



# TEXTAN

Bc. Petr Fanta

Duc Tam Hoang, B.Sc.

Bc. Adam Huječek

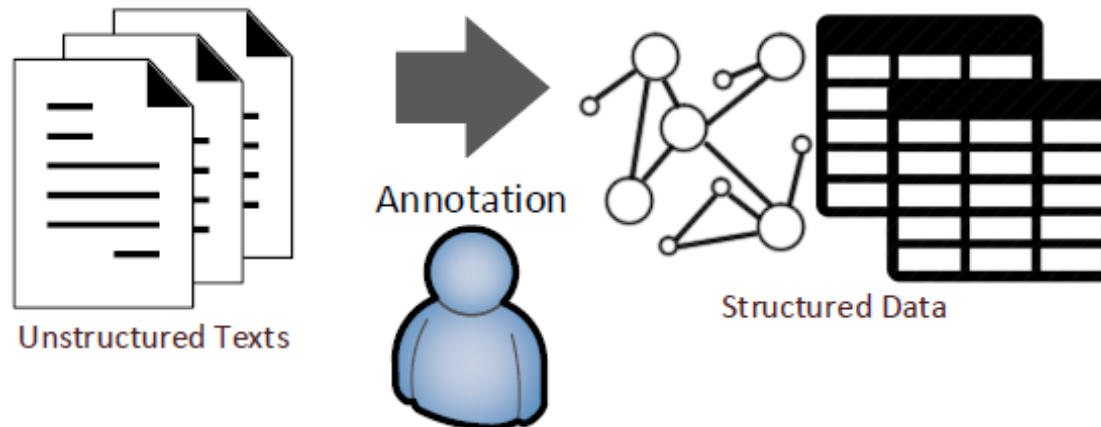
Bc. Václav Perníčka

Bc. Jakub Vlček

Vedoucí projektu: RNDr. Ondřej Bojar, Ph.D.

# Motivace

- Velké množství textů v různých oblastech
- Strukturované informace X nestrukturované dokumenty
- Objekty a vztahy zachycené v dokumentech
- Příklad
  - Policejní zprávy
  - Proces zpracování zpráv v Policii ČR



# Vzorová zpráva

Č.j.: NPC-707-37/2005

14.09.2005

Ú ř e d n í z á z n a m

Dnešního dne byl ve věznici PRAHA 4 PANKRÁC vytěžen obv. Kuka David, nar. 11.6.1976. Předmětem vytěžení bylo získání Kuky ke spolupráci a získání informací k mezinárodnímu obchodu s OPL. Kuka se asi po hodinovém rozhovoru rozhodl spolupracovat s NPC jak na objasnění v jeho trestní věci, tak i na odhalení dalších skupin a osob, které dováží do Afriky kokain, hašiš, marihuanu a vyrábí a vyváží drogu extázi. Jako důvod spolupráce uvedl snahu o snížení trestu, který určitě dostane, a také proto, že mu ve vězeňské nemocnici zjistili, že pravděpodobně trpí diabetem. Schůzka byla omezena časově, z toho důvodu není vytěžení úplné, avšak na dalších návštěvách sdělí další informace.

# Vzorová zpráva - informace

Č.j.: NPC-707-37/2005

14.09.2005

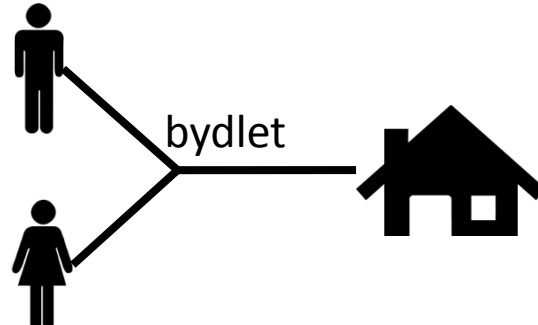
Ú ř e d n í z á z n a m

Dnešního dne byl ve věznici PRAHA 4 PANKRÁC vytěžen obv. Kuka David, nar. 11.6.1976. Předmětem vytěžení bylo získání Kuky ke spolupráci a získání informací k mezinárodnímu obchodu s OPL. Kuka se asi po hodinovém rozhovoru rozhodl spolupracovat s NPC jak na objasnění v jeho trestní věci, tak i na odhalení dalších skupin a osob, které dováží do Afriky kokain, hašiš, marihuanu a vyrábí a vyváží drogu extázi. Jako důvod spolupráce uvedl snahu o snížení trestu, který určitě dostane, a také proto, že mu ve vězeňské nemocnici zjistili, že pravděpodobně trpí diabetem. Schůzka byla omezena časově, z toho důvodu není vytěžení úplné, avšak na dalších návštěvách sdělí další informace.

# Terminologie TextAnu

- „Adam a Eva žili v Ráji.“
- Entita
  - Adam a Eva žili v Ráji.

- Objekt
  - Osoby: Adam  a Eva 
  - Adresa: Ráj 

- Relace

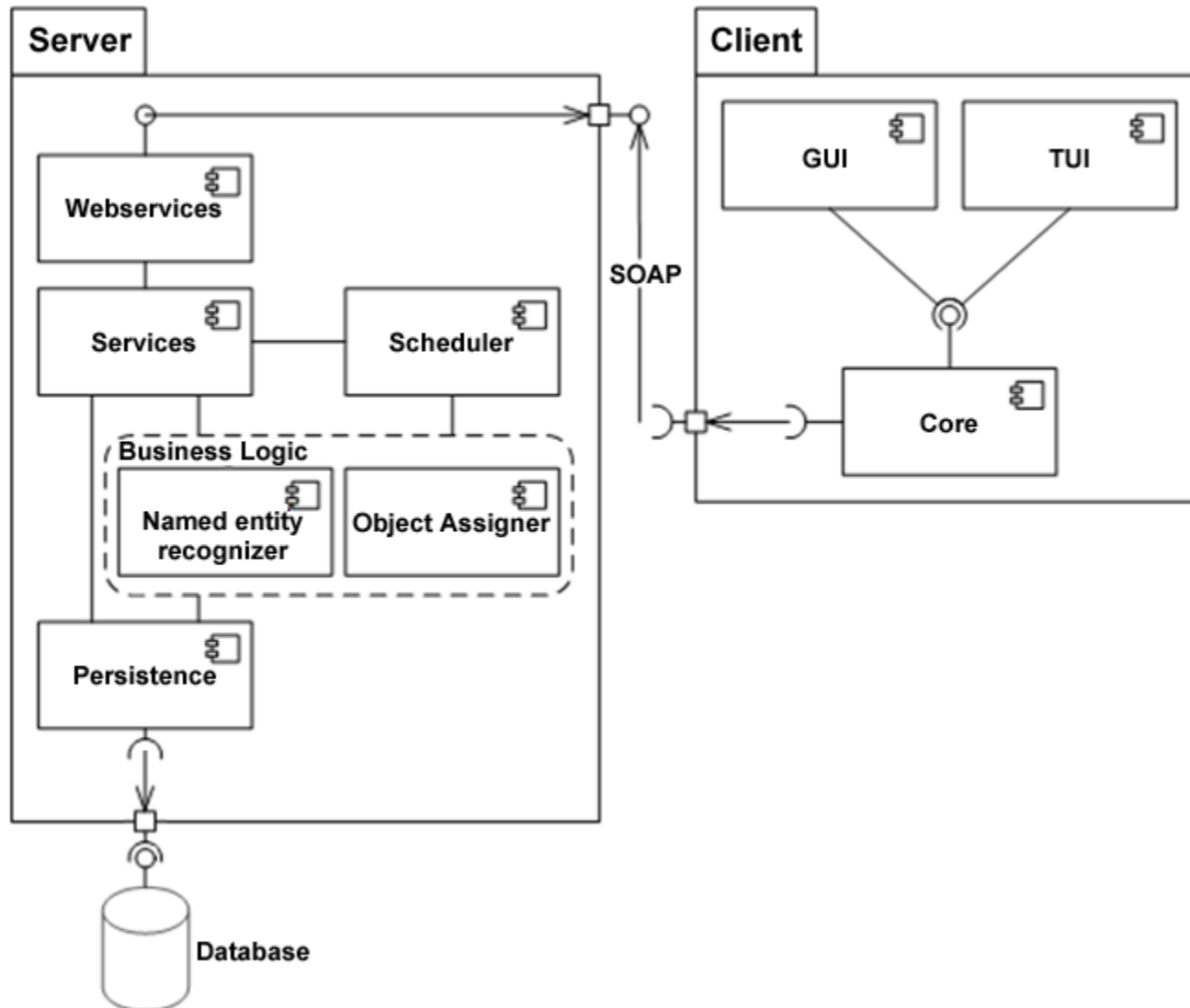
The diagram illustrates the relationship between the entities and the object. It shows two human figures (Adam and Eva) on the left, connected by lines to a house icon on the right. The word 'bydlet' (live) is written above the lines, indicating the relationship.

# Co je to TextAn?

- (Polo)automatický analyzátor textů a databáze textů
- Jazyk – čeština (další jazyky konfigurovatelné)
- Cílová doména
- Zpracování dokumentu
  - Entity (automatické rozpoznání)
  - Objekty (automatické přiřazení)
  - Vztahy (manuální vyznačení)
- Procházení dat
  - Sekvenční
  - Grafy



# Architektura



# Technologie

- Server
  - NameTag & MorphoDiTa
  - WEKA
  - Spring Framework
  - Apache CXF
  - Jetty
  - Hibernate ORM
  - Hibernate Search
  - SLF4J & Logback
- Klient
  - ControlsFX
  - JFXtras
  - JUNG & PretopoLib
  - JCommander



# Požadavky na spuštění

- Server
  - Java 8 & C++11
  - Podporované platformy
    - Windows (32 bit / 64 bit)
    - Linux (32 bit / 64 bit)
- Klient
  - Platformově nezávislý
  - Java 8

# Zpracování dokumentu - Pipeline

- Výběr zdroje zprávy
- Editace textu
- Editace entit
- Editace objektů
- Editace vztahů

# Ukázka



# Rozpoznávání pojmenovaných entit

- NameTag (a MorphoDiTa)
  - 2 iterace
  - features
- Přetrénování modelu rozpoznávače
  - asynchronní
  - konfigurovatelnost

# Přiřazení entit k objektům

- Problémy
  - jednoznačnost
  - neúplnost dat
- Klasifikace i ranking
- Features
- Trénování
  - falešné vztahy

# Projekt v číslech

- Leden 2014 – září 2014
- 1573+ commitů
- 80+ schůzek
- 42204 řádků kódu

# Praktické využití

- Testování na skutečných policejních zprávách
- 28000+ zpráv
- Příprava napojení na existující policejní systémy přes webservice

Děkujeme za pozornost