# Challenges for the Multilingual Semantic Web

**Walther v. HAHN**
Universität Hamburg • Arbeitsbereich
Natürlichsprachliche Systeme
Vogt-Kölln Straße 30
Hamburg, Germany, 22527
vhahn@informatik.uni-hamburg.de

**Cristina VERTAN**
Universität Hamburg • Arbeitsbereich
Natürlichsprachliche Systeme
Vogt-Kölln Straße 30
Hamburg, Germany, 22527
vertan@informatik.uni-hamburg.de

**Abstract**

In this paper we give an overview of Semantic Web technologies and the impact of these ones for multilingual Web. We present a possible solution for improving the quality of on-line translation systems, using mechanisms and standards from Semantic Web. We focus on Example based machine translation and the automatization of the translation examples extraction by means of RDF-repositories.

## 1 Basic Principles of Semantic Web

In WWW are used each year more than 5 Billion Documents from more than 800 million active users. However, up to now is WWW self-organised System without any predefined or standardised structure or administration.

The huge the quantity of information in WWW becomes the more difficult its administration is. Especially the update and targeted retrieval of information are more and more difficult. Usually the information is retrieved following a key-word search. This lexical search creates several problems:

- too strong specialisation (better information about „violin“ can be found maybe under „string instruments“ or 2instruments“), because of
- too many different specialised meanings, here for e.g. „instruments“ are only „music instruments“ but no surgery instruments.
- No possibility of searching also synonyms of the keywords (in German e.g. Geige/Violine)
- No possibility of multilingual search (e.g. German documents which contain the word „Geige “ will not be found)
- Limitation of the search mechanism, excluding semantics of multiword expressions
-

The Semantic web (Berners-Lee 1998) aims to support a better access to the information in WWW through references from the site to a standard common semantic meta-data represented by:

- Ontologies as relations between domain objects: e.g. „mammal“ is a sub-concept of „animal“

- Inferences among ontologies objects: e.g. “If A $\subset$ B and B $\subset$ D, than A $\subset$ D (Transitivity): “A bear is an animal, because all bears are mammals and all mammals are animals”

The idea of Semantic Web is the following: when systematic conceptual (partially also terminological) description of a fact exists (e.g. transactions by the bourse) or an entire domain (Bourse) together with the relations between concepts is available then:

- each information provider can relate the information with this ontology, and describe accordingly the content and

the user-query will be mapped on this ontology, and not only on the text.

For this approach an ontology developed by domain-specialists, or (semi-) automatic extracted, encoded in a standard language (OWL) is required. The information provider must then link and annotate his text with this ontology. The text is described than semantically in RDF. A search machine on the server can compare the RDF-annotated text as well as the query with the OWL-Ontology and retrieve appropriate information, also in cases when lexical search would have been unsuccessful.

## 2 Languages in the Semantic Web.

Tim Berners-Lee (Berners-Lee 1998) has developed the Semantic Web and described it as a hierarchy of formalisms, which are all based of Unicode Texts and Web-addresses (URIs). On top of these are classical web languages from the XML-family. The next two layers are the descriptions of texts with RDF and RDF-schema. These are in connection with ontologies described in OWL. The upper most 3 layers are seen by Berners-Lee’s Inferences and Proof-procedures.

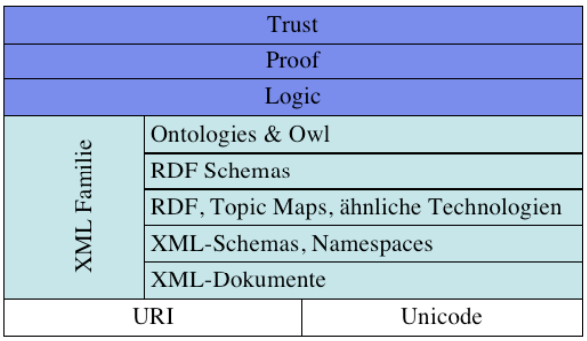



Figure 2. Berners-Lee's Layer Cake

### 2.1 The RDF-Model (Resource Description Framework

RDF is a data-model with the help of Web-based resources can be described. The basic idea of the model is that each resource can be described by means of a triple <Subject, Predicate, object>. The Subject of the triple must always be a resource, which can be unique identified through an URI (Universal Resource Identifier). The object can be a string or another resource. The predicate describes the relation between them.

A paragraph in Web, which for e.g. described who is he author of a certain Book, can be modelled in RDF as in the followings:

(1)
authorOf('http://www.w3.org/employee/id1321', 'http://www.books.org/ISBN0621')

*Prädikat* *Objekt* *Subjekt*

There are different formal languages with which we can serialise the RDF-model, among them, and the most appropriate for the goal of the Semantic Web is XML. The expression (1) is serialised in XML as follows:

```
(2) <rdf:Description
rdf:about:"http://www.w3.org/employee
e/id1321">
  <authorOf rdf:
ressource="http://www.books.org/ISBN0
621"/>
</rdf:Description>
```

### 2.2 RDFS (RDF Schema) and OWL

For the purposes of the Semantic Web only with RDF-modelling is not sufficient, as no information between the various predicates are given. For example: somebody searches all persons, which are authors or editors of a book about Semantic Web. With RDF-representation, one knows who is Author and who is Editor, but there is no information telling that "author" and „editor" are semantic related. (which means they are for example subclasses of a class „writer"). For this goal was developed RDFS. The language gives the basic elements with which one can describe Classes, subclasses, properties and sub properties, i.e. basic elements in ontology. RDFS gives no syntactical restrictions.

There are keywords such as : class, subPropertyof, subClassOf, etc.

<rdfs: Class rdf: about="Autor">

<rdfs: subClassOf rdf:ressource=#writer"/>

</rdfs: Class>

Unfortunately the expression power of RDFS is quite limited, therefore complicated class hierarchies and relationships between concepts cannot be expressed. Consequently was OWL designed, as stand alone language for ontology description. OWL has exchange syntax with RDF/XML. With OWL following descriptions are possible:

- taxonomical relationships between classes
- properties and data-types
- object properties
- instances of classes and properties

A collection of OWL expressions and the corresponding modelling of inference rules generate a knowledge base.

## 3 Ontologies and Multilinguality in Semantic Web Section

Ontologies and Text annotations can be used very successful for Web search. However when building ontological meta-data an important aspect to consider is the multilingual character of Web data. The number of documents in WWW written in other language than English increased dramatically during the last years. A recent study made by Netz-tipp.de (Netz-tipp 2002), based on the analysis of 2 million web sites, shows the increasing importance of German, French and Japanese among other languages.

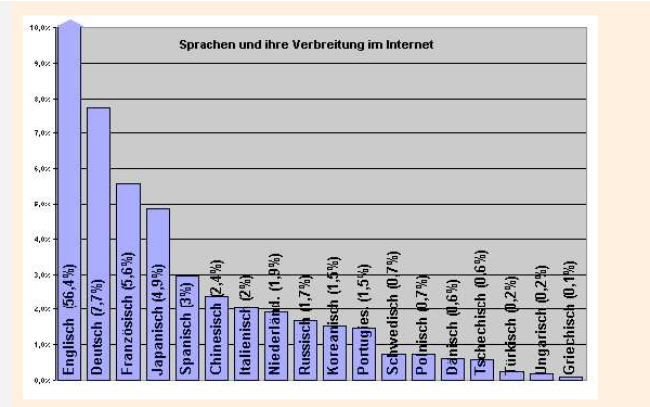



Figure 3 Distributions of languages in WWW

In the following we will explain how multilingual use of Semantic Web will function:

- the translation of web sites can be supported especially through the use of ontologies. The ontology provides the semantic connection between the represented objects and their properties. Examples for MAT-systems including ontological information are DBR-MAT as well as knowledge-based MAT-Systems or term-bases.
- Knowledge management can be also improved through web sites. Such an example is he development of resources for group, project or company knowledge, especially in multilingual form for international institutions
- International communication base for industry and commerce. Such an example are international lists of products, names of products or custom regulations

Basically, is the multilingual characteristic of WWW alone not enough motivation for development of multilingual ontologies. Until now the approach in Semantic Web is the following:

- Either the website makes reference to an English ontology.

  Advantage: Unambiguity

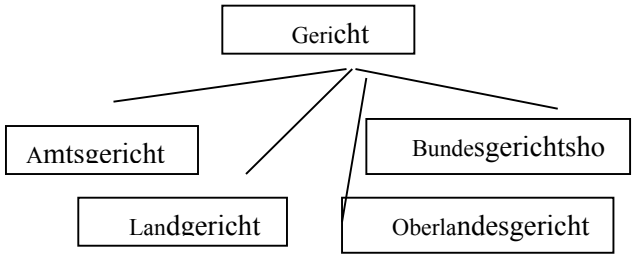
  Disadvantage: each non-English site must as a first Step make the mapping from its own language to the ontology. This mapping is sometimes difficult and when dealing with lexical gaps even impossible
- Or an ontology for each language is developed

  Advantage: no mapping between languages is necessary

  Disadvantage: no cross-lingual search, except some lexical search is possible.

Traditional Gruber's ontologies (Gruber 1992) do not distinguish between hierarchy of concepts and labels in each language ( names, terms). This is also the case with WordNet. Only recent developments in EuroWordnet use another approach, those one of creating a language independent ontology, on which the lexical material is mapped.

The following example illustrate what kind of problems may appear in cases where cultural specific facts (for e.g. in law, terms depending on the local juridical system), lexical/morphological terms and conceptual rules („Tribunals are classified according to an Instance", „there is always a national revision tribunal") cannot be inferred from the ontology




The situation would have been better if there would have been one hierarchy of facts or concepts, and the language dependent lexical terms (which are language and cultural specific) would have been linked through specific relation to the ontology.

This approach was followed in the MAT-system DBR-MAT (v. Hahn 1998). In this system recursive explanations in several languages were generated from a language independent ontology.

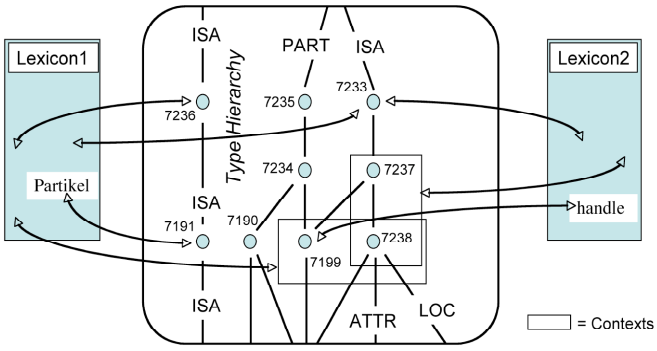



Figure 5. Ontology and lexical mapping in DBR-MAT

## 4 RDF Annotations and on-line translation in Semantic Web

Machine Translation, and in particular Example-based Machine Translation (Way and Carl 2003) can make use of the RDF additional annotations for three purposes:

1. For the achievement of parallel aligned corpora. Small languages still suffer from lack of linguistic resources, and especially multilingual resources. On-line documents are main source for machine-readable corpora; however, with few exceptions (explicitly translations of the same Web page) it is difficult to determine automatically which part of a document is a translation of another document. RDF annotations can be used for such purposes
2. For Example based rough translation: As mentioned in section 1 on-line translation is made for assimilation purposes, therefore, meaning preservation is much more important as an exact translation. RDF model aims to enrich documents with information about their content. This can help in the process of "example based rough translation". Until now, the trials in this field were done only on the basis of retrieval and translation of content-words (Shimhata, Sumita and Matsumoto 2003).
3. For disambiguation: the current example based translation systems make use only of syntactic annotation. These can be insufficient in disambiguation cases like the following:

Let us assume that we have in the database of translation examples:

Große Besonderheiten ↔ important peculiarities

Große Städte ↔ big cities

The translation choice for große Schlößer as important castles or big castles is context depending. For the moment the disambiguation is done only statistical. Semantic annotation of the examples, as well as the input text would increase the translation accuracy. This makes sense especially for translation of on-line resources, which are supposed to be correspondingly annotated

Although the advantages of Semantic Web annotations (in particular RDF-model) are transparent from the points mentioned above, the main question, which arises, is

Who will decide which semantic information has to be included, at what level (sentence /paragraph/document), and in which language?

Following information is needed for increasing the translation quality:

- translation equivalents of words /expressions
- transfer rules for syntactic structures
- semantic classes for the candidate solutions.

The main problem to be solved is the consistency between different RDF annotations corresponding to different users. Let us assume that in the German text the annotation for Große Städte is .

<rdf.description rdf. about:"http…..>
<user1: Messung > Große </user1: Messung >

and in the English one
<rdf.description rdf. about:"http…..>
<user2: size >big</user2: size >

A relationship between "size" and "Messung" has to be established showing that they refer to the same concept. This has to be done via mapping on an ontology. The main challenge in the design of ontologies with multilingual instances is that, very often words in one language overlap concepts in the ontology, and there is no one-to-one mapping to the meaning in the other language

The architecture in figure 2 proposes a framework for extracting translation correspondences, taking into account their RDF annotations (Vertan 2004). We propose the organisation of the RDF annotation scheme in two parts: syntactic annotation and semantic annotation. The concepts to be instantiated for this annotations will be organised in two correspondent ontologies.

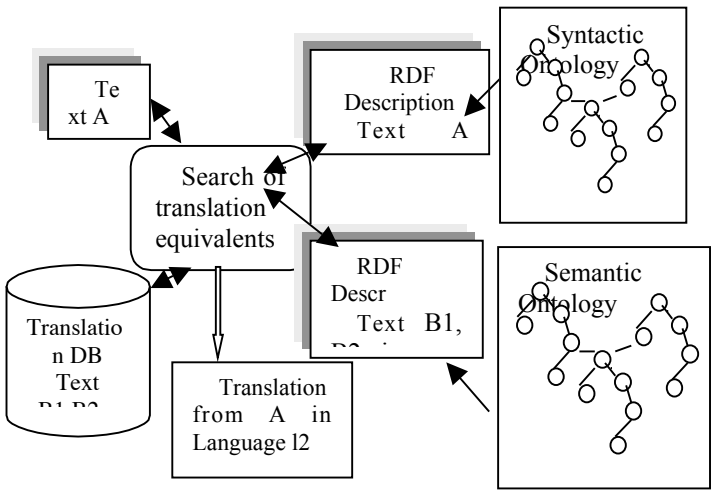



Figure 5: Extraction of Translation Equivalents from RDF annotated texts.

Assuming that input is a text A in language L1, a search process will identify fragments from A in the translation database and obtain one or more translations, namely Texts B1, B2,…Bn. During the next step the RDF descriptions of the input text and the translation candidates are compared by mapping the RDF annotations on the syntactic and semantic ontology, and the most similar one is chosen as output.

**References**

Berners-Lee, T.(1998). Semantic Web Road Map. An attempt to give a high-level plan of the architecture of the Semantic WWW. 1998. [cited 24.6.2005]. Available: http://www.w3.org/DesignIssues/Semantic.html.

A-Way, and M. Carl (2003), "Introduction to Example-based machine Translation", Kluwer Academic Press, 2003

M. Shimohata, and E Sumita, and Y. Matsumoto (2003), "Retrieving Meaning-equivalent Sentences for Example-based Rough Translation", HLT-NAACL Workshop: Building and using Parallel Texts. Data Driven Machine Translation and Beyond, Edmonton, May-June 2003, pp. 50-56

Netz-tipp (2002). Das Internet spricht Englisch und neuerdings auch Deutsch Sprachen und ihre Verbreitung im World-Wide-Web. [cited 24.6.2005]. Available: http://www.netz-tipp.de/sprachen.html.

C. Vertan (2004). Language Resources for the Semantic Web-Perspectives for the Machine Translation", Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training, Coling Conference Geneva, pg. 37-42.

T. Gruber (1992). A Translation Approach to Portable Ontology Specifications. Knowledge Systems Laboratory. Stanford Univ. Tech. Report KSL 92-71.

W. v.Hahn, (1998). Handling Multilingual Technical Terms in a Knowledge Based Translation System" In: Hansen, G. (Hrsg) LSP Texts and the Process of Translation. Kopenhagen. S. 31 – 57.

W. v.Hahn, C. Vertan (2005), „Mehrsprachiges Semantic Web und damit verbundene linguistische Aufgaben", in Proceedings of VAKKI'05, Finland (to appear)