Extrakce informací z úplných textů

Jan Žbirka, Tovek Partner

INFORUM 2006

Praha, 23.5.2006

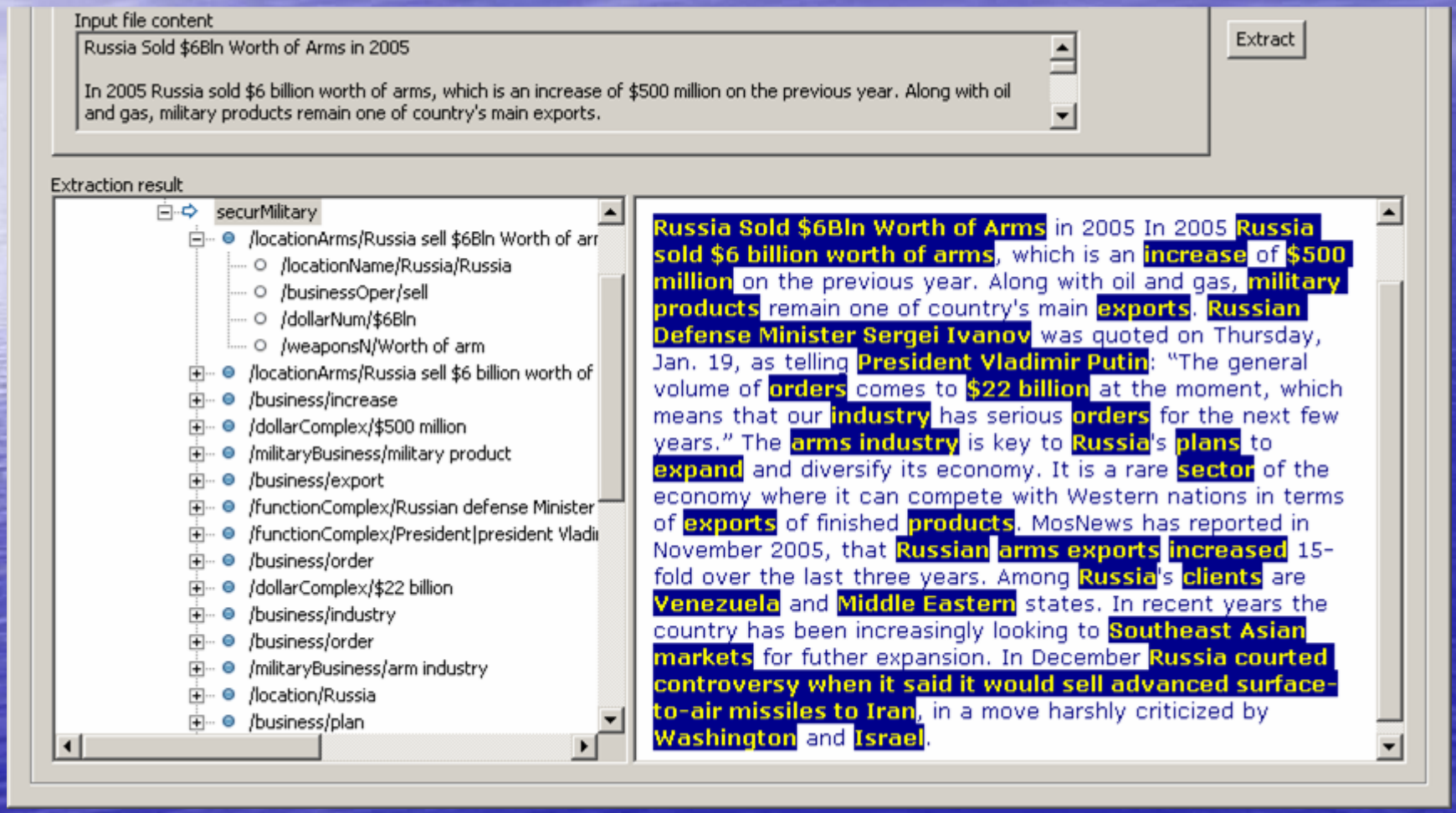
Zpracování úplných textů

- Fulltextové vyhledávání (3. generace)
- Vizuální analýzy
- Extrakce informací

Podstata extrakce informací

- nahrazování určitých slov (výrazů)
- „nálepkami“ „pojmů“
- odkazování na nálepky pojmů ve vyšších vrstvách
- získávání strukturovaných faktů
- za cenu ztráty (nepodstatného?!) textu

Extrakce informací z textů



Architektura systému

- Segmentace a preprocesor
- Filtr
- Parser, Kombinace fragmentů
- Sémantická interpretace, odstranění víceznačností
- Generování šablon

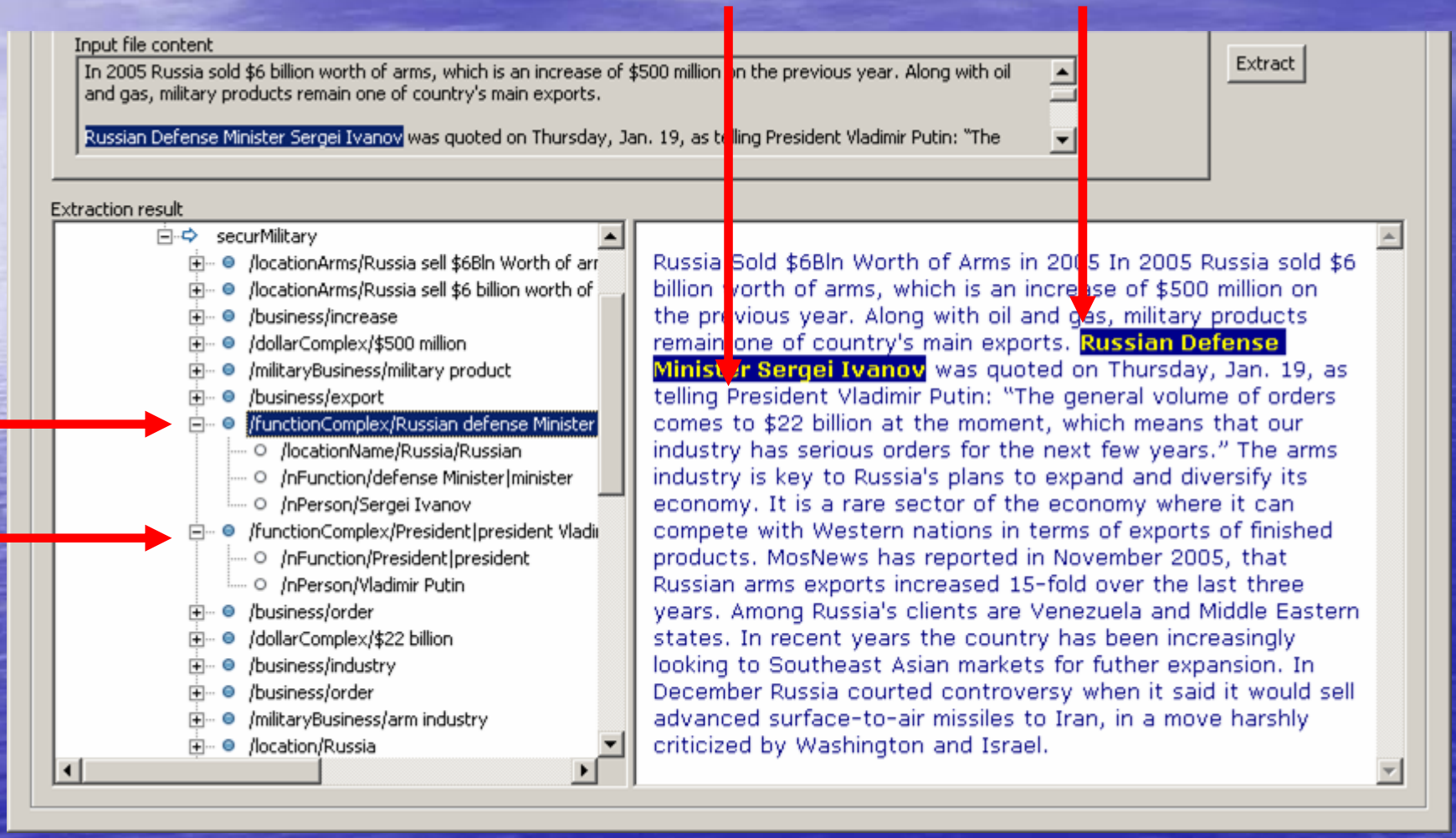
Základní extrakce

- místa
  - oblasti, státy, města, ...
- osoby
  - stát, instituce, organizace, firmy, ...
- čísla
  - peníze, datum, telefonní čísla
- funkce

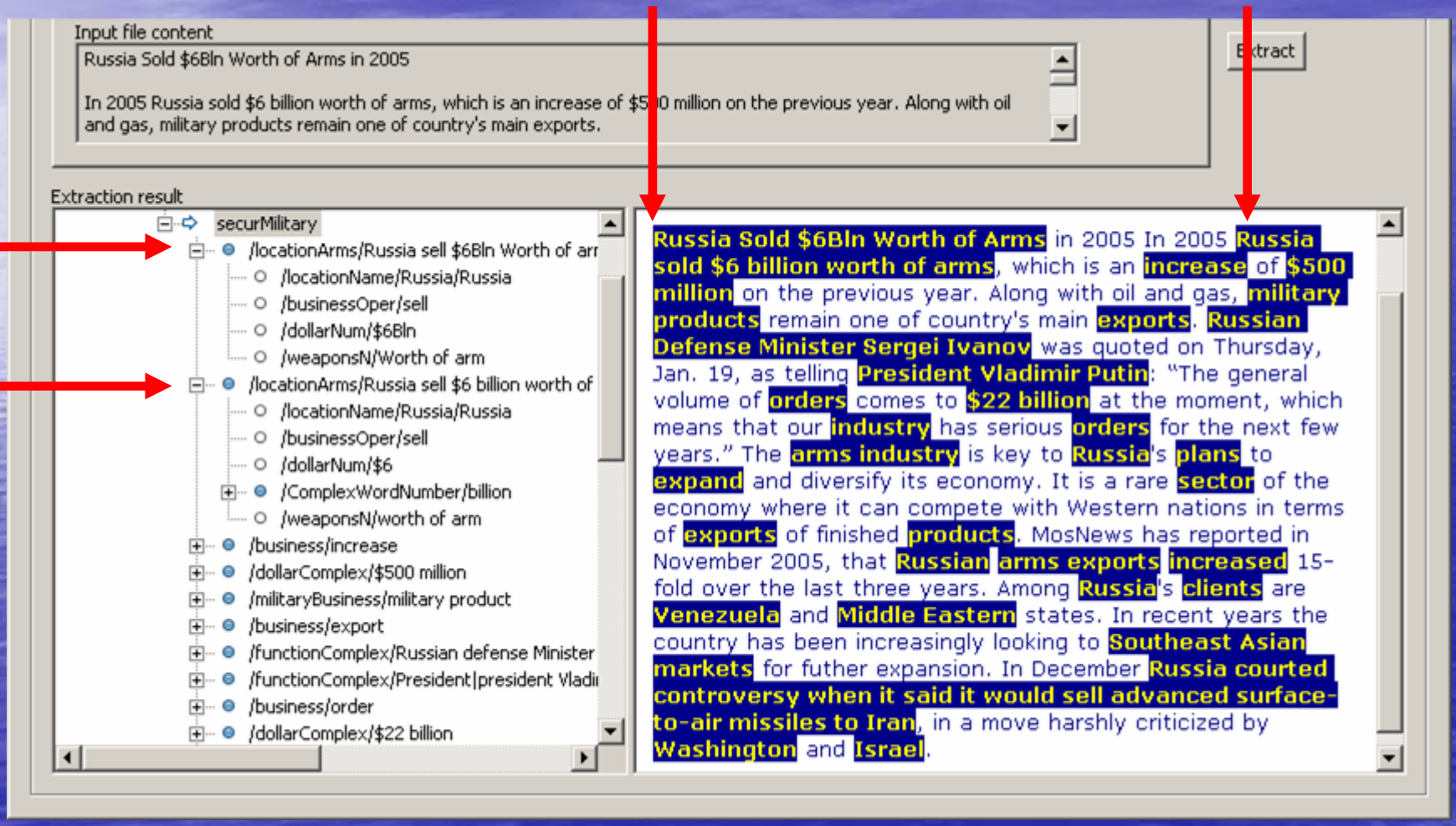
Komplexní extrakce

- obecně: entita – relace – entita
- instituce (firma) – funkce – osoba
- firma – obchod – zboží
- firma – obchod – zboží – firma
- atd.

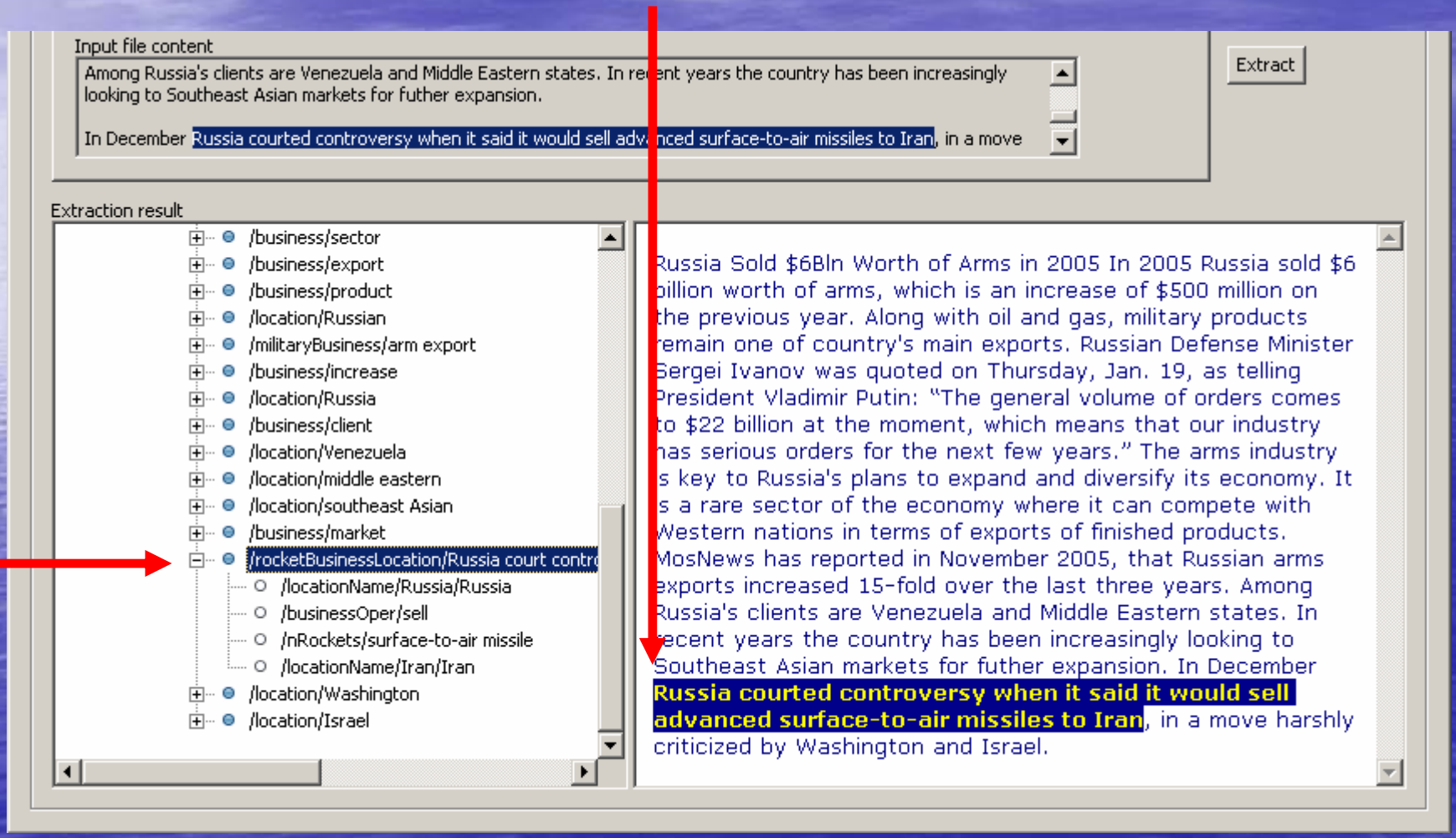
stát – funkce – osoba



stát – obchod – zbraně



stát – obchod – zbraně – stát



Testování extrakcí

- Jazyky
  - Angličtina
  - Čeština
- Tematické oblasti
  - Bezpečnost
  - Competitive Intelligence

Závěry

- Flektivní jazyky obtížněji zpracovatelné nežli jazyky izolační
- Lingvistiku je třeba korigovat
- Odhadnout kdy ukončit rozvoj
  - dalším rozvojem – zhoršení výsledků
- Nasadit skutečně na extrakce
  - ne na kategorizaci, shlukování, atd. kde je fulltext lepší