

# 4IZ210

## Sémantické anotování zpráv z CI

LS2007/2008

Jan Nemrava, KIZI VŠE

květen 2008

# Výchozí stav

- Každá skupina (2-5 členů) zpracovává jako semestrální práci analýzu okolí zadané společnosti podle 5 Porterových sil
- Výsledkem je nestrukturovaný DOC dokument, který je určen konkrétní firmě a pro ostatní firmy není zcela relevantní

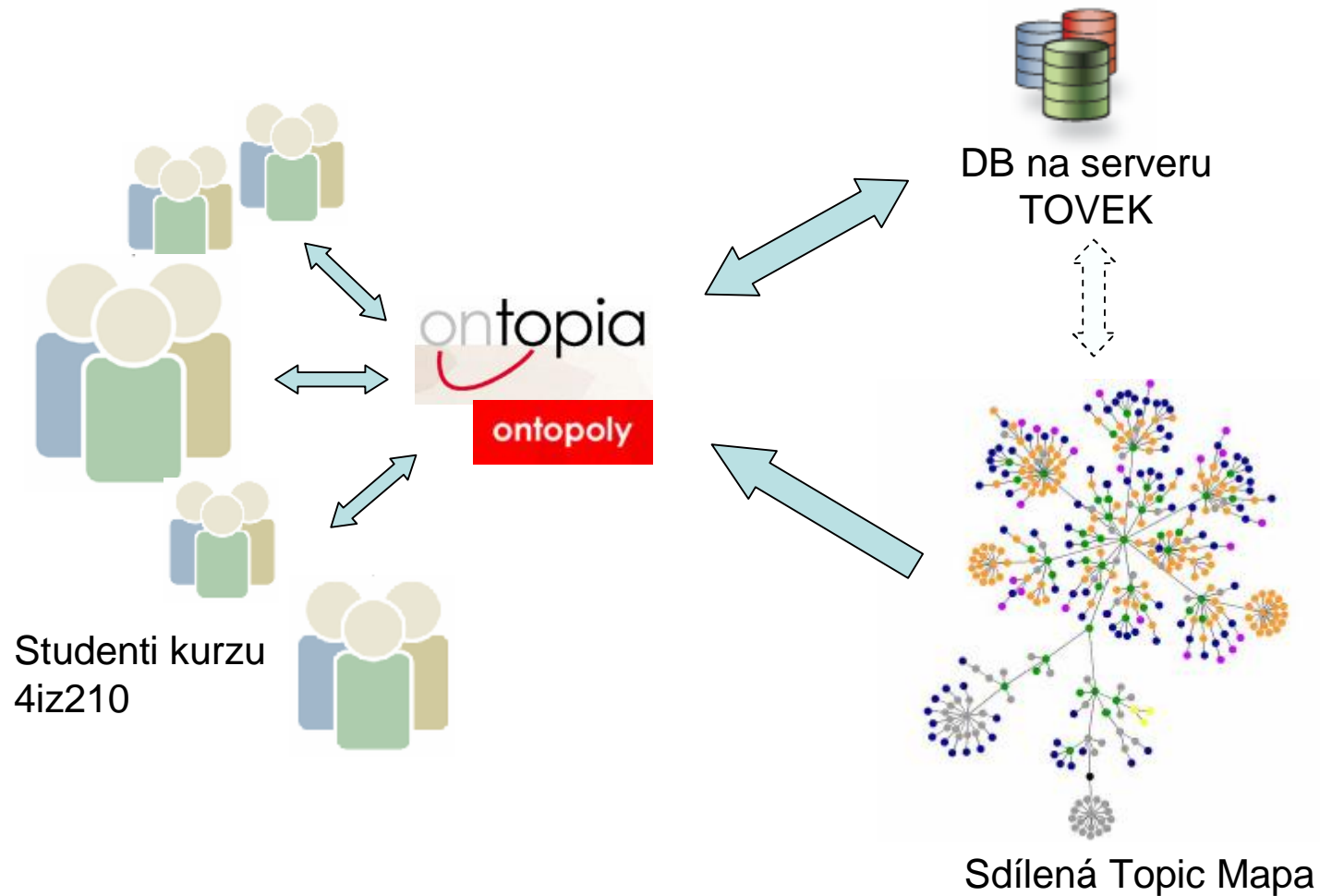
# Problém

- Dá se předpokládat, že u firem ze stejného odvětví (sklářský průmysl a infozdroje v LS2007/2008) je celá řada shodných informací a hlavně zajímavých pro konkurenci
  - Dodavatelé, Odběratelé – pro každou firmu odlišné
  - Substituty, Bariéry vstupu, konkurence – pro celé odvětví víceméně shodná, ale mohou se objevit zajímavé informace
- Prohledávání dokumentů těžko proveditelné
- Ruční procházení časově náročné

# Řešení

- Zaznamenat do dokumentu informaci o čem, jednotlivé části pojednávají a pak podle těchto anotací.
- Je nutné specifikovat strukturu (ontologii) podle které se texty anotují  
(aby jedna skupina neoznačila „odběratele“ a jiná „zákazníci“ text pojednávající o jednom a tom samém)
- Nároky a požadavky každé skupiny jsou jiné, proto je nutné ontologii vytvářet kolaborativně (použitím nástroje OKS)
- Anotace textů pomocí Topic Mapperu

# Sdílená Topic Mapa



# Postup řešení

- Zprávy z Wordu převést do HTML, načíst v Topic Mapperu a pak uložit jako Topic Mapu v XTM
- (tím se zajistí spojení s linkovacím souborem)



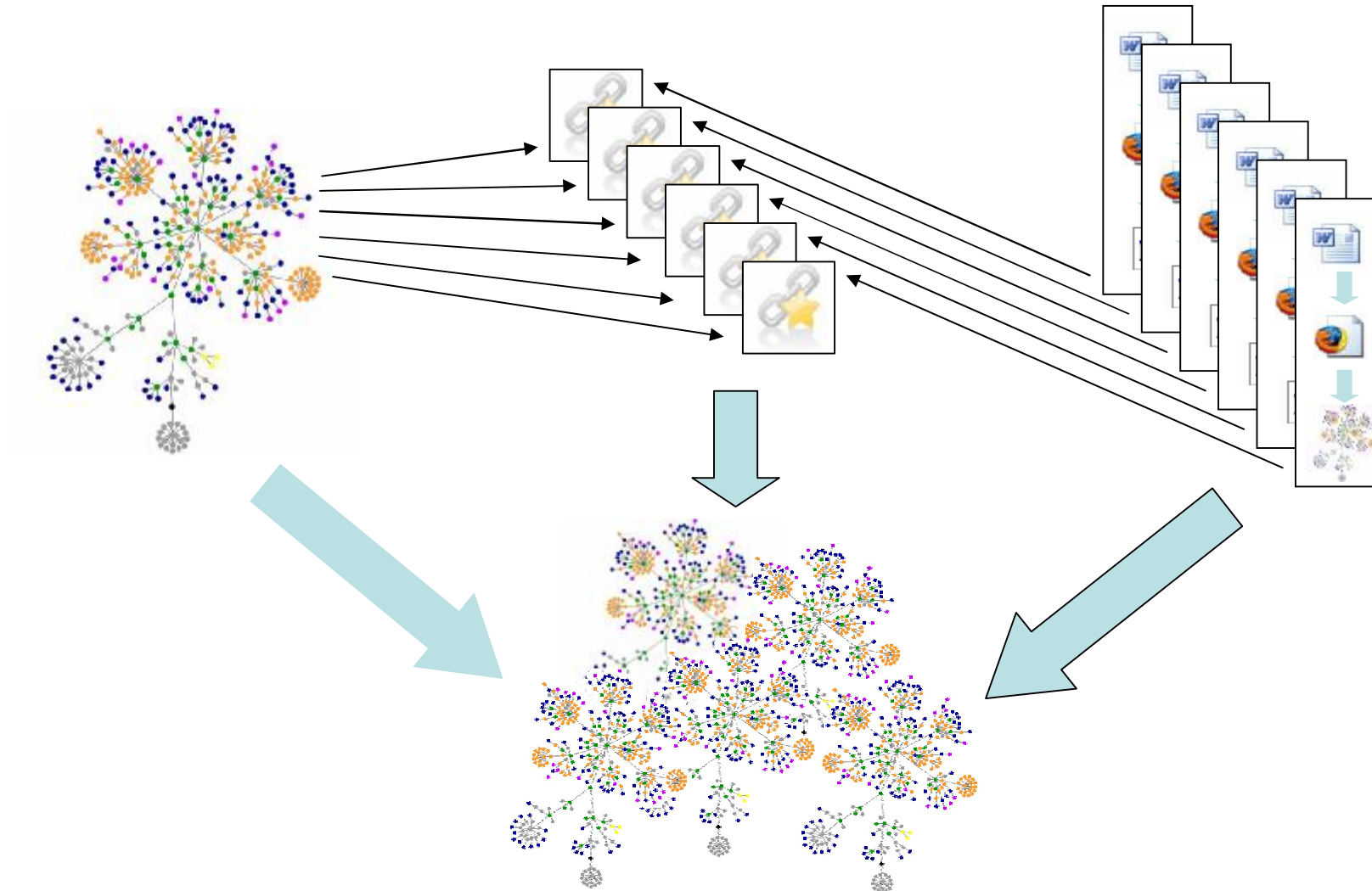
The screenshot shows the Tavek TopicMapper interface. On the left, a colorful topic map is displayed. The main window is divided into several panes:

- Left Pane:** A list of companies and their assets. The first entry is "Computer Press, a.s., IČ 256529" with 21045 assets.
- Right Pane (New Topic Map):** A list of topics including "Advertising campaign", "Anopress", "Appliance Selling", "Assets", "Immaterial assets", "Material", "Assets and Liabilities", and "Association role type".
- Bottom Pane (Topic Properties):** A table showing properties for a selected topic, "Assets".

Assets	
<b>Description</b>	skupina 007 tým č.2
<b>Has cardinality</b>	Exactly one
<b>Has field</b>	Name
<b>InstanceOf</b>	Topic type
<b>Is Tag of</b>	Analýza porterových sil - Aspe...
<b>Subclass of</b>	Assets and Liabilities
<b>Superclass of</b>	Immaterial assets Material

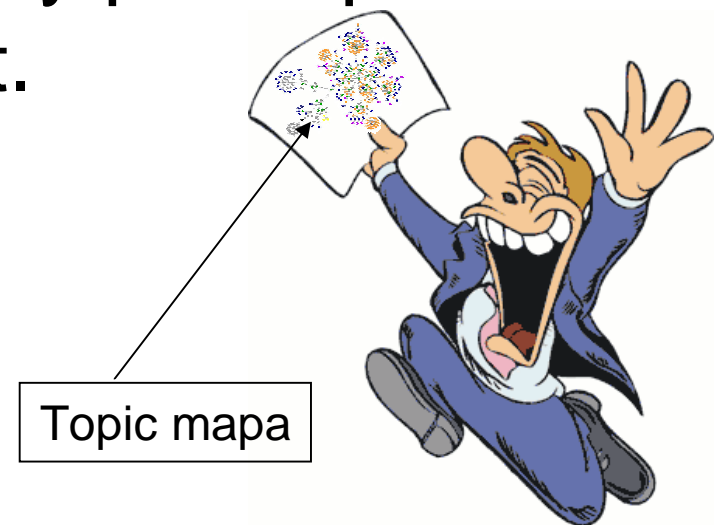
Linkovací soubor spojí Topic Mapu s dokumentem a je výsledkem anotování

# Spojování Topic Map



# Výsledek

- Jedna velká topic mapa obsahující
  - Topic mapu odvětví
  - Všechny texty dokumentů
  - Linkovací soubory
- Možnost procházet všechny práce podle strukturované mapy témat.



# Problémy

V LS 2007/2008  
experimentální nasazení,  
řešení problému „za  
pochodu“

- **Sdílená tvorba topic mapy pomocí Ontopia Knowledge Suite**
  - Při práci s OKS se při nestandardním ukončení operace (stačí stisknutí tlačítka ZPĚT v prohlížeči) vytvoří topic type, který nejde ani editovat ani smazat
  - Jednou se topic mapa sesypala => export do XTM => import
- **Převod hotové topic mapy do TopicMapperu**
  - TopicMapper neumí pracovat s instancemi
  - v XTM souboru není u instanci PSI
  - nutné vytvořit XSLT transformaci, která PSI „ručně přidala“ => problem s mezerami ve slovech a čj.

# Problémy, část 2

- **Anotace topic mapperem**
  - Rozpadlá struktura topic mapy
  - Občas problém při znovunačítání prací
- **Spojování topic map**
  - Každý linkovací soubor a každá topic mapa dokumentu se musí zvlášť pospojovat
  - HTML značky v textu

## Subject Identifiers (1)

- [http://psi.tovek.cz/document1206090562810/docpart\\_1206695489748](http://psi.tovek.cz/document1206090562810/docpart_1206695489748)

## Internal Occurrences (1)

### Text of Document Bit

- o [8] Výpis dostupných informací z obchodního rejstříku (databáze ARES) o firmě Anopress IT, a.s.</p><p class=MsoNormal><a href="http://wwwinfo.mfcr.cz/cgi-bin/ares/darv\_or.cgi?ico=26694484&odp=html&jazyk=cz&xml=true">http://wwwinfo.mfcr.cz ...</a></p><p class=MsoNormal>&nbsp;</p><p class=MsoNormal>[9] Výpis dostupných informací z obchodního rejstříku (databáze ARES) o firmě TOVEK, spol. s.r.o.</p><p class=MsoNormal><a href="http://wwwinfo.mfcr.cz/cgi-bin/ares/darv\_or.cgi?ico=49687981&odp=html&jazyk=cz&xml=true">http://wwwinfo.mfcr.cz ...</a></p><p class=MsoNormal>&nbsp;</p><p class=MsoNormal>[10] Výpis dostupných informací z obchodního rejstříku (databáze ARES) o firmě SUWECO CZ, s.r.o.</p><p class=MsoNormal><a href="http://wwwinfo.mfcr.cz/cgi-bin/ares/darv\_or.cgi?ico=25094769&odp=html&jazyk=cz&xml=true">http://wwwinfo.mfcr.cz...</a></p><p class=MsoNormal></p><p class=MsoNormal>[11] Výpis dostupných informací z obchodního rejstříku (databáze ARES) o firmě BONUS-EBSCO, s.r.o.</p><p class=MsoNormal><a href="http://wwwinfo.mfcr.cz/cgi-bin/ares/darv\_or.cgi?ico=49621823&odp=html&jazyk=cz&xml=true">http://wwwinfo.mfcr.cz/...</a></p></li></ul>

# Závěr a poučení

- Veškeré problémy byly konzultovány s firmou TOVEK
- Většina problémů se vyřešila (víceméně)
- Pro příští semestr máme dostatek materiálů a zkušeností
- Co se nevyřešilo:
  - HTML kód v anotacích
  - Vytváření

**Děkujeme za trpělivost**