

# 4IZ440 Reprezentace a zpracování znalostí na WWW

## Seznámení s předmětem

Doc. Ing. Vojtěch Svátek, Dr.

*Letní semestr 2010*

<http://nb.vse.cz/~svatek/rzzw.html>

# Hlavní otázky, které by měl předmět zodpovědět

- Jak je možné využít strukturu WWW ke sdílení a propojování strukturovaných dat a znalostí mezi lidmi a organizacemi?
- Jak takové strukturované znalosti vytvářet?
- Jak je možné v takto propojených datech a znalostech vyhledávat, a automaticky nad nimi odvozovat pomocí logických postupů?
- Jak je možné získávat strukturované znalosti z nestrukturovaných textů, zejména na webu?
- Co z výše uvedeného se v současnosti nebo v blízké budoucnosti uplatní v praxi, a jak?

# Motivační scénáře

- Digitální knihovna, ve které lze pro daný dotaz obdržet nejen dokumenty obsahující zadaná klíčová slova, ale i jiné dokumenty řešící stejnou nebo podobnou problematiku
- Repozitář obrázků nebo videí, ve kterém lze vyhledávat podle různých obsahových hledisek
- Portál integrující informace o osobách, organizacích, místech apod., získané z různých, nepředdefinovaných zdrojů
- Webový portál, který automaticky mění svou strukturu na základě dat, které do něj poskytovatelé dodávají
- Agregátor blogů využívající souvislosti mezi jednotlivými vyslovenými tvrzeními
- Aplikace, která za běhu vyhledává adekvátní webové služby a kombinuje jejich výstupy za účelem vyhovění požadavku uživatele (např. zabezpečení zahraniční cesty se vším všudy – doprava, hotel, pojištění...)
- Některé ze scénářů využití znalostních technologií na webu lze již dnes využívat v rámci **veřejně** přístupných služeb, některé se testují **neveřejně** v rámci projektů pro komerční sféru, a některé jsou zatím jen na úrovni omezených, **čistě výzkumných** prototypů

# Co předmět je a není?

- Nejde o ustálenou disciplínu, pro kterou by existovaly „tradiční“ osnovy a „závazné“ učebnice
- Obsahem výuky není jedna přesně vymezená technologie (jazyk, metodika, software...)

## ALE

- Jde o velmi volné propojení výzkumných iniciativ, softwarových nástrojů, jazykových standardů...
- „Správné odpovědi“ na klíčové otázky v oboru zatím nikdo na 100% nezná
- Vzhledem k tomu je u studentů velmi vítána vlastní iniciativa, zvědavost, zkoumání toho, co se nově objevuje na webu, na konferencích apod.!

# Klíčové pojmy

- Sémantický web
- Mikroformáty
- RDF (Resource Description Framework)
  - Úložiště, SPARQL, koncové body
  - Propojená data (Linked Data)
  - RDFa
- Topic Maps (mapy témat)
- Ontologie
  - OWL
  - Deskripční logika
  - Ontologické návrhové vzory
- Sémantické anotování
  - Extrakce informací

# Klíčové pojmy (zjednodušeně)

- **Sémantický web:** hlavní náplň předmětu – jde o využití webové infrastruktury pro zpřístupnění „strojům srozumitelných“ informací (sémantika = význam, smysl)
  - Termín zaveden kolem r.2000 pro oblast vzniklou spojením
    - nástrojů a standardů sítě WWW
    - technologie reprezentace a zpracování znalostí, zejména modelování znalostí (ontologického inženýrství) a formální logiky (deskripční, event. Hornova logiky)
  - Později se zapojily i další komunity: zpracování přirozeného jazyka, text/web mining, databáze, filosofie, zpracování neurčitosti, ...

# Klíčové pojmy (zjednodušeně)

- **Mikroformáty**: prostředek pro jednoduché zachycení sémantiky v HTML
- **RDF (Resource Description Framework)**: hlavní jazyk pro reprezentaci dat v sémantickém webu, standard W3C; spíše „databázově“ orientovaný
  - **Úložiště RDF (RDF store)**: „databáze“ pro data ve struktuře RDF
  - **SPARQL**: dotazovací jazyk analogický SQL
  - **Koncové body (SPARQL endpoints)**: webová rozhraní pro dotazování do (centralizovaných nebo distribuovaných) úložišť
  - **Linked Data**: soubor principů pro propojování dat v RDF, pocházejících z různých zdrojů, a množina takto již přizpůsobených zdrojů
  - **RDFa**: formát pro vkládání dat RDF do stránek v HTML, „vylepšení“ mikroformátů
- **Topic Maps (TM)**: alternativní jazyk pro reprezentaci sémantických dat, standard ISO; spíše „dokumentově“ orientovaný, omezené možnosti odvozování

# Klíčové pojmy (zjednodušeně)

- **Ontologie:** formální konceptuální model určité problémové oblasti; pojmy a vztahy z modelu slouží k dodání sémantiky datům jak pro RDF, tak i pro TM
  - **OWL:** jazyk pro reprezentaci ontologií, zpravidla používaný ve spojení s RDF
  - **Deskripční logika:** logický kalkul umožňující odvozování nad výroky v RDF/OWL
  - **Ontologické návrhové vzory:** opakovaně použitelné fragmenty ontologií, zpravidla opatřené vysvětlujícím komentářem, příklady apod.
- **Sémantické anotování:** přiřazování významu (tj. pojmů z ontologie) částem textu – může být ruční nebo automatické
  - **Extrakce informací:** automatické sémantické anotování textu, směřující ke zpřístupnění nalezených informací ve strukturované podobě

# Tématické bloky

- Sémantický web – přehled, ukázky aplikací
- Reprezentace znalostí pro sém.web
  - RDF/S, OWL, pravidla...
- Nástroje pro RDF
  - Úložiště, SPARQL, koncové body
- Sémantika v HTML
  - Mikroformáty, RDFa, GRDDL...
- Úlohy v sém.webu a nástroje na jejich řešení - přehled
- Linked Data
- Sémantické anotování
  - ruční a automatické
- Deskripční logika
  - formalismus, tablové odvozování
- Ontologické inženýrství
  - návrhové vzory
- Topic Maps
- Rozšiřující témata
  - sém.webové služby
  - sémantika multimédií
  - mapování ontologií
  - ...

# Co se od Vás očekává?

- Získat rámcový přehled o oblasti
  - Absolutní minimum: [slidy](#) k přednáškám, budou postupně vystavovány na webu (od 2. týdne)
  - Základní tištěnou literaturou jsou [knihy](#)
    - *A Semantic Web Primer* od G. Antoniou a F. van Harmelena
    - *Semantic Web for the Working Ontologist* od D. Allemanga a J. Hendlera
    - případně *XML Topic Maps* od J. Parka a S. Huntinga
    - dostupné ve studovně a ve skladu (není nutné znát všechny podrobnosti)
  - Definice používaných [jazyků](#) jsou dostupné na WWW
  - V případě zájmu mnoho další tištěné literatury u vyučujícího, články na WWW a v digitálních knihovnách
- Bude ověřeno dvěma písemnými [testy](#) (40% hodnocení), kombinujícími požadavky na konkrétní znalosti se schopností samostatně uvažovat a formulovat teze
- Testy jsou předběžně v 7. a 12. týdnu semestru

# Ale také a především...

- Zpracovat samostatně nebo (v případě rozsáhlejšího projektu) ve dvojici **semestrální projekt** (40% hodnocení)
- Některé možné varianty projektu:
  - vývoj nové nebo adaptace existující aplikace pracující se sémantikou dat
  - rozsáhlé testy hotové volně šířené aplikace tohoto typu
  - rozsáhlý a pečlivě zpracovaný konceptuální model určité oblasti (OWL nebo Topic Maps)
  - nový logický nebo obsahový návrhový vzor pro ontologie
  - přehledová komparativní studie určitého tématu zpracovaná na základě nejnovější literatury
- Kromě poslední varianty se požaduje i textový dokument rozebírající postup práce, provedená rozhodnutí při realizaci projektu atp.
- Projekt je potřeba odevzdat do 25.5., pracovní verze s neformální prezentací ve 13. týdnu výuky (11.5.)
- Zpracovat ústní a písemný **minireferát** ze zvoleného odborného článku a vytvořit k němu **formální relační anotaci** (20% hodnocení)
- V průběhu semestru (prezentace 11.-12. týden)

# A k čemu Vám to dále může být?

- Pro studenty bezprostředně směřující do praxe
  - Nový pohled na techniky, které se v praxi běžně využívají (integrace datových zdrojů, textová a webová analytika, architektury orientované na služby, metamodelování IS, groupware, ...)
  - Přípravenost na novinky, které do praxe proniknou v příštích 5-10 letech
- Pro studenty se zájmem o prozkoumávání neprozkoumaného
  - Přehled významné části výzkumných témat, která jsou na KIZI řešena a mohou být předmětem doktorských disertací, ale i diplomových prací

# Pro hlubší zájemce

- Neformální výzkumný seminář KIZI
  - resp. pracovní skupiny KEG (Knowledge Engineering Group)
- Téměř každý čtvrtek v semestru
  - od 10.30 do cca 12.00
- Viz <http://keg.vse.cz/seminars>

