# Web Semantization

Jan Dedek[1], Alan Eckhardt[1,2], Leo Galambos[1], and Peter Vojtáš[1,2]

[1] Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic
{dedek, eckhardt, galambos, vojtas}@ksi.mff.cuni.cz
[2] Academy of Sciences of the Czech Republic, Institute of Computer Science

**Abstract.** We understand Web Semantization as an automated process of increasing degree of semantic content on the web. Our idea is supported by models, methods, prototypes and experiments with a web repository, WIE with assisted learning, automated annotation tools producing third party semantic annotations, a semantic repository – a sample of semantized web and a proposal of an intelligent software agent.

**Key words:** Web Content Mining, Web Content Machine Processing, Annotation, Linguistic Analysis

## 1 Idea of Web semantization, prototypes and experiments

Our main idea is to fill a semantic repository with information that is automatically extracted from the web and make it available to software agents. We are working on a proof of concept that this idea is realizable and we give results of several experiments in this direction.

Our web crawler downloads a part of the web to the web repository. Page classifier selects those parts of web archive which are suitable for further semantic enrichment (we are able to enrich only a part of resources). More semantic content is created by several extractors and annotators in several phases.

**(1) The idea of a web repository.**

The web repository is a temporal repository of web documents crawled by a crawler. We use the web crawler Egothor 2.x (http://www.egothor.org/) and it's web repository. For pages from hidden web we have used RSS feeds.

**(2) The second idea is to split annotation process into two parts**, the first is domain independent intermediate annotation and the second is domain dependent user directed annotation.

**Domain independent intermediate annotation** can be done with respect to general ontologies. We distinguish two cases: (1) textual resources and (2) structured resources. For *textual resources* the ontology we use is the general linguistic PDT tectogrammatical structure [1] which captures semantic meaning of ordinary language sentences in Czech.

For *structured survey or product summary pages* we assume that their structure is often similar and the common structure can help us to detect data regions and data records and possibly also attributes and their values from detailed product pages. **Current solution** uses similarities of DOM structure of pages.

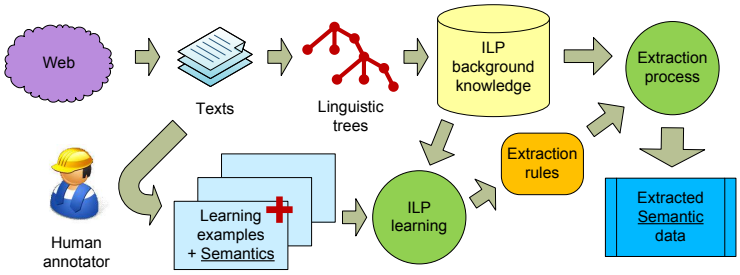


**Fig. 1.** ILP Learning of Extraction Rules

**Domain dependent annotation** is concerning only pages previously annotated by general ontologies. This makes second annotation faster and easier.

Repetitions in *textual pages* make possible to learn a mapping from structured tectogrammatical instances to an ontology. **Current solution** uses ILP tool `PROGOL` over annotations obtained in the first domain independent part. It is demonstrated in Fig. 1. For traffic accident reports, we were able to learn rules for finding sentences reporting on injuries.

For *structured pages* the domain dependent annotation **current solution** uses simple ontology in a form of relational schema provided by user.

**(3) Idea of semantic repository**. It should store all the semantic data extracted by extraction tools and accessed through a semantic search engine. **Current solution** uses [2]. It supports RDF storage and SPARQL querying.

**(4) Design of an software agent**, which will give the evidence that our semantization really improved general web search. Besides using annotated data it should also contain some user dependent preference search capabilities. **Current solution** exploits user preference modelling technique from our work [3].

### Conclusion

In this paper we have presented our work on web semantization – models and prototypes making the web to the web of things ([4]). Preliminary experiments are promising. This work was partially supported by Czech projects 1ET100300517, 201/09/0990 GACR and MSM-0021620838.

## References

1. Mikulová et al, M.: Annotation on the tectogrammatical level in the PDT. Technical Report 30, ÚFAL MFF UK, Prague, Czech Rep. (2006)
2. Dokulil, J., Tykal, J., Yaghob, J., Zavoral, F.: Semantic web infrastructure. In Kellenberger, P., ed.: First IEEE International Conference on Semantic Computing, Los Alamitos, California, IEEE Computer Society (2007) 209–215
3. Eckhardt, A., Horváth, T., Vojtáš, P.: Learning different user profile annotated rules. In: SUM. LNCS 4772, Springer (2007) 116–130
4. Berners-Lee, T.: The web of things. ERCIM News - Special: The Future Web **72** (January 2008) 3