



# Sémantická extrakce a anotace informací

→ z Webu,

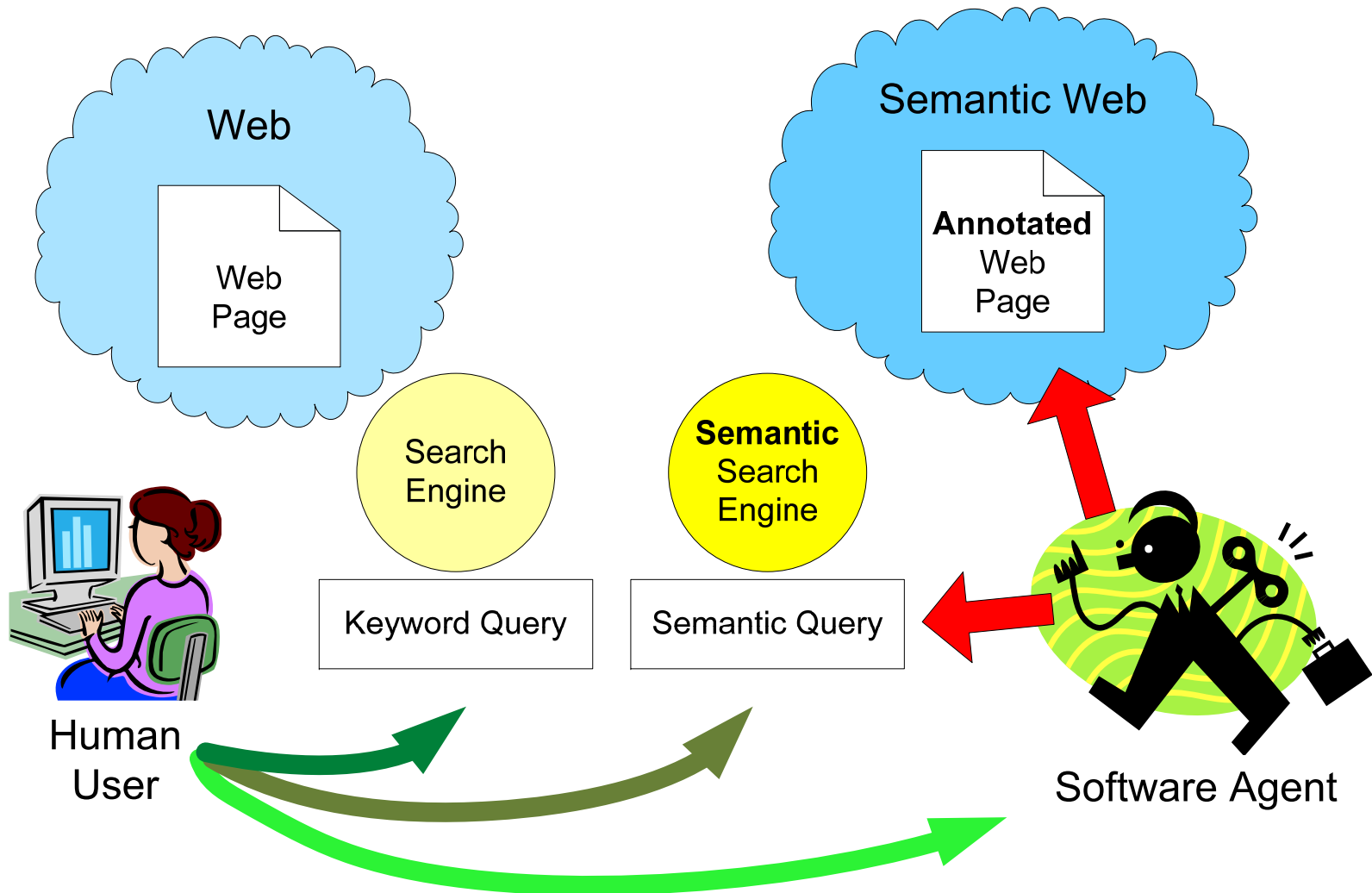
→ pro Sémantický Web

Jan Dědek, (Prof. Peter Vojtáš)

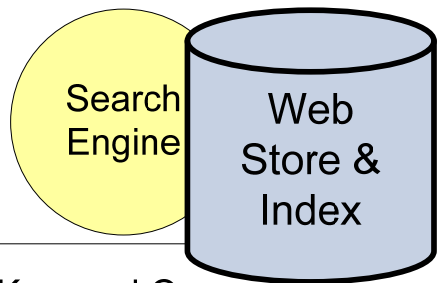
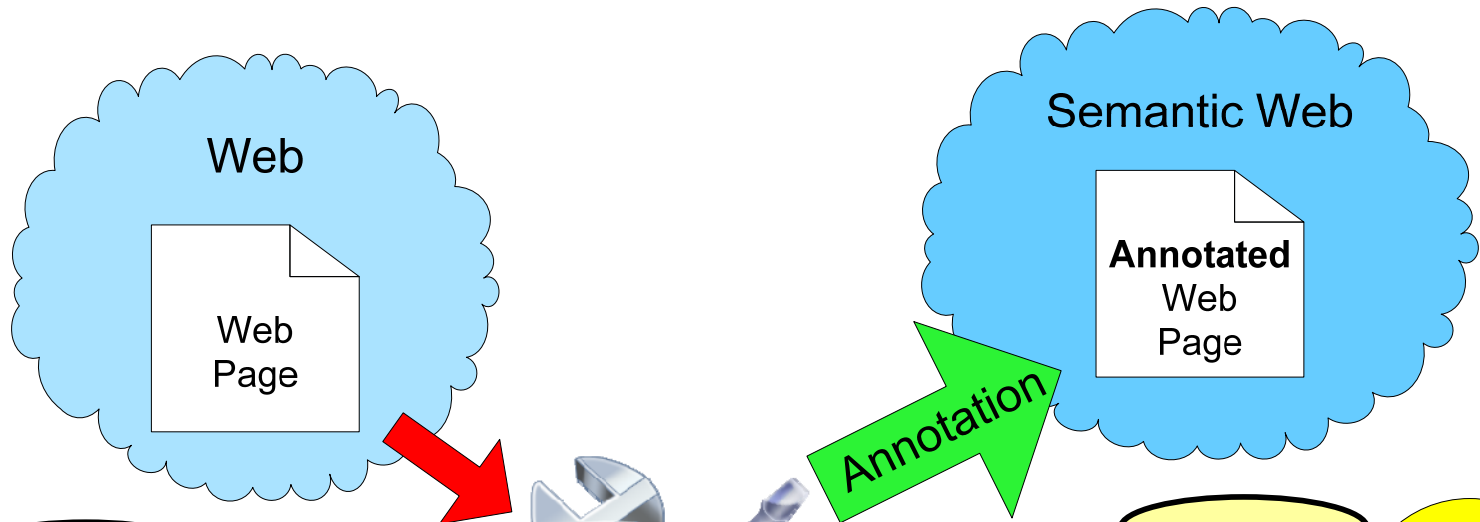
Res Informatica – Informační schůzka

10. 4. 2009

# Looking for information on the Web



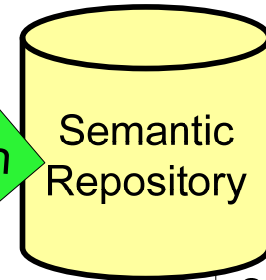
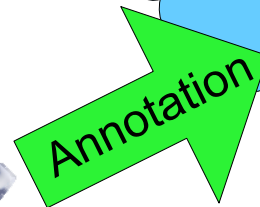
# Extraction & Annotation Tool



Keyword Query



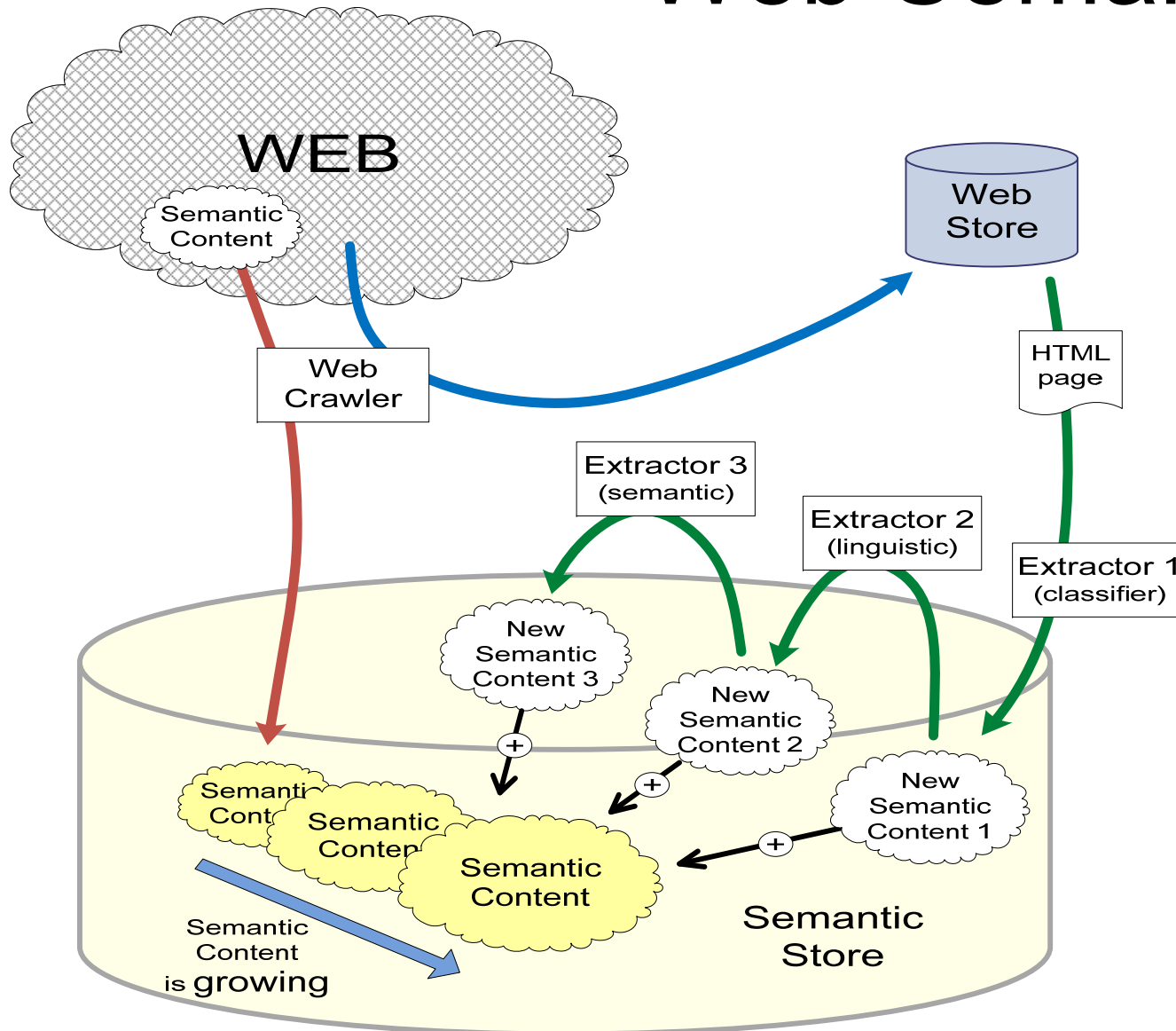
Extraction & Annotation Tool



Semantic Search Engine

Semantic Query

# Web Semantization



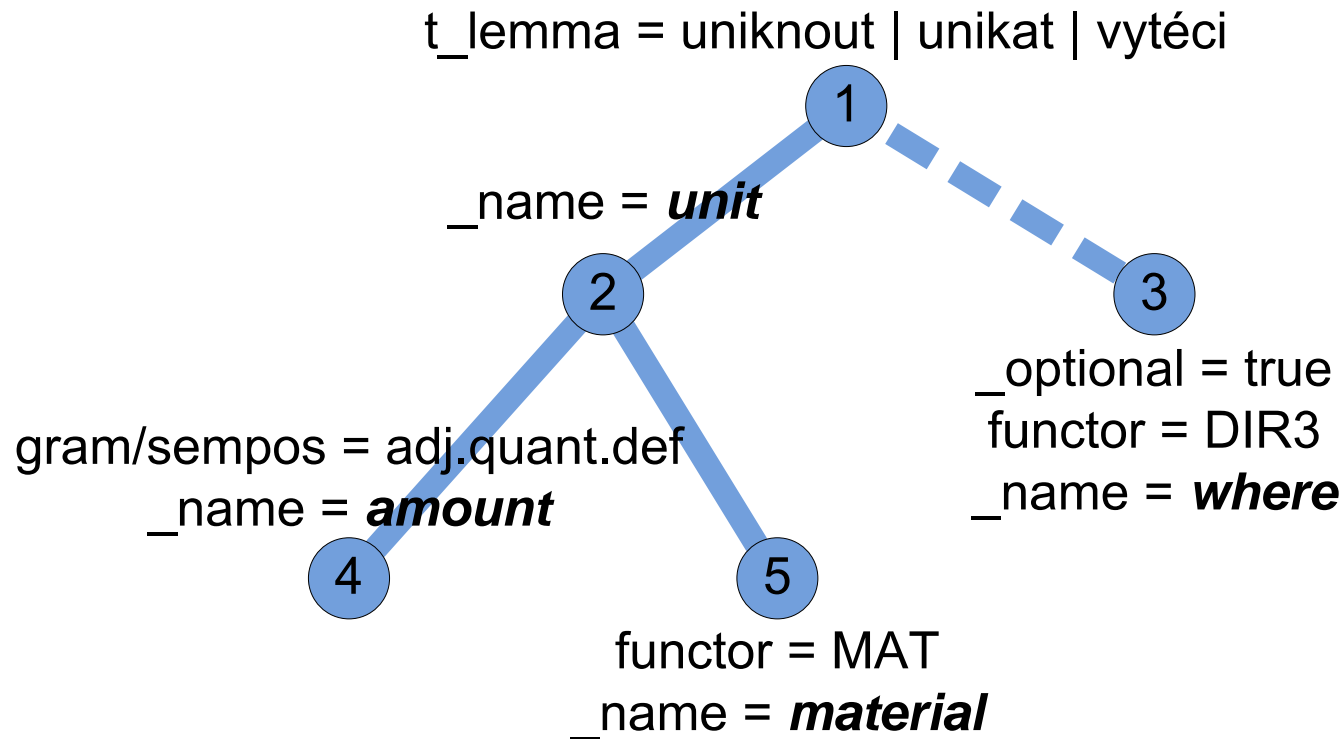
# Konkrétněji – co děláme, *chceme dělat*

- 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 stránek e-shopů**
  - Detekce datových regionů a datových záznamů
  - Heuristika založená na opakování podobných vzorů v DOM stromu

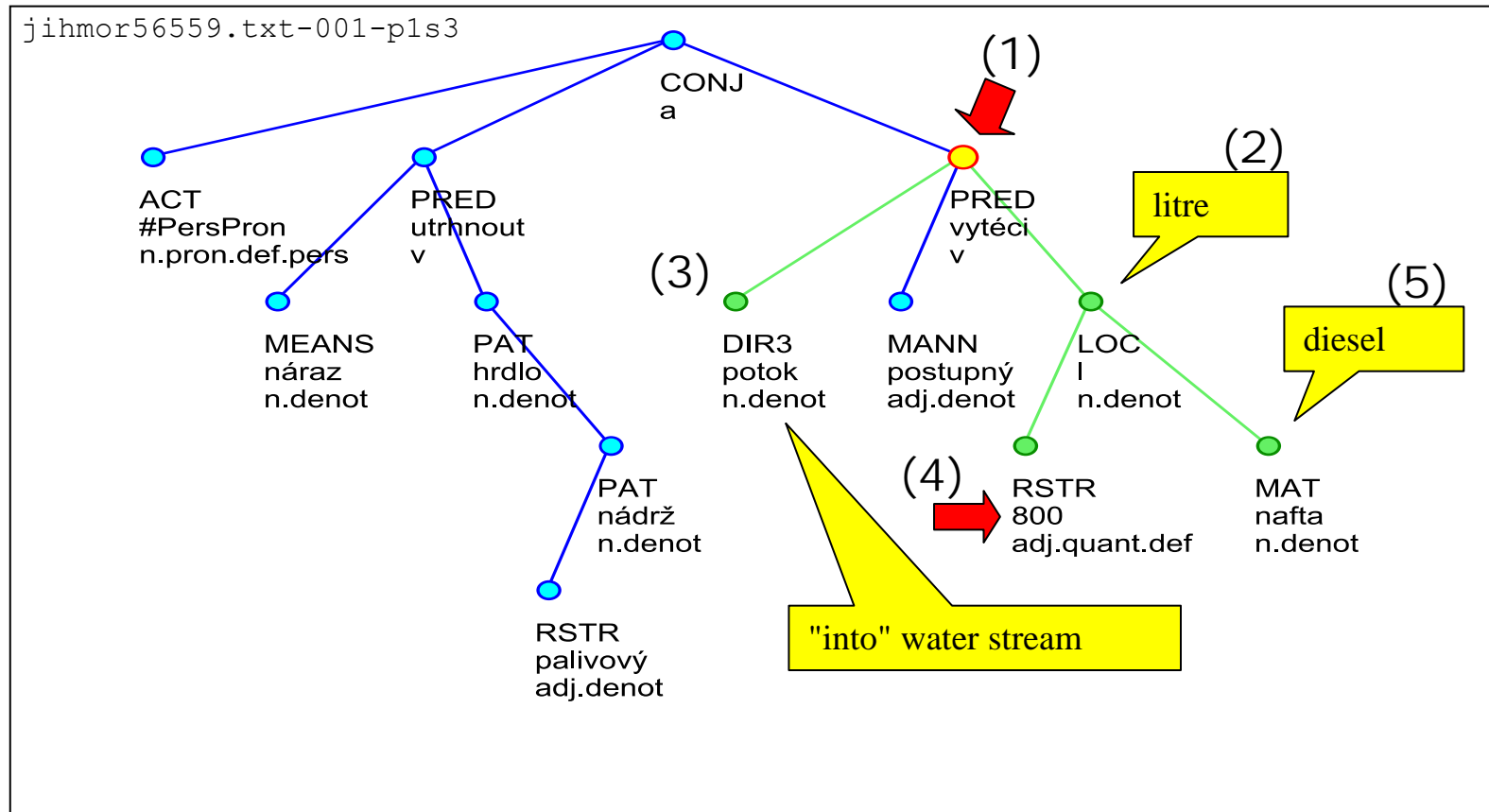
# Example of an extraction rule.



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





# 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 utrhł 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



# Zpravodajství

## 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

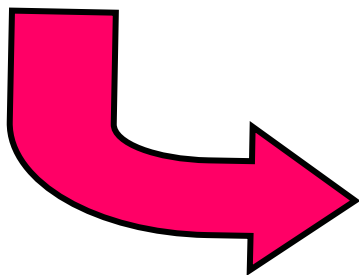
*K tragické nehodě dnes odpoledne hasiči vyjžděli na silnici z obce Česká do Kuřimi na Brněnsku.*

Nehoda byla operačnímu středisku HZS ohlášena ve 13.13 hodin a na místě zasahovala jednotka profesionálních hasičů ze stanice Tišnově. Jednalo se o čelní srážku autobusu Karosa s vozidlem Trabant 601. 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 udělali na vozidle protipožární opatření a po vyšetření zadokumentování nehody dopravní policii vrak trabantu zaklesnutý pod autobusem pomocí lana odtrhli. Po odstranění střechy trabantu pak kabiny vyprostili těla obou mužů. Obě vozidla – trabant i autobus, postupně odstranili na kraj vozovky a uvolnili tak jeden jízdní pruh. Únik provozních kapalin nebyl zjištěn. Po 16. hodině pomohli vrak trabantu naložit k odtahu a asistovali při odtažení autobusu. Po úklidu vozovky krátce před 16.30 hod. místo nehody předali policistům a ukončili zásah.



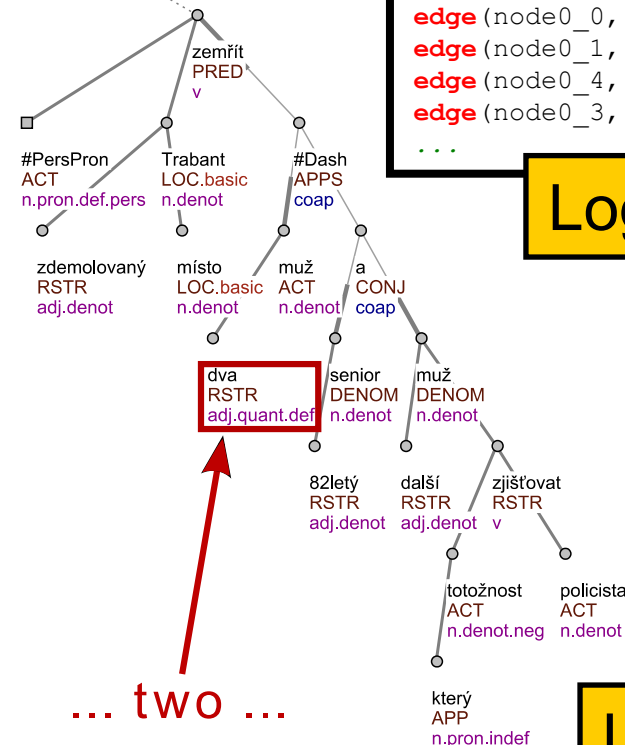
Source web page



### Odkazy

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

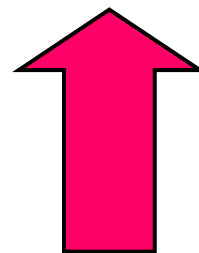
T-jihomoravsky49640.txt-001-p1s4  
root



... two ...

```
tree_root(node0_0). node(node0_0).
id(node0_0, t_jihomoravsky49640_txt_001_p1s4).
node0_1 node0_1
node(node0_1).
functor(node0_1, pred).
gram_sempos(node0_1, v).
t_lemma(node0_1, zemrit).
node0_2 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 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