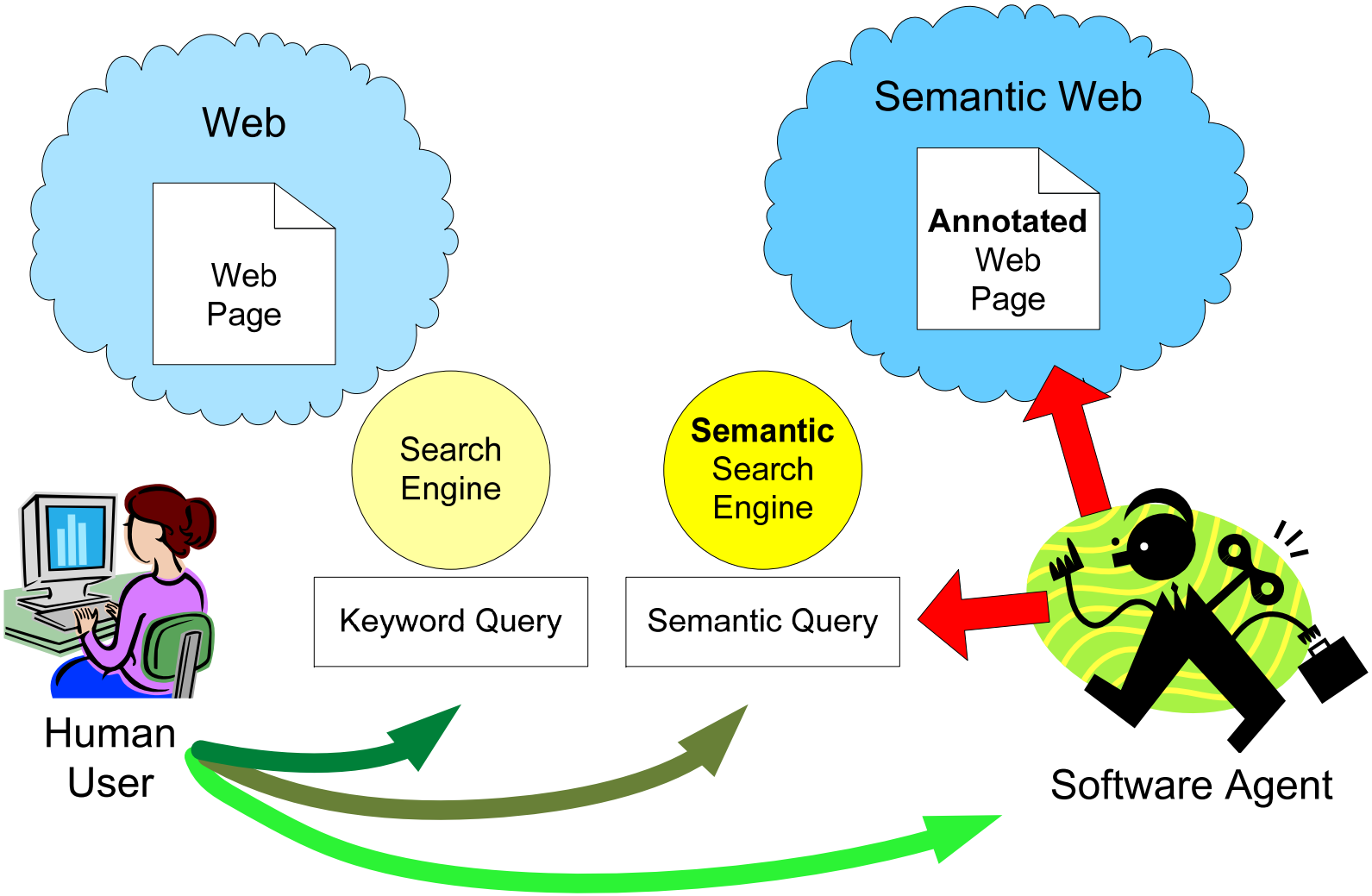
Sémantická extrakce a anotace informací
→ z Webu,
→ pro Sémantický Web

Jan Dědek, (Prof. Peter Vojtáš)

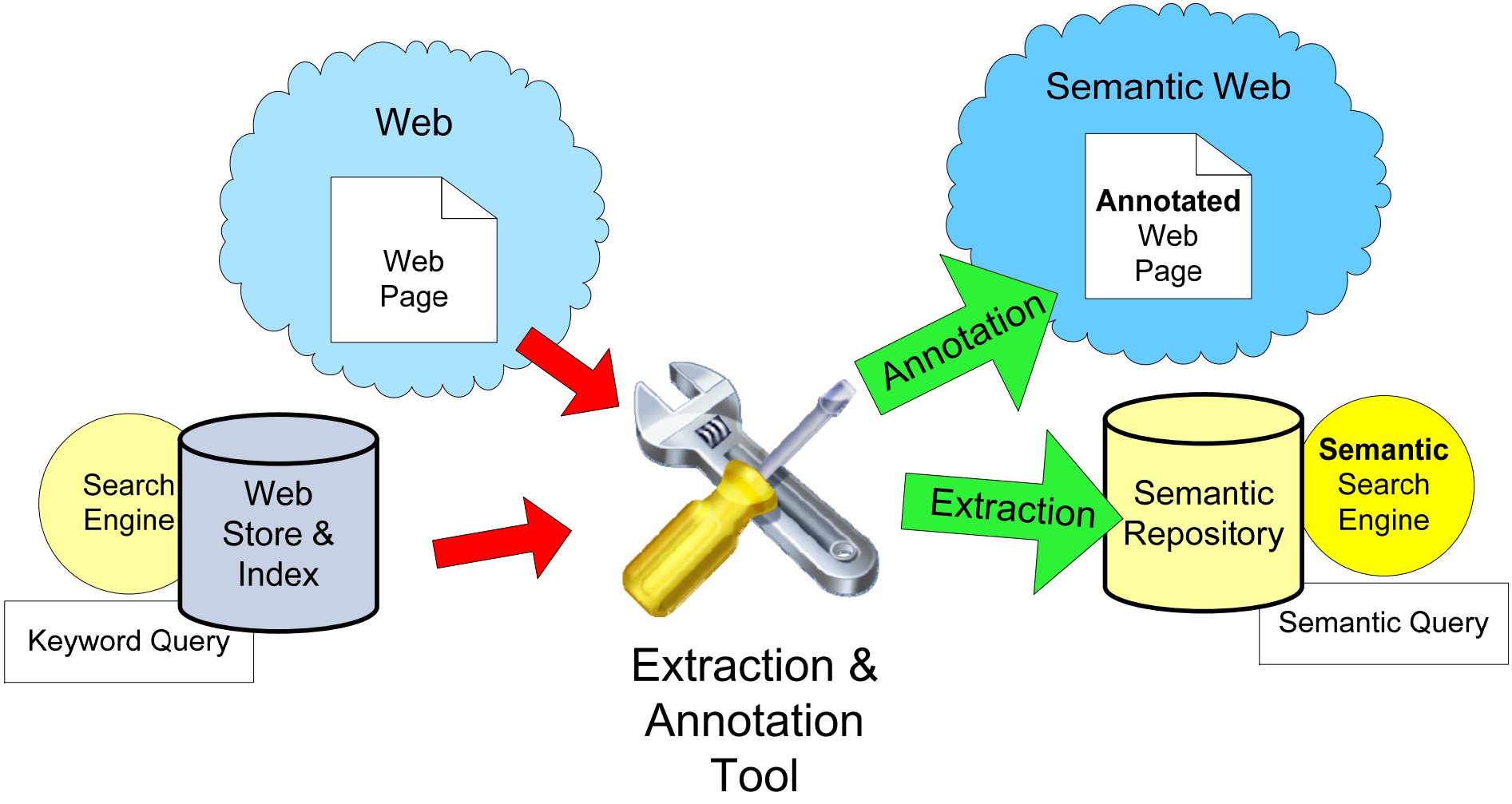
Res Informatica – Informační schůzka

10. 4. 2009

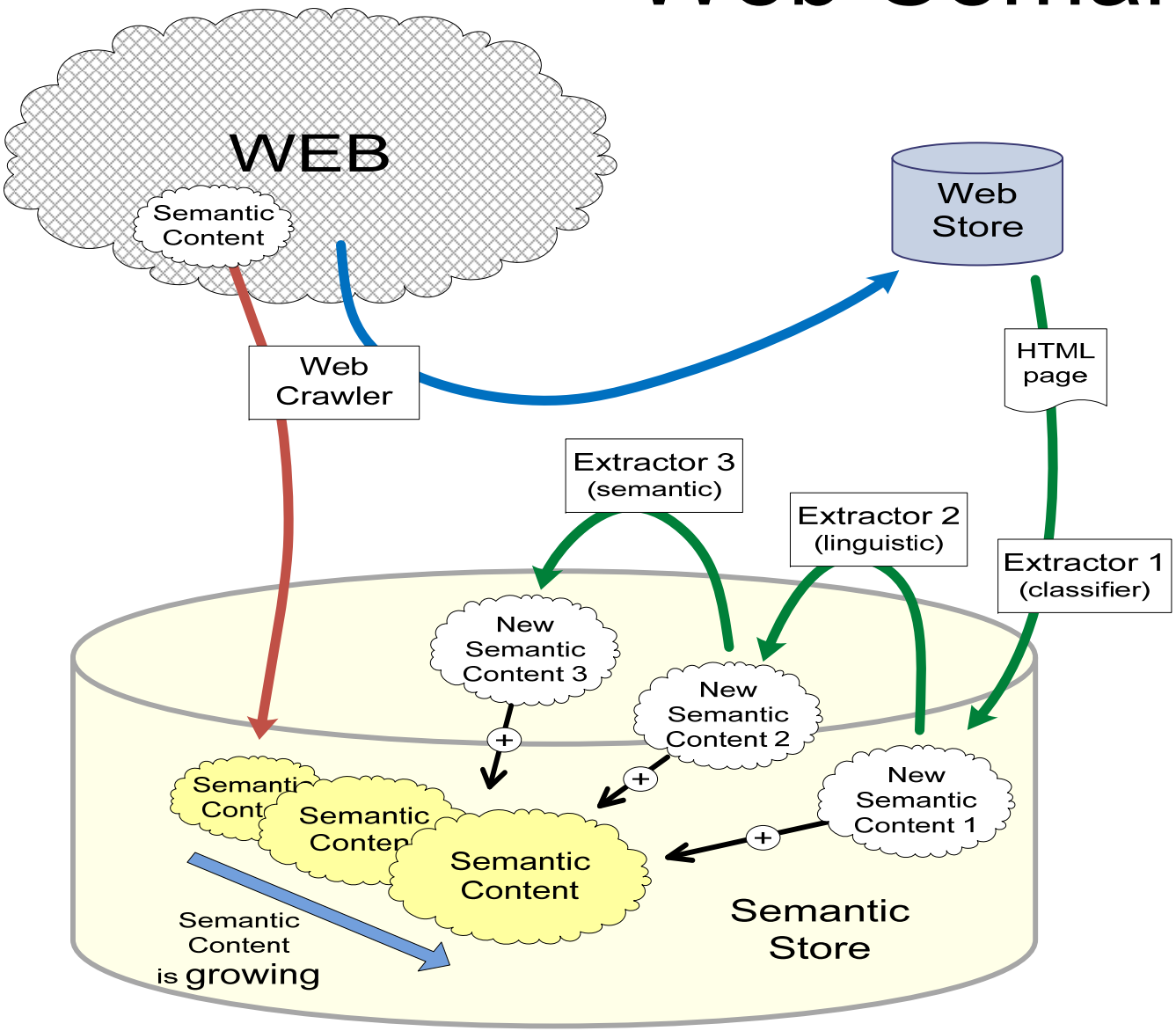
Looking for information on the Web



Extraction & Annotation Tool



Web Semantization



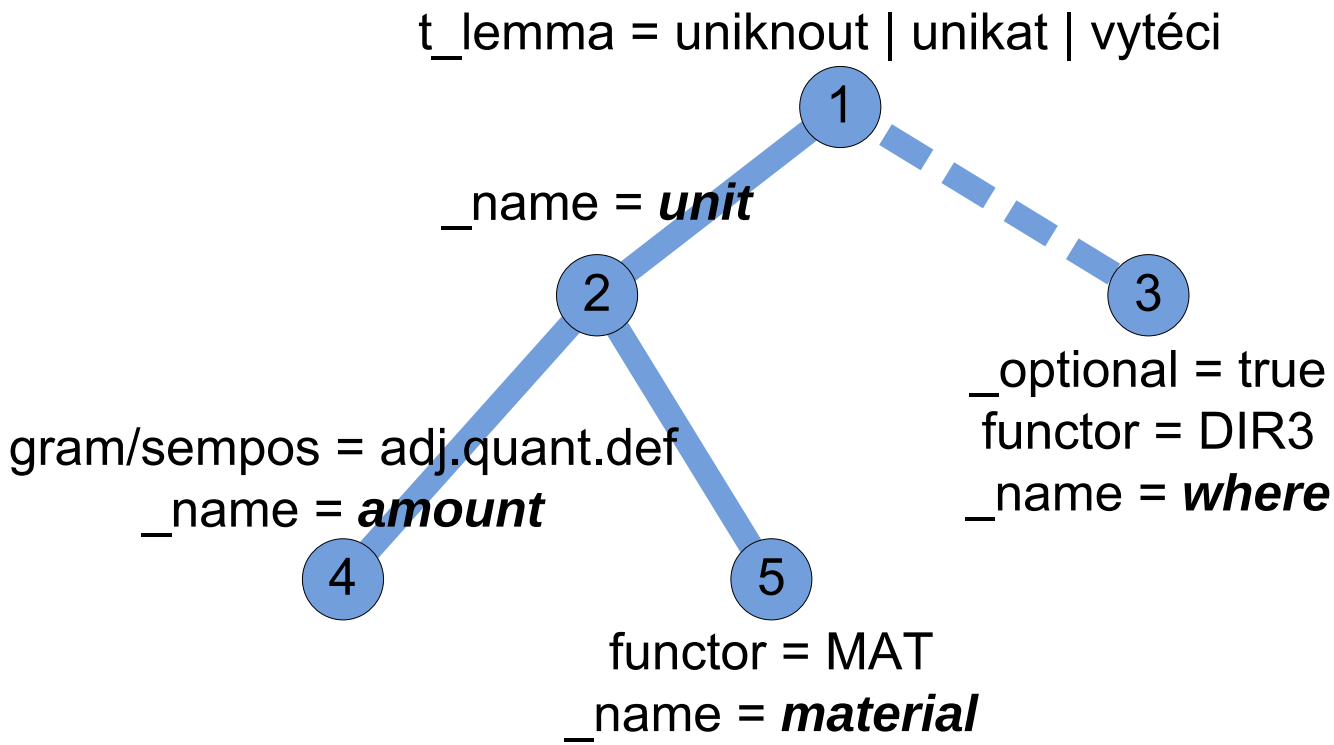
Konkrétněji – co děláme, *chceme dělat*

- Vývoj nástroje pro extrakci informací
  - Z textových stránek – lingvistické metody
  - Z HTML tabulek, stránek se stejnou „šablonou“
  - Sémantická reprezentace výsledků (ontologie)
- *Práce na infrastruktuře pro **sémantizaci***
  - *Krawlování a archivace webových stránek*
  - *Klasifikace stránek*
    - *Podle domény*
    - *Podle struktury (tabulka nebo text)*
  - *Extrakce efektivního textu stránky*

Současné zkušenosti

- Lingvistická analýza textů
  - Čeština (PDT nástroje)
  - Angličtina (GATE framework, Stanford parser)
- Rule-based lingvistická extrakční metoda
  - Pravidla převzatá z aplikace Netgraph (ÚFAL)
- Induktivní Logické Programování
  - Převod lingvistické struktury na „logický program“ pro ILP
  - Učení pravidel pro detekci relevantních slov
    - Počty zraněných při nehodách
- Extrakce z webových stránkánek e-shopů
  - Detekce datových regionů a datových záznamů
  - Heuristika založená na opakování podobných vzorů v DOM stromu

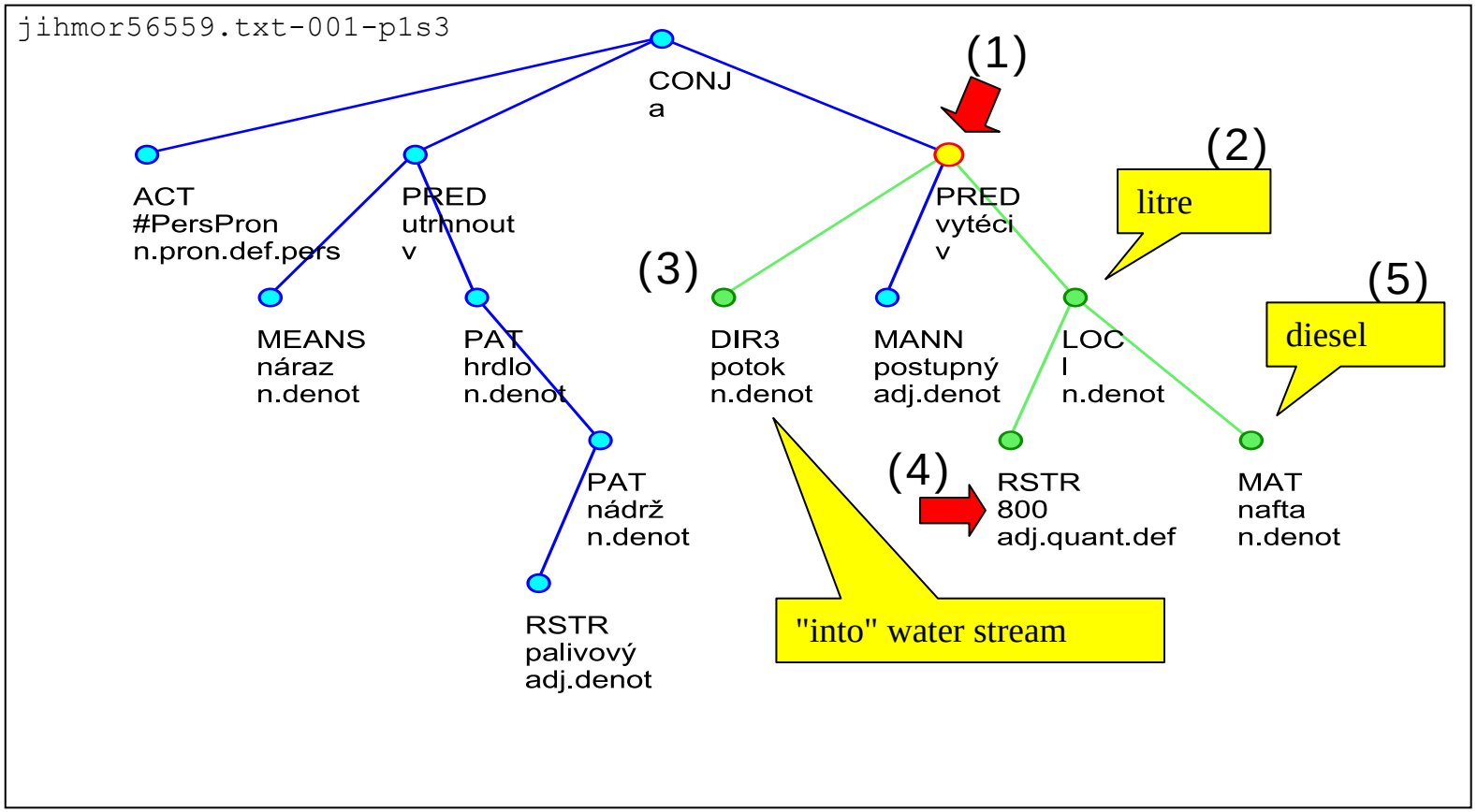
Example of an extraction rule.



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."



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>
...
```

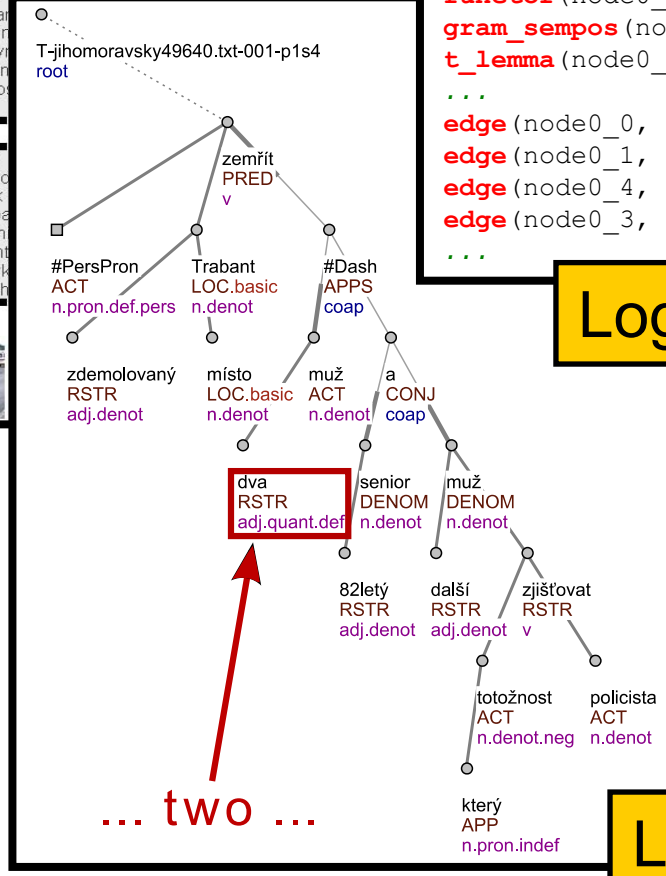
litre

water stream

diesel

soil

Source web page




```
tree_root(node0_0). node(node0_0).
id(node0_0, t_jihomoravsky49640_txt_001_p1s4).
%%%%%%%% node0_1 %%%%%%%%%%%%%%%%%%
node(node0_1).
functor(node0_1, pred).
gram_sempos(node0_1, v).
t_lemma(node0_1, zemrit).
%%%%%%%% node0_2 %%%%%%%%%%%%%%%%%%
node(node0_2).
functor(node0_2, act).
gram_sempos(node0_2, n_pron_def_pers).
t_lemma(node0_2, x_perspron).
%%%%%%%% node0_3 %%%%%%%%%%%%%%%%%%
node(node0_3). id(node0_3,
functor(node0_3, loc).
gram_sempos(node0_3, n_denot).
t_lemma(node0_3, trabant).
...
edge(node0_0, node0_1). edge(node0_1, node0_2).
edge(node0_1, node0_3). edge(node0_3, node0_4).
edge(node0_4, node0_5). edge(node0_3, node0_6).
edge(node0_3, node0_7). edge(node0_3, node0_8).
...
```

Logic representation

Linguistic trees