

Web Information Extraction for eEnvironment

Jan Dědek

Peter Vojtáš

- Department of software engineering, Charles University in Prague
- Institute of Computer Science Czech Academy of Sciences

European conference of the Czech Presidency of the Council of the EU

TOWARDS eENVIRONMENT 2009

Opportunities of SEIS and SISE: Integrating Environmental Knowledge in Europe

The Environment and Information on the Web



World

Real existence of



Dangers to the Environment

Partially Reflected on the Web



WWW

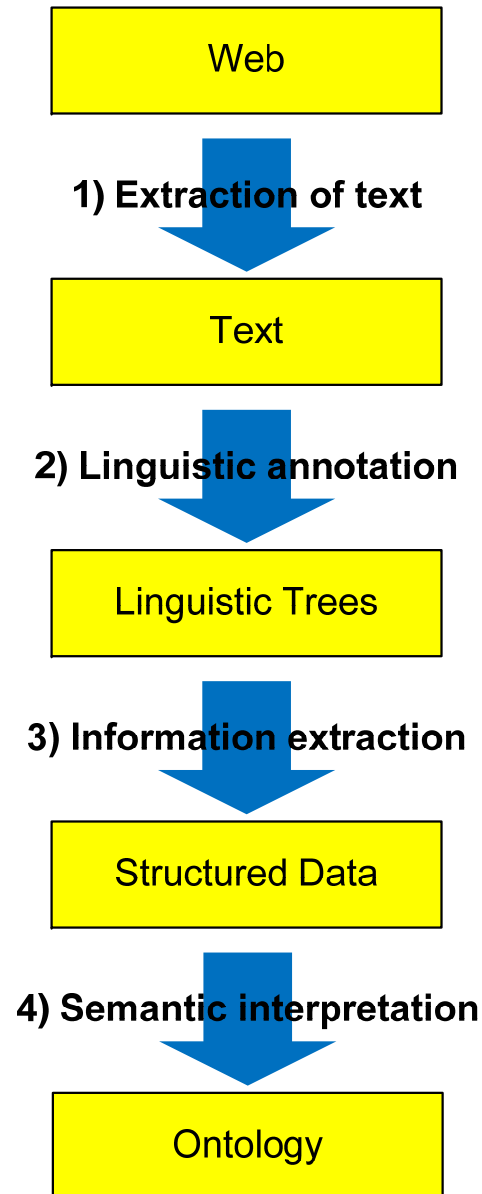
Extraction



Web Information Extraction
Tool

Partial Evidence of
Dangers

The Data Flow



Example of processed web page



Zpravodajství

Informace z resortu o tom, co se stalo, co se děje i co se připravuje

home navigace vyhledávání změna vzhledu

HZS Jihomoravského kraje

Zubatého 1, 614 00 Brno, telefon 950 630 111,
<http://www.firebrno.cz>
Zpravodajství v roce 2006



15.05.2007

V trabantu zemřeli dva lidé

K tragické nehodě dnes odpoledne hasiči vyjžděli na silnici z obce Česká do Kuřimi na Brněnsku.

Nehoda byla operačnímu středisku HZS ohlášena ve 13.13 hodin a na místě zasahovala jednotka profesionálních hasičů ze stanice v Tišnově. Jednalo se o čelní srážku autobusu Karosa s vozidlem Trabant 801. Podle dostupných informací trabant jedoucí ve z Brna do Kuřimi zřejmě vyjel do protisměru, kde narazil do linkového autobusu dopravní společnosti ze Žďáru nad Sázavou. Ve zdemolovaném trabantu na místě zemřeli dva muži – 82letý senior a další muž, jehož totožnost zjišťují policisté.

Hasiči udelali na vozidle protipožární opatření a po vyšetření a zadokumentování nehody dopravní policií vrak trabantu zaklesnutý pod autobusem pomocí lana odtrhli. Po odstranění střechy trabantu pak z kabiny vyprostili těla obou mužů. Obě vozidla – trabant i autobus, pak postupně odstranili na kraj vozovky a uvolnili tak jeden jízdní pruh. Unik provozních kapalin nebyl zjištěn. Po 16. hodině pomohli vrak trabantu naložit k odtahu a asistovali při odtažení autobusu. Po úklidu vozovky krátce před 18.30 hod. místo nehody předali policistům a ukončili zásah.



Odkazy

Hasiči

- Generální ředitelství
- hl. m. Praha
- Jihočeský kraj
- Jihomoravský kraj
- Karlovarský kraj
- Královéhradecký kraj
- Liberecký kraj
- Moravskoslezský kraj
- Olomoucký kraj
- Pardubický kraj
- Píseňský kraj
- Středočeský kraj
- Ústecký kraj
- kraj Vysočina
- Zlínský kraj



V této rubrice Zpravodajství

- Aktualizace stránek
- Archiv zpravodajství
- Bleskové zpravodajství
- RSS
- Boj proti korupci
- Digitální televize
- Hasiči
- Hlavní zprávy
- Ministerstvo
- Od dopisovatelů (neoficiální)
- Police
- Regiony
- Servis nejen pro novináře
- Schengenská spolupráce
- WebEditorial

Na našem serveru v jiných rubrikách

- Aktuality Národního archivu

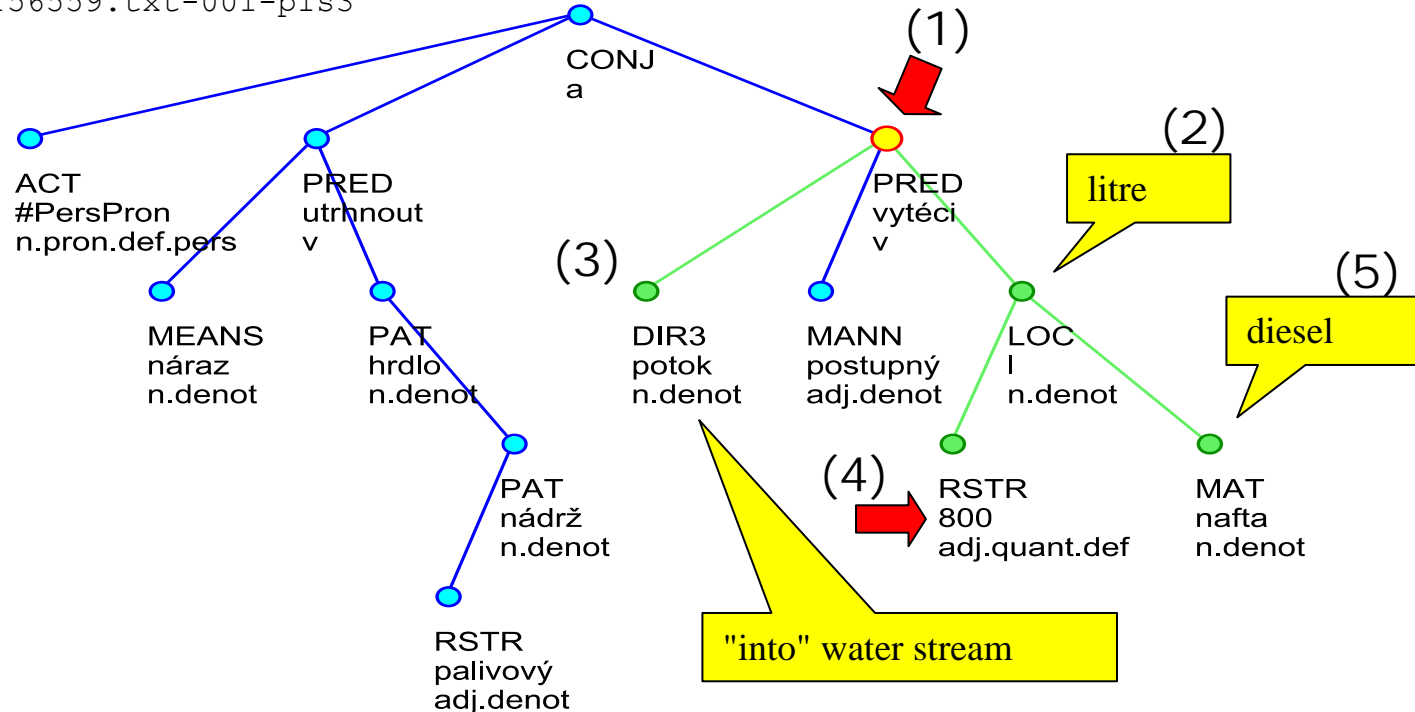
Relevant text

Example of a linguistic tree

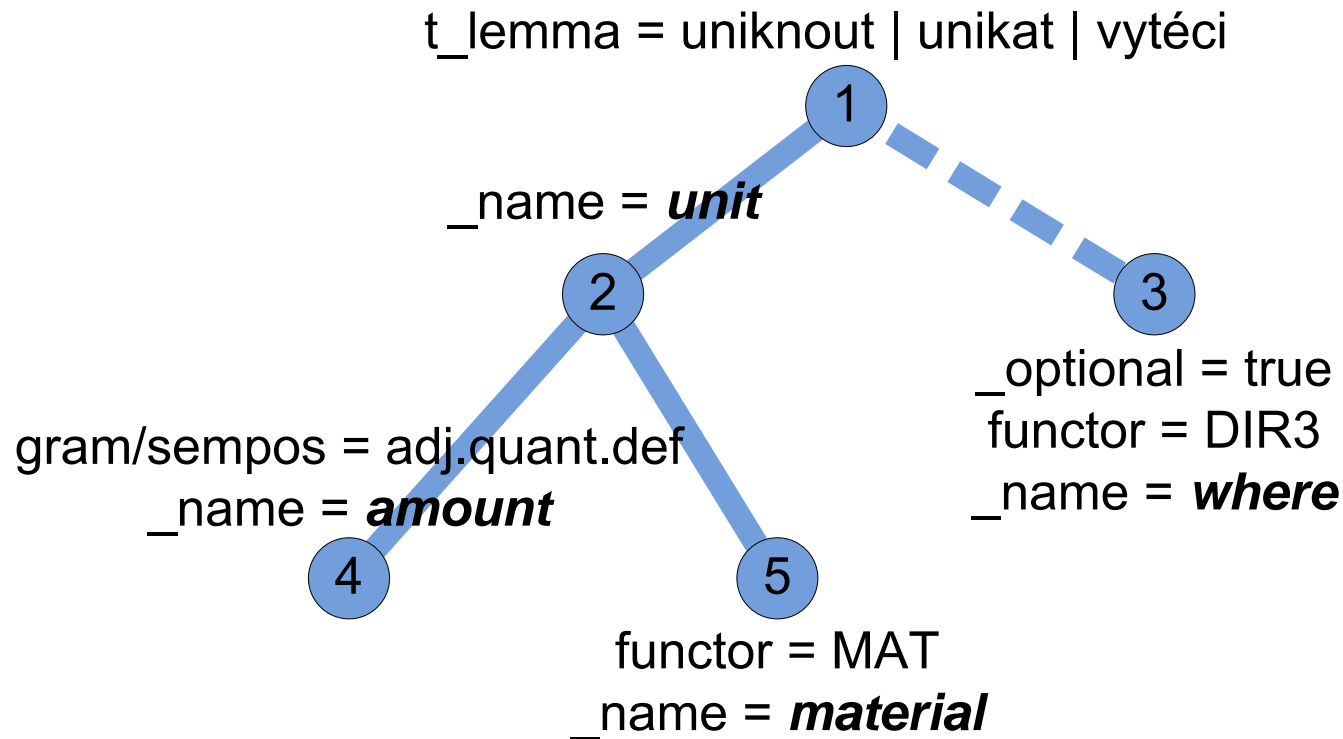
"Due to the clash the throat of fuel tank tore off and 800 litres of oil (diesel) has run out to a stream."

"Nárazem se utrhlo hrdlo palivové nádrže a do potoka postupně vyteklo na 800 litrů nafty."

jihmor56559.txt-001-pls3



Example of an extraction rule.



Experimental results (1)

```
<QueryMatches>
  <Match root_id="jihmor56559.txt-001-pls3" match_string="15:0,16:4,22:1,23:2,27:3">
    <Sentence>Nárazem se utrhlo 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">1</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-pls3" 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">1</Value>
      <Value variable_name="material" attribute_name="t_lemma">nafta</Value>
      <Value variable_name="where" attribute_name="t_lemma">půda</Value>
    </Data>
  </Match>
```

litre

water stream

diesel

soil

Experimental results (2)

```
...
<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>
```

other material

gear oil

What is interesting on the Web?

- For **environment specialists**?
- What information from the Web can help with the **evidence**, **inspection** and **care** for the environment?
- Perhaps our method can provide it!