed at available public SPARQL endpoints to discover adequate concepts. Expected

\textsuperscript{11}https://www.mediawiki.org/wiki/Parser\_functions
4.3 New ABox Pages

Assuming that the categories of the LD-Wiki are linked to the according concepts in LOV as discussed in Section 4.2, we can assist the user on creating new instances for the ABox in the wiki. Instances of a concept are represented as pages within the category that represents that concept in the wiki. Therefore, whenever a page is created, the knowledge management module is triggered to query for existing individuals in LOD with the same label as the new page and the same concept as the category of the new page. If one or more individuals are found, the user of the wiki can select the individuals that represent the same instance as the new page. The great benefit for this kind of interlinkage is that we can query directly for properties of these individuals in LOD or retrieve a summary of entity data using entity summarization tools like LinkSUM (Thalhammer et al., 2016).

Figure 4 shows how this is done in the LD-Wiki. For creating a new instance within the local knowledge management, the user opens the special page for creating new instances in MediaWiki, provides the string that labels that new instance and selects the category of which the new page should be an instance of. When submitting this string, MediaWiki again sends it to LDaMM in the knowledge management layer together with the identifier of the selected category. LDaMM invokes SPARQL queries to search for instances in LOD that are labeled with the same string and are instances of any of the concepts that the given category is linked to. If, for example, the user would like to create a new instance of the category “Stadt” for the German-language terminology in our example, he would enter the name of this city as string, e.g., “Karlsruhe”, and select the category “Stadt” for it. To find instances related to that string and category in LOD, LDaMM produces the query string 'SELECT * WHERE { ?instance rdf:type http://www.wikidata.org/entity/Q515; rdfs:label "Karlsruhe". } limit 100' to discover any instance that has the label “Karlsruhe” and type http://www.wikidata.org/entity/Q515, as this is one of the concepts which is linked to the category “Stadt”. This query string is then executed at available public SPARQL endpoints to discover adequate instances. An expected result would be for example http://www.wikidata.org/entity/Q1040 which describes the German city in the state of Baden-Wuerttemberg. LDaMM returns these results to MediaWiki where the user can select the adequate instance. On creation of the new instance in MediaWiki, the information of the new instance and the linked instances is send back to LDaMM and stored to the local knowledge graph, including all properties that are retrieved from the linked entity and also their provenance information. All this information is now available in the local knowledge graph without further action.
5 EXPERIMENTAL RESULTS

For a first evaluation of our approach, we use the SPARQL endpoints of DBpedia\(^\text{12}\) and Wikidata\(^\text{13}\) as two instances of LOD resources. Due to the different implementation of these endpoints, the query string has to be mapped to meet the individual characteristics.

5.1 New TBox Pages

The first step is to run the query \(\text{SELECT * WHERE \{ ?category rdf:type rdf:Class; rdfs:label "Stadt"@de . \}}\) on the SPARQL endpoints of DBpedia and Wikidata.

**Wikidata**

Wikidata uses the property http://www.wikidata.org/prop/direct/P279 (subclass of) to describe subclasses of other classes. We therefore map the property-value pair “rdf:type rdf:Class” to this wikidata property which results in the following query:

\[
\text{SELECT * WHERE \{ ?category <http://www.wikidata.org/prop/direct/P279> \?class ; rdfs:label "Stadt"@de . \}}\]

When executing this query at the SPARQL endpoint of Wikidata, we receive two classes: http://www.wikidata.org/entity/Q515 which describes a city as a large and permanent human settlement and http://www.wikidata.org/entity/Q15253706 which is the class for a more specific definition of a city by country that holds the size of cities and towns in Korea, Japan, the USA, China, North Korea and France.

**DBpedia**

For DBpedia, we map the class of rdfs:Class to owl:Class as DBpedia makes use of Web Ontology Language (OWL) and the default configuration of this endpoint does not imply superclasses which would include rdfs:Class as well. The result is the following query:

\[
\text{SELECT * WHERE \{ ?category rdf:type owl:Class ; rdfs:label "Stadt"@de . \}}\]

When executing this query at the SPARQL endpoint of DBpedia, we receive again two classes: http://dbpedia.org/ontology/City and http://dbpedia.org/ontology/Town.

5.2 New ABox Pages

Next, we test the retrieval of instance data for a given concept. In our example, we want to execute the query \(\text{SELECT * WHERE \{ ?instance rdf:type <http://www.wikidata.org/entity/Q515>; rdfs:label "Karlsruhe"@de. \}}\) on the SPARQL endpoints of DBpedia and Wikidata.

**Wikidata**

Wikidata uses the property http://www.wikidata.org/prop/direct/P279 (subclass of) to describe subclasses of other classes. We therefore map the property-value pair “rdf:type rdf:Class” to this wikidata property which results in the following query:

\[
\text{SELECT * WHERE \{ ?category <http://www.wikidata.org/prop/direct/P279> \?class ; rdfs:label "Stadt"@de . \}}\]

When executing this query at the SPARQL endpoint of Wikidata, we receive two classes: http://www.wikidata.org/entity/Q515 which describes a city as a large and permanent human settlement and http://www.wikidata.org/entity/Q15253706 which is the class for a more specific definition of a city by country that holds the size of cities and towns in Korea, Japan, the USA, China, North Korea and France.

**DBpedia**

For DBpedia, we map the class of rdfs:Class to owl:Class as DBpedia makes use of Web Ontology Language (OWL) and the default configuration of this endpoint does not imply superclasses which would include rdfs:Class as well. The result is the following query:

\[
\text{SELECT * WHERE \{ ?category rdf:type owl:Class ; rdfs:label "Stadt"@de . \}}\]

When executing this query at the SPARQL endpoint of DBpedia, we receive again two classes: http://dbpedia.org/ontology/City and http://dbpedia.org/ontology/Town.
prop/direct/P31 (instance of) to indicate that an instance belongs to a specific category. We therefore map the property rdf:type to the Wikidata-specific term:

\[
\text{SELECT} * \text{WHERE } \{ ?\text{instance} \text{ rdf:type } \text{http://www.wikidata.org/ontology/City} ; \text{rdfs:label "Karlsruhe"@de .} \}
\]

For this query, we get two matching instances: http://www.wikidata.org/entity/Q1040, the German city in the state of Baden-Württemberg, and http://www.wikidata.org/entity/Q1026577, a city in North Dakota. Depending on the instance the user wants to refer to, he or she has to select the appropriate one. This example does also show that a completely automatic information retrieval is difficult to control and therefore human supervision of this process is still reasonable. If we run the query with the more definition of a city by country using the query string \[\text{SELECT} * \text{WHERE } \{ ?\text{instance} \text{ rdf:type } \text{http://www.wikidata.org/ontology/Town} ; \text{rdfs:label "Karlsruhe"@de .} \} \], we do not get any result.

**DBpedia**

For DBpedia, we run the query for instances of http://dbpedia.org/ontology/City or http://dbpedia.org/ontology/Town:

\[
\text{SELECT} * \text{WHERE } \{ ?\text{instance} \text{ rdf:type } \text{http://dbpedia.org/ontology/City} ; \text{rdfs:label "Karlsruhe"@de .} \}
\]

The single result of this query is the instance of http://dbpedia.org/resource/Karlsruhe.

6 CONCLUSION

With the LD-Wiki approach we have shown how we can assist users of organizational wikis with creating new links to LOD entities. As we have separated the knowledge management module from the knowledge representation in MediaWiki, we are able to keep track of the provenance of statements in our knowledge base without affecting the knowledge representation. Especially the synchronization of wiki data and the organizational knowledge base as it was the case in other semantic wiki approaches like SMW is not required any longer. The great benefit for this kind of interlinkage is that we can enrich the information value of individuals in LD-Wiki by querying for properties of these individuals in LOD or retrieve a summary of entity data using entity summarization tools to exploit the power the continuously growing amount of LOD for corporate knowledge bases. Open issues for future work include the privacy for confidential data on the one hand while publishing parts of the corporate knowledge base as LOD on the other hand.

This requires a proper implementation of Access Control Lists (ACLs) with carefully designed access roles for each statement in the knowledge base. In order to benefit from features of semantic Web technologies besides reusing information from LOD, the knowledge management module should also enable advanced reasoning over organizational data which would also help to evaluate and interpret potential uncertain, incomplete, inconsistent or redundant LOD correctly. This would further increase the informative value of the organizational knowledge base.

**REFERENCES**


