STI INNSBRUCK

COMBINING AND EASING THE ACCESS OF THE ESWC SEMANTIC WEB DATA

Dieter Fensel, and Alex Oberhauser

STI Innsbruck, University of Innsbruck,
Technikerstraße 21a, 6020 Innsbruck, Austria
firstname.lastname@sti2.at

SEMANTIC TECHNOLOGY INSTITUTE INNSBRUCK

STI INNSBRUCK
Technikerstraße 21a
A - 6020 Innsbruck
Austria
http://www.sti-innsbruck.a

Combining and easing the access of the ESWC Semantic Web Data

Alex Oberhauser, Dieter Fensel

Semantic Technology Institute (STI)
6020 Innsbruck, Austria
{firstname}.{lastname}@sti2.at

Abstract. The following paper describes the combination of *conference metadata* stored on the conference website with *Semantic Web Dog Food metadata* retrieved from a paper submission system. The main goal is to increase the visibility of publicly available research papers and related conference information.
The outcome of this paper is an alignment algorithm and a related implementation that supports the alignment process of the two metadata sets.
Keywords: ESWC, easychair.org, conference, linked open data, metadata

1 Introduction

In order to increase the visibility of research papers and conference information on the web this paper presents an approach that aligns *conference metadata*, retrieved from the official conference website, with *Semantic Web Dog Food (SWDF) metadata*, retrieved from the paper submission system easychair.org. Where applicable the *SWDF metadata* vocabularies are changed to vocabularies that are understandable by all major search engines[1]. This step increases further the discoverable via search engines.

The alignment process is supported by a mapping algorithm and a related implementation that should ease the combination of the two metadata sets. Additional it describes how to setup the content management system (CMS) of the conference website in order to define the mappings from the CMS to RDF/RDFa/Microdata.

This paper describes the technical approach with an implementation in section 3 and the mapping of the vocabularies in section 4. The last section, section 5, concludes the paper.

2 Motivation

The main motivation behind this approach is to increase the visibility of research publications by combining the two related, but physically separated metadata sources and publishing them in a easy accessible form. The first data set is the metadata from the paper submission system and the second is the metadata provided by the conference website. As shown in Figure 1, the two data sets are mainly overlapping, but described with different vocabularies. This makes

[1] schema.org was developed by Bing, Google, Yahoo! and Yandex

the automatic alignment, with current technology, impossible and decreases the visibility and automatic discoverable. It increases also the effort of an human reader to check if concepts, in the two sets, are equal.

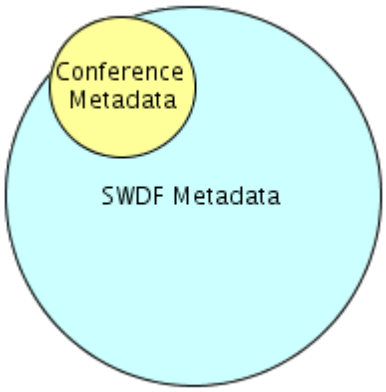

Figure 1. Overlapping of metadata sets

Additional the import of paper submissions into the conference website was in the past a time consuming, manual effort that results in most cases to human-only readable data. The presented approach overcomes these weaknesses with the help of an mapping algorithm that takes the available information from the Semantic Web Dog Food website and imports them to the content management system of the related conference website. This approach is possible, because the Semantic Web Dog Food dataset gets the information from the widely used conference system easychair.org. Easychair is used by the most conferences for the submit process.

The import of semantic-enriched information has additional the advantage that it could be used for the search engine optimization process.

3 Conceptual Metadata Integration

In order to be able to import the data into the CMS and later one back to the Semantic Web Dog Food (SWDF) a mapping has to be defined. Figure 2 visualizes the different metadata set. For this purpose the input source (SWDF data dump) is mapped with the help of the following rules:

1. schema.org vocabulary for the mapping from SWDF vocabulary.
2. If no concept in schema.org, fallback to sti2/oc vocabulary.
3. If no concept in sti2/oc vocabulary, fallback to original ontology.

Table 1 and 2 outline a few, representative examples how such a mapping looks like.

Semantic Web Dog Food	**Drupal Content Type**	**RDF Concept**
foaf:Organization	Organization	schema:Organization
foaf:Person	Person	schema:Person
swc:ConferenceEvent	ConferenceEvent	sti2:ESWC
swc:KeynoteTalk	KeynoteTalk	swc:KeynoteTalk

swc:Chair	Chair	swc:Chair

Table 1. Class mapping examples

Semantic Web Dog Food	**Drupal Content Type**	**RDF Concept**
rdfs:label	title	schema:name
ical:description	body[Body]	schema:description
swrc:affiliation	affiliation	schema:affiliation
ical:summary	body[Summary]	ical:summary
swc:booktitle	booktitle	swc:booktitle

Table 2. Property mapping examples

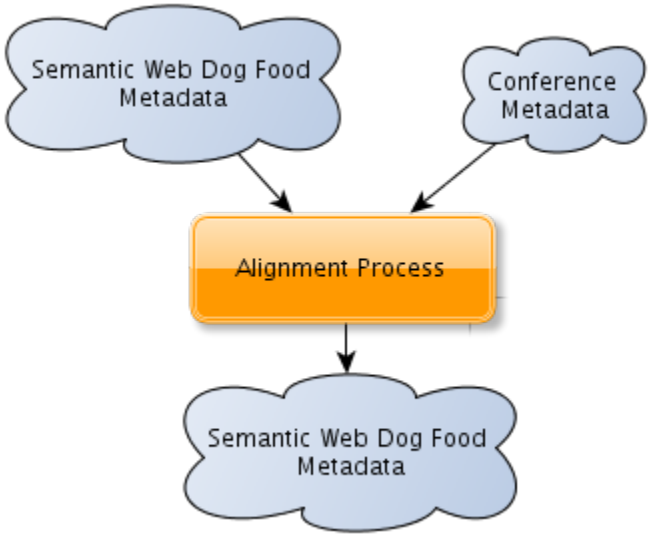

Figure 2. Metadata Sets

The reason why schema.org was chosen for the primary mapping is the use of this vocabulary by search engines. After the import into the CMS the metadata could be used for search engine optimization (SEO). The good tool support for microdata makes the metadata easy publicly accessible. As intermediate step the data has to be imported into the CMS. For this purpose the SWDF data has to be mapped to the internal representation. This mapping must match with the internal class structure of the implementation. The CMS internals are not visible from the outside and have no effect of the end result. Important are only the annotation of the properties and classes. That is the reason why the ContentTypes in the CMS have the same name as the concepts in the source vocabulary, but without the namespace.

4 Technical Integration

The ESWC to CMS process consists of different steps that are partially supported by an implementation. The first step is to get the conference data from the Semantic Web Dog Food platform. For the current implementation the following URL is used: http://data.semanticweb.org/conference/eswc/2012/complete. During the development the RDF/XML dump was imported to OWLIM and exported as N3. This newly created dump can be used by the implementation that takes this data and imports it to the CMS Drupal 7.x. For this purpose the CMS must be setup correctly. That means that the ContentTypes of the CMS have to match with the ContentTypes used by the implementation. Another step that must be performed in the CMS is the annotation of the properties with the right vocabularies. This step is supported by a script that is able to extract all classes from the dump.

In order to use the conference integration implementation the content management system Drupal 7.x has to be enriched by the following modules:

RESTful Web Service [2]: This module provides a REST API that is used for the import process.
RDFx [1]: Enriches the CMS with RDF capabilities for the mapping from internal properties to RDF vocabularies. In combination with the RESTful Web Services module it allows the serialization in different formats, e.g. RDF/XML, NTriples, Turtle.
SPARQL [3]: The SPARQL module allows the data extraction with the help of SPARQL 1.0 queries. This extension is helpful for the export of the metadata.

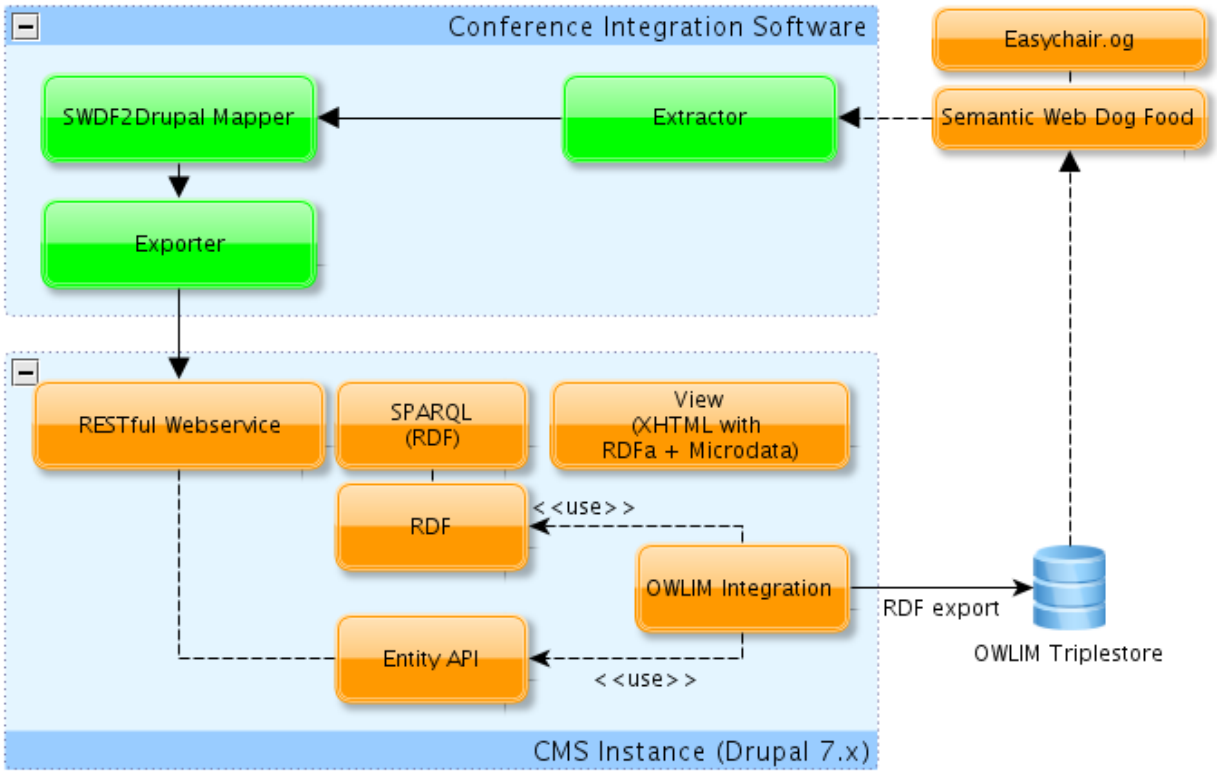

Figure 3. Implementation Architecture Diagram

The related implementation extracts from the gained RDF dump the needed concepts and export them via REST API into the CMS. For this purpose an in-memory triplestore and SPARQL is used. Each class instance with related properties is mapped (with the help of SPARQL SELECT queries) to the code internal classes. Each class represents the mapping

from one class from the ontology to one ContentType in the CMS. The instantiated objects are then converted to an XML message that is understandable by the content management systems.

Such an XML messages represents the internal ContentType with related fields. The architecture diagram in Figure 3 describes graphical the import process.

5 Conclusion

The approach presented in this paper outlines a strategy how to improve the visibility of conference data. For this purpose a basic alignment process was defined and supported by an implementation that should ease the alignment. The alignment algorithm describes a strategy how to improve the visibility and could be adapted to other use cases.

As data sources the data from Semantic Web Dog Food and the internal data from the conference website was used. The algorithm was demonstrated on the base of easychair.org data and the content management system Drupal 7.x. The implementation was written as Ruby library. This allows the integration of the implementation into a bigger platform.

A Mapping Definitions

Prefix	**Namespace**
swdf	http://data.semanticweb.org/ns/swc/ontology#
foaf	http://xmlns.com/foaf/0.1/
swrc	http://swrc.ontoware.org/ontology#
oc	- (Online Communication Ontology, no URL yet)

Semantic Web Dog Food	**Drupal Content Type**	**RDF Concept**
swdf:ConferenceEvent	ConferenceEvent	oc:ConferenceSeries (or for ESWC: sti2:EWSC)
foaf:Organization	Organization	schema:Organization
foaf:Person	Person	schema:Person
swdf:KeynoteTalk	KeynoteTalk	swdf:KeynoteTalk
swrc:InProceedings	InProceedings	swrc:InProceedings
swdf:Chair	Chair	swdf:Chair
swdf:Presenter	Presenter	swdf:Presenter
swrc:Proceedings	Proceedings	swrc:Proceedings

Table 3. Class Mapping

References

1. Drupal RDFx module. http://drupal.org/project/rdfx, accessed: 14.05.2012
2. Drupal RESTful Web Services module. http://drupal.org/project/restws, accessed: 14.05.2012
3. Drupal SPARQL module. http://drupal.org/project/sparql, accessed: 14.05.2012
4. Semantic web dog food. http://data.semanticweb.org/, accessed: 14.05.2012