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

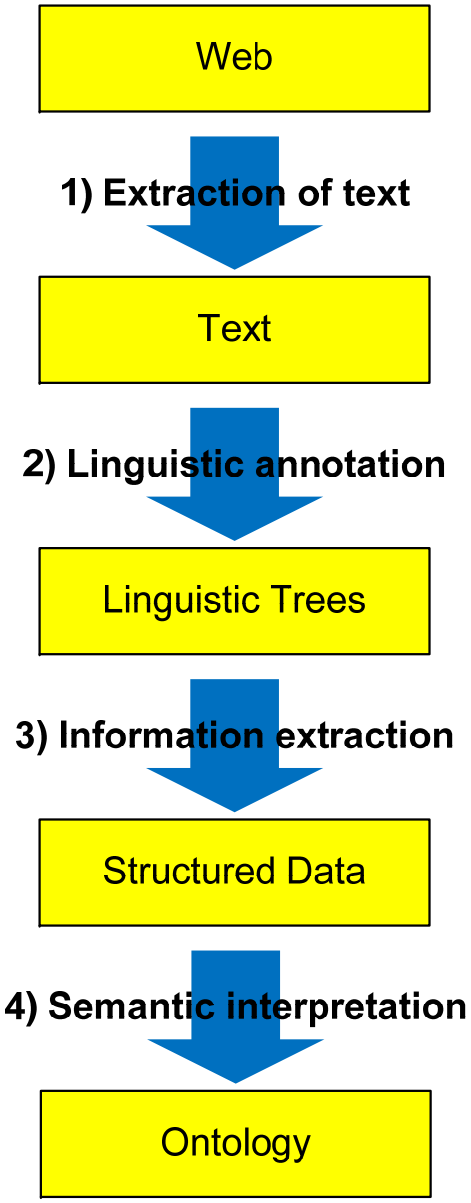
WWW

Extraction

Web Information Extraction Tool

Partial Evidence of Dangers

The Data Flow



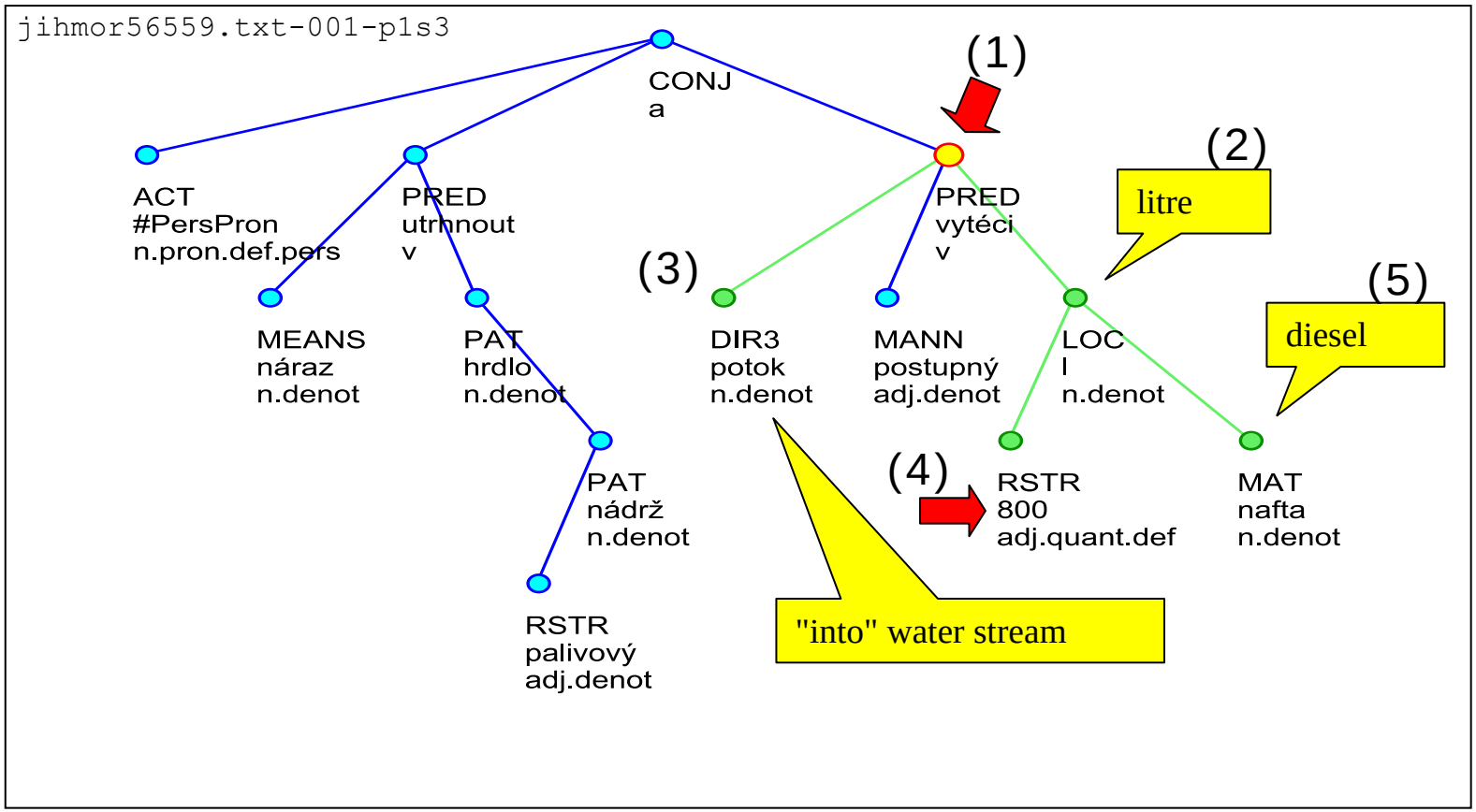
Example of processed web page

Relevant text

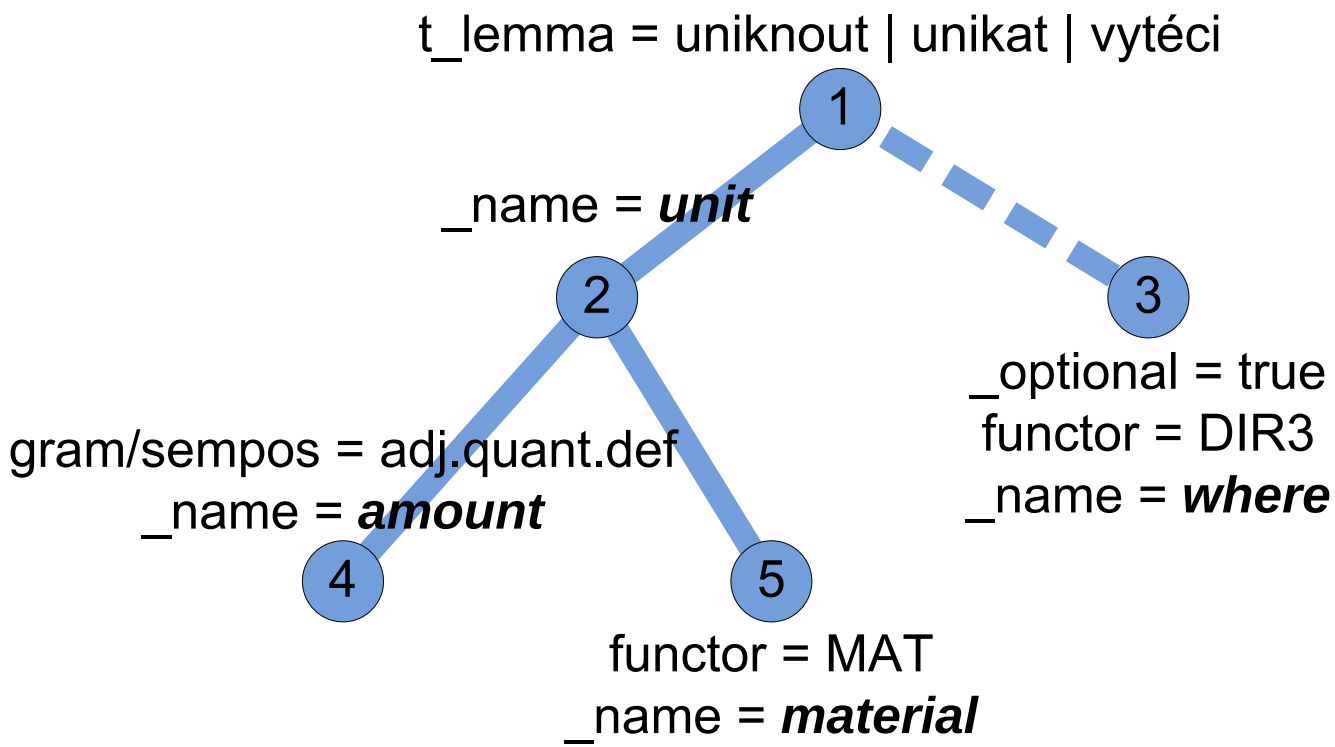
Example of a linguistic tree

"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 utrhl hrdlo palivové nádrže a do potoka postupně vyteklo na 800 litrů nafty."



Example of an extraction rule.



Experimental results (1)

```
<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>
...
```

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!