European conference of the Czech Presidency of the Council of the EU
TOWARDS eENVIRONMENT (Challenges of SEIS and SISE: Integrating Environmental Knowledge in Europe)
http:/www.e-envi2009.org/proceedings/
J. Hřebíček, J. Hradec, E. Pelikán, O. Mírovský, W. Pilmmann, I.Holoubek, R. Legat (eds.)
Masaryk University, 2009

Web Information extraction for e-environment

Jan Dědek[a,b] **and Peter Vojtáš**[a,b]

[a] *Department of software engineering, Charles University, Prague, Czech Republic*
(jan.dedek@mff.cuni.cz, peter.vojtas@mff.cuni.cz)

[b] *Institute of Computer Science, Czech Academy of Sciences, Prague, Czech Republic*

Abstract: We will discuss possibility of using web information extraction methods for improving understanding eEnvironment relevant information on the web. Main contribution is in automated information extraction from web resources and annotation by an ontology.

Keywords: Web; Information extraction; Natural Language Processing.

1 INTRODUCTION AND MOTIVATION

Today a big amount of information is published electronically on the web. Good examples are pages of public institutions, which are publishing plenty of information on their web pages. This does not except information about the environment. In the Czech Republic for example the public right to information about the state of environment is guaranteed by the *Charter of Fundamental Rights and Basic Freedoms* (article 35/2) and the information is provided for example by *Ministry of the Environment* and by *local governments*.

This information is published in the form of natural language texts, which are suitable for human readers but not for computer processing. Computer processing of the information can be beneficial in many directions – statistical purposes, easy information search, integration of information from different sources, artificial intelligence automatic reasoning for new derived knowledge, automatic detection of contrary claims, complex visual presentation and publication, etc. All the benefits of machine understandable information sketched Tim Berners-Lee et al. [2001] in the famous article about the *Semantic Web*. Since then these ideas are being intensively developed within many activities of the *Semantic Web* foundation[1].

In the present paper we try to describe our method how the machine understandable data can be obtained form the web. We concentrated ourselves on the data and information relevant to the environment and we also present a practical experiment with extraction of information related to risks of environment's damage.

As described on the Figure 1 the web contains just partial information about environment and environmental risks. And just a part of it can be extracted by web information extraction tools but even though such information can make up interesting and important evidence. The fact that this evidence is kept in machine understandable form brings us all the advantages mentioned above.

The paper is structured as follows. Next section describes our extraction system in detail. Then section 3 is concentrated on the extraction method itself. The section 4 presents our experiment with environmentally relevant data and section 5 concludes the paper.

[1] http://www.w3.org/2001/sw/

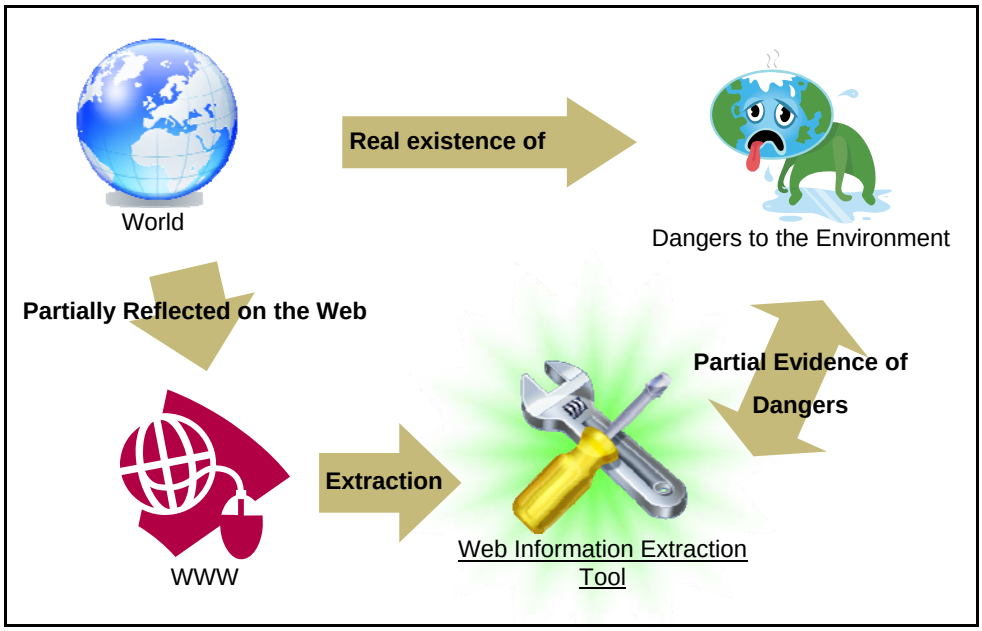


Figure 1. The Environment and Information on the Web

2 DESCRIPTION OF OUR TOOL / SYSTEM

We are developing a system, which produce machine (computer) processable information from the web. Our system captures text of web pages, annotates it linguistically, extracts data and stores the data in ontology data structure[2].

Our system covers an extraction process that starts on the web and ends in an ontology. This process consists of four steps. The Figure 2 describes them.

1) Extraction of relevant text
In this phase we have to extract the text from a page on the web. We use RSS feeds of the target web site. From the RSS we obtain URLs of particular articles (web pages) and we download them. From downloaded web pages we extract the desired text by means of a regular expression. This text is an input for the second phase.

2) Linguistic annotation
In this phase the linguistic annotation tools process the extracted text and produce corresponding set of linguistic trees representing the deep syntactic structure of individual sentences. We have used third party linguistic tools, which will be described in next section.

3) Information extraction
Our extraction method will be described in next section. This method uses the structure of linguistic trees and special extraction rules. The extraction rules define the target of extraction.

4) Semantic interpretation
This phase consists of transformation or conversion of the structured extracted data to the desired

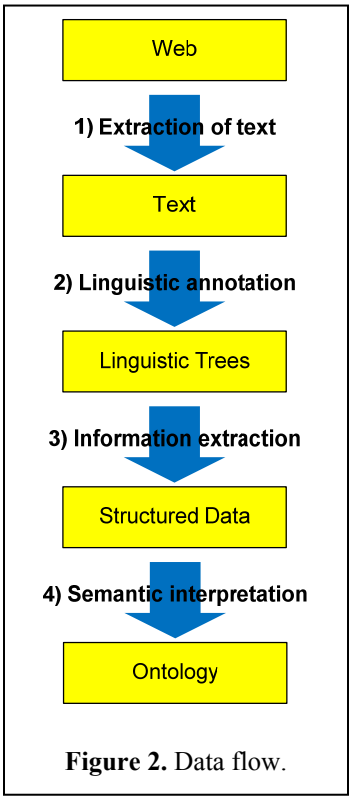


Figure 2. Data flow.

[2] Web Ontology Language http://www.w3.org/TR/owl-guide/

ontology format. It is quite important to choose suitable ontology that will properly represent semantics of the data. The interpretation expresses how to transform matching nodes of an extraction rule (and the available linguistic information connected) to the format of output ontology. Complexity of the transformation varies form simple (e.g. setting value of a data-type property to the value of some linguistic attribute) to complex. More details can be found in our previous work [Dědek and Vojtáš, 2008].

3 EXTRACTION METHOD

Our extraction system exploits set of linguistic tools published in [Hajič et al. 2006] and [Klimeš 2006]. These tools make tokenization, segmentation, morphological analysis, and linguistic parsing on so called *analytical* and *tectogrammatical* level of Czech. Although this linguistic analysis is addressed to the Czech language our extraction method is not limited to Czech. The project PEDT[3] uses the same linguistic theory for English and we plan to demonstrate usability of our method also with different linguistic approaches[4].

Our extraction method is based on extraction rules. These rules correspond to query requests of Netgraph[5] application. The Netgraph application is a linguistic tool used for searching through a syntactically annotated corpus of a natural language. Jiří Mírovský [2008] finished the development of Netgraph application (as his doctoral thesis) recently. Netgraph queries are written in a special query language. An example of Netgraph query will be described in next section.

The extraction works as follows: the extraction rule is in the first step evaluated by searching through a set of linguistic trees. Matching trees are returned and the desired information is taken from particular tree nodes.

Figure 3 shows linguistic (tectogrammatical) tree of sentence: "*Due to the clash the throat of fuel tank tore off and 800 litres of oil (diesel) has run out to a stream.*" (In Czech orignal: *Nárazem se utrhl hrdlo palivové nádrže a do potoka postupně vyteklo na 800 litrů nafty.*) Some of the nodes are emphasised – this should demonstrate how the extraction rule matches the tree.

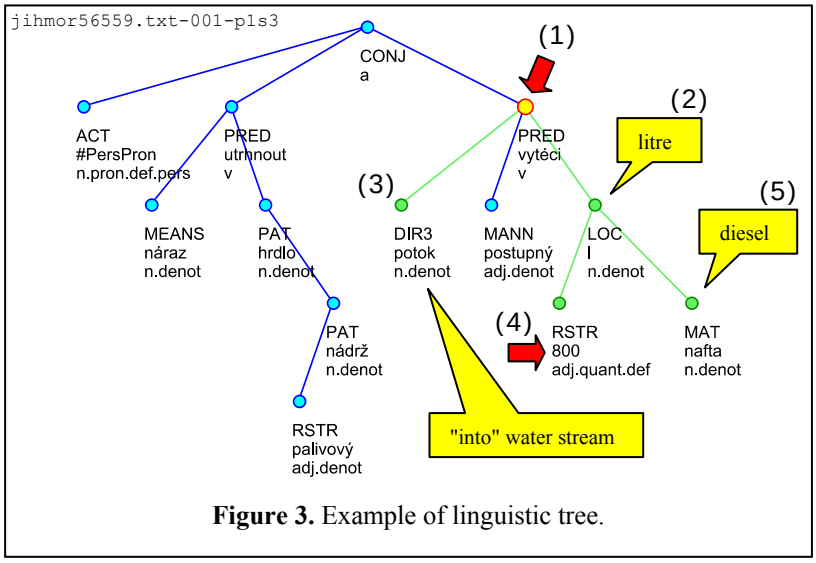


Figure 3. Example of linguistic tree.

[3] Prague English Dependency Treebank, http://ufal.mff.cuni.cz/pedt/
[4] *The Penn Treebank Project*, http://www.cis.upenn.edu/~treebank/ represents one of the suitable approaches. Its linguistic annotations are supported by many available automatic tools e.g. the *Stanford POS Tagger* and *Parser*, http://nlp.stanford.edu/software/
[5] http://quest.ms.mff.cuni.cz/netgraph/

4 EXPERIMENTS WITH ENVIRONMENTALLY RELEVANT DATA

For the purpose of the *TOWARDS eENVIRONMENT* conference we have made some experiments with acquisition of environmentally relevant information form the web.

We have used web reports from fire departments of several regions of the Czech Republic. These reports are written in Czech language and can be accessed through the web of *General Directorate of the Fire and Rescue Service of the Czech Republic*[6]. These reports describe interventions of fireman's units during different accidents (car accidents, fire accidents, etc.). These reports are rich in information, e.g. where and when a traffic accident occurred, which units helped, how much time it took them to show up on the place of accident, how many people were injured, killed etc. There is also information about endangered environment in the reports. Here we present an experiment with acquisition of environmentally relevant information from the fireman's web reports.

In the experiment we were interested in monitoring dangerous liquids (or materials) that have run out (spilled) during an accident. For this purpose we have designed an extraction rule that is depicted on the Figure 4. This rule consists of five nodes. Each node of the rule will match with some node in each suitable tree. So we can investigate the relevant information by reading values of linguistic tags of matching nodes. The fact, that this extraction rule match with some linguistic tree in most cases means that the tree (and corresponding sentence) deals with an amount of something which have run out (spilled). This is partially ensured by the node (1) – root of the extraction rule, which is restricted to contain one of the verbs "`uniknout`", "`unikat`" and "`vytéci`", which all have similar meaning to the English verbs *run out* or *spill*. From the other nodes we can find out the amount (node number 4), metric units (2) and material (5), which have run out (spilled) during an accident. And we can also identify the location where the dangerous material ended (expressed by the optional node number 3).

The extraction rule on the Figure 4 is designed manually by human expert, who has to be familiar with Netgraph application and linguistic formalism used. In [Dědek, Eckhardt, Vojtáš, 2008] we have presented an approach based on *inductive logic programming*. This approach makes it possible for almost unskilled user to tag relevant words in a sentence. Extraction rules are then learned automatically by the system.

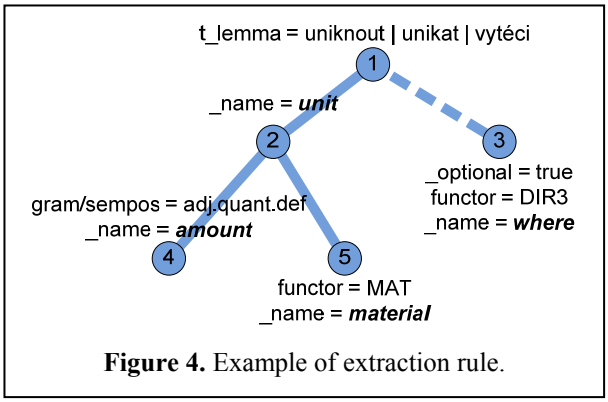


Figure 4. Example of extraction rule.

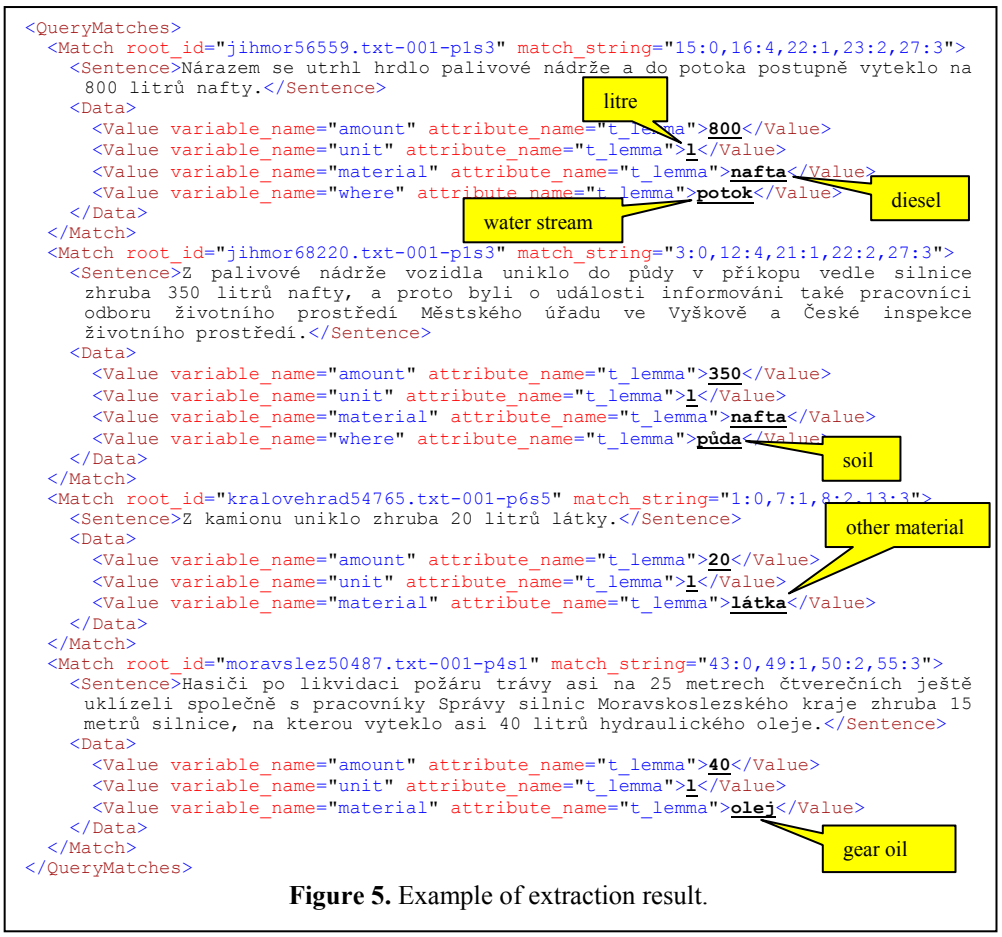
4.1 Results

Some example pieces chosen form the XML results are shown in the Figure 5. This example results contains four pieces of information extracted from four articles using the extraction rule on the Figure 4. Each piece of extracted information corresponds to a match of the extraction rule. Each match is closed in the `Match` element and each contains values

[6] http://www.hzscr.cz/

of some linguistic attributes closed inside the `Value` elements. Each value comes from some of the nodes of the extraction rule. Name of corresponding node is saved in the `variable_name` attribute of the `Value` node. The original text of matching sentence is closed inside the `Sentence` tag. We can see that the first matching sentence corresponds with the tree presented on the Figure 3 (with emphasized matching nodes).

Different materials (diesel, gear oil and some unspecified material – probably mentioned in related sentences) are showed up and different endangered types of location (water stream and soil) are identified in presented examples.




```
<QueryMatches>
  <Match root_id="jihmor56559.txt-001-p1s3" match_string="15:0,16:4,22:1,23:2,27:3">
    <Sentence>Nárazem se utrhl hrdlo palivové nádrže a do potoka postupně vyteklo na
     800 litrů nafty.</Sentence>
    <Data>
      <Value variable_name="amount" attribute_name="t_lemma">800</Value>
      <Value variable_name="unit" attribute_name="t_lemma">l</Value>
      <Value variable_name="material" attribute_name="t_lemma">nafta</Value>
      <Value variable_name="where" attribute_name="t_lemma">potok</Value>
    </Data>
  </Match>
  <Match root_id="jihmor68220.txt-001-p1s3" match_string="3:0,12:4,21:1,22:2,27:3">
    <Sentence>Z palivové nádrže vozidla uniklo do půdy v příkopu vedle silnice
     zhruba 350 litrů nafty, a proto byli o události informováni také pracovníci
     odboru životního prostředí Městského úřadu ve Vyškově a České inspekce
     životního prostředí.</Sentence>
    <Data>
      <Value variable_name="amount" attribute_name="t_lemma">350</Value>
      <Value variable_name="unit" attribute_name="t_lemma">l</Value>
      <Value variable_name="material" attribute_name="t_lemma">nafta</Value>
      <Value variable_name="where" attribute_name="t_lemma">půda</Value>
    </Data>
  </Match>
  <Match root_id="kralovehrad54765.txt-001-p6s5" match_string="1:0,7:1,8:2,13:3">
    <Sentence>Z kamionu uniklo zhruba 20 litrů látky.</Sentence>
    <Data>
      <Value variable_name="amount" attribute_name="t_lemma">20</Value>
      <Value variable_name="unit" attribute_name="t_lemma">l</Value>
      <Value variable_name="material" attribute_name="t_lemma">látka</Value>
    </Data>
  </Match>
  <Match root_id="moravslez50487.txt-001-p4s1" match_string="43:0,49:1,50:2,55:3">
    <Sentence>Hasiči po likvidaci požáru trávy asi na 25 metrech čtverečních ještě
     uklízeli společně s pracovníky Správy silnic Moravskoslezského kraje zhruba 15
     metrů silnice, na kterou vyteklo asi 40 litrů hydraulického oleje.</Sentence>
    <Data>
      <Value variable_name="amount" attribute_name="t_lemma">40</Value>
      <Value variable_name="unit" attribute_name="t_lemma">l</Value>
      <Value variable_name="material" attribute_name="t_lemma">olej</Value>
    </Data>
  </Match>
</QueryMatches>
```

Figure 5. Example of extraction result.

5 CONCLUSION

We have presented our system for extraction of information from Czech text on Web pages. Our system relies on linguistic annotating tools from ÚFAL[7] and the tree querying tool Netgraph. Our method is not limited to Czech and can be used with other languages and similar linguistic approaches. Although our system is still in development it can produce interesting results today.

We have shown an example how the environmentally sensitive information can be extracted from fireman's web reports and we are convinced that our method is applicable in any similar setting. As mentioned in our motivation: just partial information about the environment can be extracted from the web but even though such information can make up interesting and important evidence. The fact that this evidence is kept in machine understandable form brings us all the advantages of the semantic web technologies.

[7] Institute of Formal and Applied Linguistics in Prague, http://ufal.mff.cuni.cz/

ACKNOWLEDGEMENTS

This work was partially supported by Czech projects IS-1ET100300517, GACR-201/09/H057 and MSM-0021620838.

REFERENCES

Berners-Lee, T., Hendler, J. and Lassila, O., The Semantic Web. *Scientific American*, May 2001, 34–43.

Dědek, J., Eckhardt, A. and Vojtáš, P. Experiments with Czech linguistic data and ILP. *ILP 2008 - Inductive Logic Programming (Late Breaking Papers)*, Železný, F. and Lavrač, N. Prague, Czech Republic: Action M, 2008, 20–25.

Dědek, J. and Vojtáš, P., Computing aggregations from linguistic web resources: a case study in Czech Republic sector/traffic accidents. *Second International Conference on Advanced Engineering Computing and Applications in Sciences*, C. Dini, Ed. IEEE Computer Society, 2008, 7–12. http://www2.computer.org/portal/web/csdl/doi/10.1109/ADVCOMP.2008.17

Hajič, J. et al. Prague dependency treebank 2.0 cd-rom. *Linguistic Data Consortium LDC2006T01*, Philadelphia 2006.

Klimeš, V. Transformation-based tectogrammatical analysis of czech. *The 9th International Conference TSD 2006*, Lecture Notes In Computer Science 4188, 135–142. Springer-Verlag Berlin Heidelberg, 2006.

Mírovský, J. Netgraph – a Tool for Searching in the Prague Dependency Treebank 2.0, Doctoral Thesis (2008), *Institute of Formal and Applied Linguistics*, Faculty of Mathematis and Physics, Charles University in Prague, 2008. http://quest.ms.mff.cuni.cz/netgraph/publications.html