Integrating Ontology-based Metadata Enrichment into a CMS-based Research Infrastructure

Dennis Spohr  
CITEC, Germany  
dspohr@cit-ec.uni-bielefeld.de

Philipp Cimiano  
CITEC, Germany  
cimiano@cit-ec.uni-bielefeld.de

Cord Wiljes  
CITEC, Germany  
cwiljes@cit-ec.uni-bielefeld.de

Keywords: linked data, RDF, ontologies, metadata harvesting, content management systems

1. General Description

This abstract discusses research under development aiming to create an ecosystem of entities connected to a research institution, such as its researchers and the resources produced. In particular, we are investigating ways of being able to enter metadata descriptions in a uniform way on the one hand, and to expose them in various different formats on the other. Here, we aim at supporting current standards for metadata exchange, such as the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH; Van de Sompel et al., 2004), as well as the Resource Description Framework (RDF1) in order to be able to interlink the descriptions with others available on the Linking Open Data (LOD) cloud. For the whole process to integrate smoothly into our existing research infrastructure, the approach presented here relies on the Open Source Content Management System Drupal2, as it is at the center of our current infrastructure for managing metadata. While version 7 of Drupal generally provides better support for Semantic Web formalisms than its predecessor, we should mention that we base our current architecture on Drupal 6 – on the one hand due to the fact that many Drupal 7 modules are still in beta status, and on the other hand because the existing research infrastructure builds on Drupal 6.

As Corlosquet et al. (2009) have shown, Drupal offers a number of solutions for creating LOD-compliant resource descriptions, such as modules for generating RDF descriptions alongside (X)HTML webpages, and for implementing a PARQL endpoint – an RDF repository that can be directly queried using the standard Semantic Web query language SPARQL3. This means that metadata descriptions which have been imported into the Drupal CMS can be exposed not only in the form of human-readable (X)HTML webpages, but also in the form of RDF. However, it needs to be said that while there is a very strong development towards using RDF – and higher-level Semantic Web formalisms in general – for metadata annotation, many infrastructures do not support these technologies at the moment. For example, the Common Language Resources and Technology Infrastructure project (CLARIN; Varádi et al., 2008), which aims at developing an infrastructure that is capable of harvesting metadata represented in different metadata vocabularies – such as DCMI or IMDI (Broeder and Wittenburg, 2006) –, bases its harvesting technologies on OAI-PMH. Here, we can make use of the OAI-PMH Views plugin module of Drupal4, which implements a data provider supporting Dublin Core metadata that can be indexed by CLARIN and harvested using OAI-PMH.

The above descriptions interoperate with ongoing efforts to formalize all resources connected to our research institution in terms of an OWL ontology, which is being developed on the basis of existing ontologies, such as the AKT Reference Ontology5 and the OWL-Time Ontology6. To give an example of the benefits of such formalizations, consider that research publications are linked

---

1 http://www.w3.org/RDF/  
2 http://drupal.org  
3 http://semantic-drupal.com/node/4  
4 http://drupal.org/project/views_oai_pmh  
5 http://www.aktors.org/ontology/portal  
6 http://www.w3.org/2006/time
to their authors, who are in turn affiliated with specific departments and associated with projects, which in turn have specific data sets and publications – such as the initial publication – as research outputs. It is clear to see that it is not feasible to annotate all this information directly when entering the metadata of the publication. Rather, the link to the URI identifying the author establishes a connection to all the other information. Moreover, in the context of the Semantic Web, we can make use of reasoning in order to infer links which have not been asserted explicitly. For example, in cases where a publication has not been annotated with the project in whose context it was created, we can use the Semantic Web formalisms in combination with the aforementioned ontologies and the Dublin Core vocabulary to find potential research outputs of these projects. Here, we can define that if the date of a dct:BibliographicResource is within the time interval of an akt:Project which has an author of the bibliographic resource as a project member, then the bibliographic resource may potentially be a research output of the project. As temporal overlap is not sufficient, however, we do not assert such inferred links explicitly, but instead use them to speed up the process of detecting actual research outputs of projects.

FIG. 1 summarizes the proposed architecture of the system. At the time of writing, both the RDF export and the OAI-PMH data provision have been implemented in a prototype system that is currently being tested, while the ontological formalization is still at an early development stage. In the poster, we present the key concepts of this architecture, illustrating how different modules interoperate to provide metadata descriptions in different formats, such as RDF and via OAI-PMH. Moreover, we will discuss the current state of the ontological formalization of our research institution, illustrate its major benefits, and show how it interoperates with the above architecture.

References