Ontology Lexicalization as a core task in a language-enhanced Semantic Web

Philipp Cimiano
University of Bielefeld, Germany

In order to provide language-based access to the growing amount of knowledge published in Semantic Web formalisms, e.g. as part of the so called Linked (Open) Data cloud, ontologies and vocabularies used to describe data need to be enriched with information about how classes and properties modeled therein can be expressed linguistically, also in different languages.

In this talk I will discuss the vision of a language-enhanced Semantic Web in which information about linguistic realization is modeled in Semantic Web languages and forms part of the Linked Data itself, thus becoming retrievable by standard Semantic Web search engines and indexing services as well as queryable and browseable using Semantic Web standards. This ecosystem of ontologies enriched with linguistic knowledge can then be exploited by a number of NLP applications across applications, avoiding duplication of work by people aiming at supporting language-enhanced access to the Semantic Web.

There are three important ingredients to make this vision feasible.

First of all, we need vocabularies that allow to model lexical and linguistic knowledge using Semantic Web vocabularies. In the last years, we have been developing the lemon model for this purpose that has formed the initial input for standardization activities carried on in the context of the W3C Community Group on the ontology-lexicon interface.

Second, we need practical approaches that ease the effort of creating such ontology lexica. I will present current efforts in this direction aiming at semi-automatically supporting the creation of ontology lexica by human users by exploiting a domain corpus. I will present results of experiments in which we use Wikipedia to automatically induce a lexicon for the DBpedia ontology.

Third and finally, we need people to recognize the value of ontology lexicalization so that they have incentives to contribute to the development of lexica for their favourite ontologies, and we need efficient and tested (collaborative) methodologies which incorporate semi-automatic support allowing people to develop such lexica effectively and efficiently. I have unfortunately no solutions so far for this third challenge to present.