

Lightweight Semantics and the Web



Marián Šimko

PeWe

28.9. 2011

Sémantika

- = význam
- Web so sémantikou (the Semantic Web)

"man-made woven web of data" that facilitates machines to understand the semantics, or meaning, of information on the World Wide Web
- načo?
 - usudzovanie, odvodzovanie
 - inteligentné vyhľadávanie, odporúčanie

Sémantika

- = význam
- Web so sémantikou (the Semantic Web)
"man-made woven web of data" that facilitates **machines to understand** the semantics, or meaning, of information on the World Wide Web
- načo?
 - usudzovanie, odvodzovanie
 - inteligentné vyhľadávanie, odporúčanie

Príklad

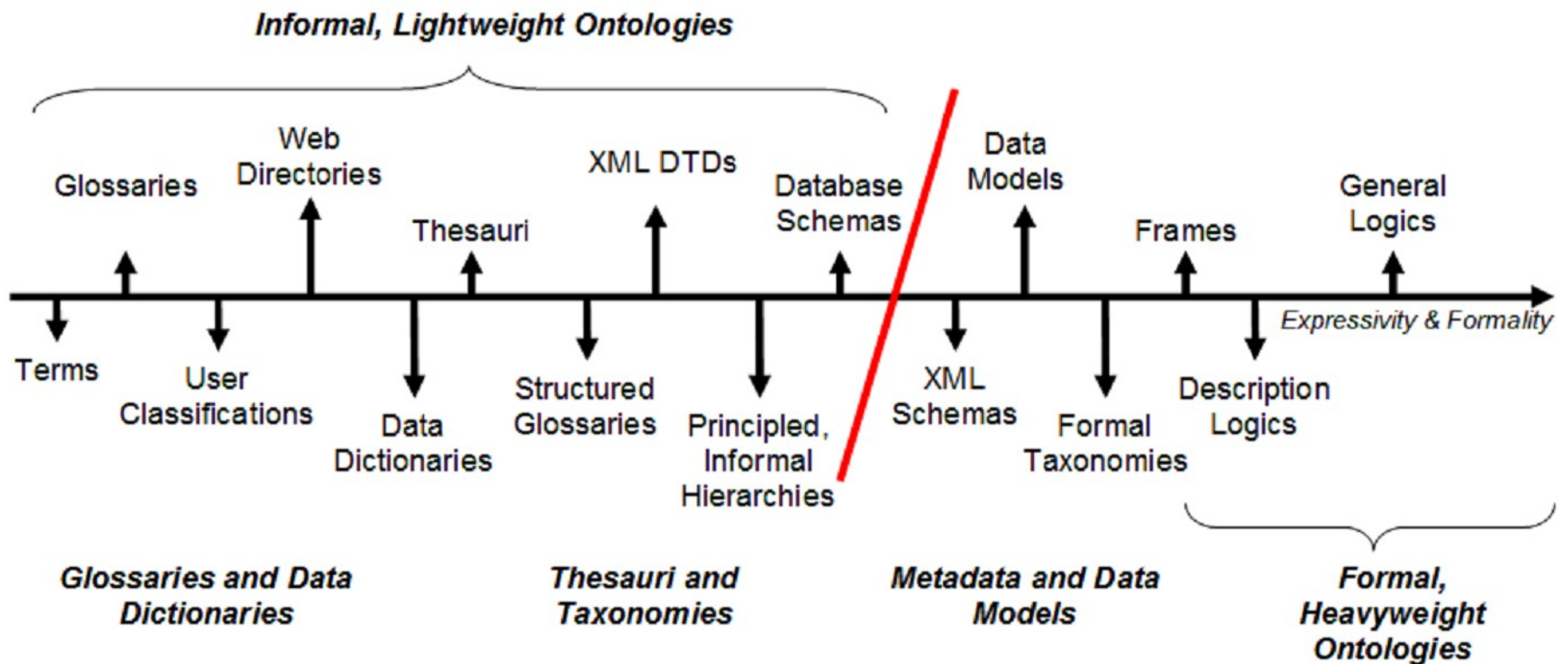
Kedy je skúška z
Procedurálneho programovania?

Web

- Web so sémantikou je riedky a nekompletný (Sabou et al., 2007)
- Nízke percento pokrytia ontológiami (Fernandez, et al., 2008)
- Náročnosť manuálneho zostrojenia ontológie
- Dynamicky meniace sa prostredie

Ľahká sémantika

- jednoduchšie poňatie konceptov
- vybrané typu vzťahov
- typicky už nič viac :)



(Wong et al., 2011)

Ľahká sémantika

- Komparatívna výhoda:

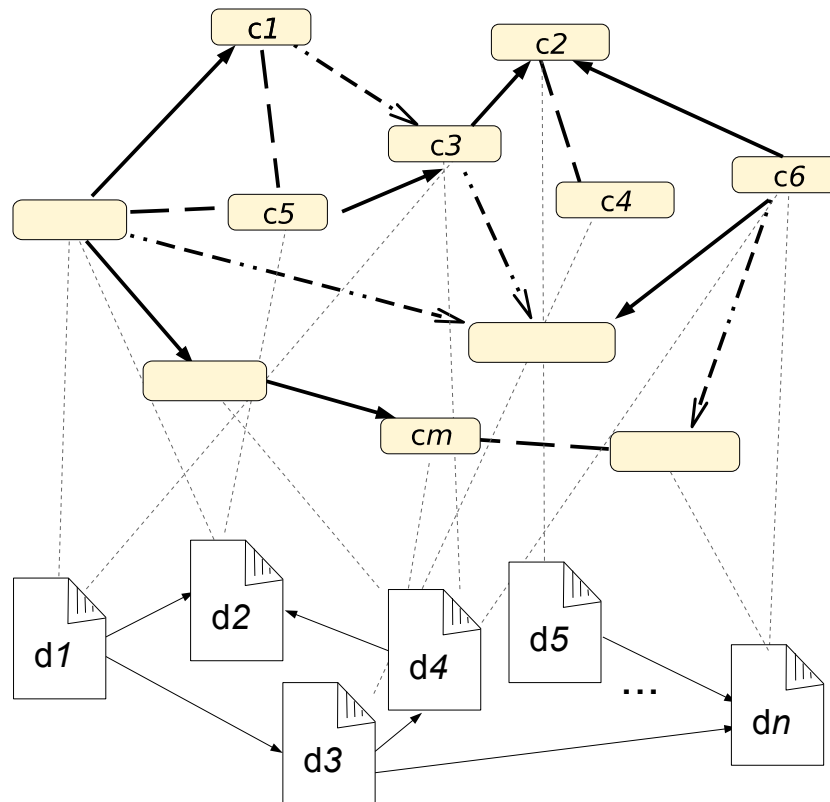
Možnosť automatizovaného získania

- Zdroje:
 - Obsah webu
 - Štruktúra webu
 - Používanie webu

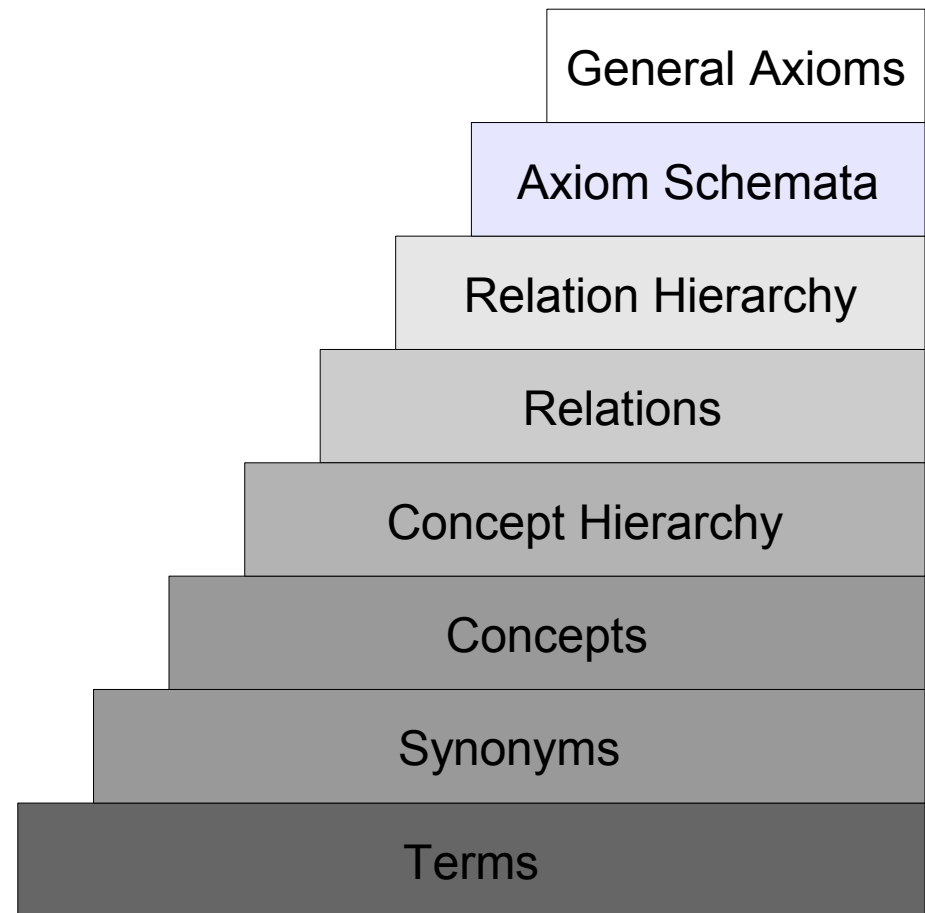
Ľahká sémantika

metadáta

dokumenty



Získanie „heavyweight“ ontológie



(Cimiano, 2006)

Získanie „heavyweight“ ontológie

$\forall x(\text{country}(x) \rightarrow \exists y \text{ capital_of}(y,x) \wedge \forall z(\text{capital_of}(z,x) \rightarrow y=z))$ General Axioms

$\text{disjoint}(\text{river}, \text{mountain})$ Axiom Schemata

$\text{capital_of} \leq_R \text{located_in}$ Relation Hierarchy

$\text{flow_through}(\text{dom:river}, \text{range:GE})$ Relations

$\text{capital} \leq_C \text{city}, \text{city} \leq_C \text{GE}$ Concept Hierarchy

$c := \text{country} := \langle i(c), \|c\|, \text{Ref}_C(c) \rangle$ Concepts

$\{\text{country}, \text{nation}\}$ Synonyms

$\text{river}, \text{country}, \text{nation}, \text{city}, \text{capital}, \dots$ Terms

Získanie „heavyweight“ ontológie

$\forall x(\text{country}(x) \rightarrow \exists y \text{ capital_of}(y,x) \wedge \forall z(\text{capital_of}(z,x) \rightarrow y=z))$

General Axioms

N/A

$\text{disjoint}(\text{river}, \text{mountain})$

Axiom Schemata

$\text{capital_of} \leq_R \text{located_in}$

Relation Hierarchy

10 - 20%

$\text{flow_through}(\text{dom:river}, \text{range:GE})$

Relations

$\text{capital} \leq_C \text{city}, \text{city} \leq_C \text{GE}$

Concept Hierarchy

40 - 50%

$c := \text{country} := \langle i(c), ||c||, \text{Ref}_C(c) \rangle$

Concepts

{country, nation}

Synonyms

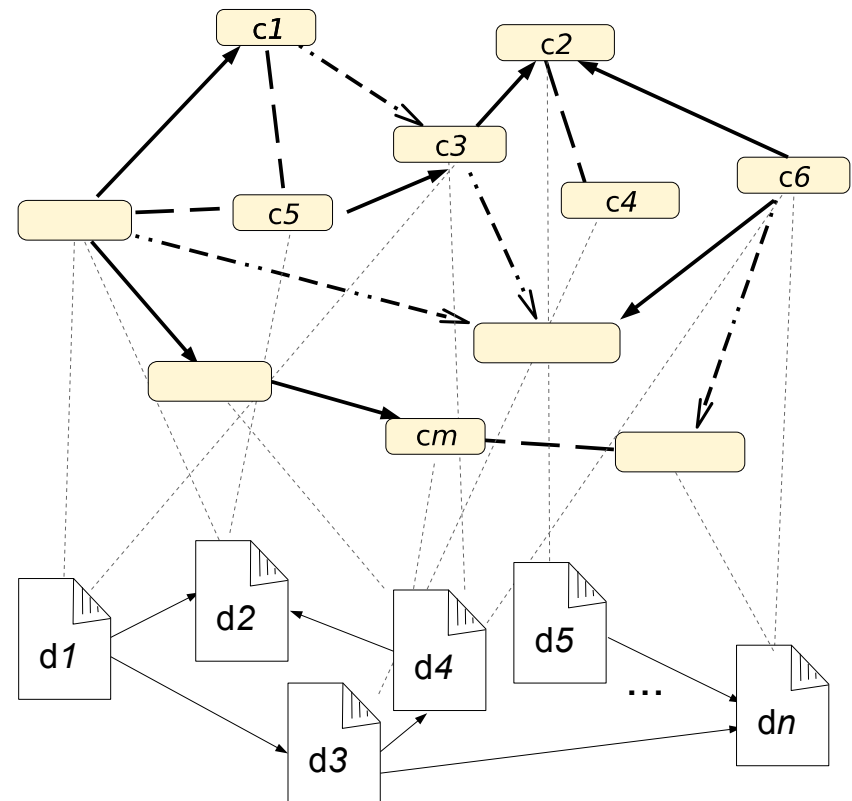
70 - 80%

river, country, nation, city, capital,...

Terms

Získanie „lightweight“ ontológie

- identifikácia dôležitých doménových pojmov
- naviazanie pojmov na obsah
- objavovanie vzťahov medzi pojmami



Predspracovanie obsahu

- tokenizácia
- morfológická anotácia
- lematizácia
- identifikácia viet
- rozvitie viet

Identifikácia doménových pojmov

- ATR, NER
- Štatistické spracovanie textu:
 - váhovanie slov (termov)
 - tf-idf =
term frequency – inverse document frequency
- Spracovanie používateľmi vytvorených dát
 - tagy, anotácie

Identifikácia doménových pojmov

- ATR, NER

Štatistické spracovanie textu

<http://peweproxy.fiit.stuba.sk/metall/>

tf-idf –

term frequency – inverse document frequency

- Spracovanie používateľmi vytvorených dát
 - tagy, anotácie

Objavovanie vzťahov

- Podobnosť
 - similarity vs. relatedness
 - štatistická analýza, grafová analýza
 - spoločné výskyty slov
 - co-occurrence, collocation
 - šírenie aktivácie, centralita uzlov
 - lingvistická analýza
 - porovnanie vektorovej reprezentácie

Objavovanie vzťahov

- Hierarchia

- is-a človek *is-a* cicavec
- lexikálno-syntaktické vzory
- hypotéza rozdelenia (angl. distributional hypothesis)
- analýza spoločného výskytu

Objavovanie vzťahov

- Nehierarchické vzťahy
 - syntaktické závislosti (analýza slovies)
 - analýza spoločného výskytu
 - asociačné pravidlá

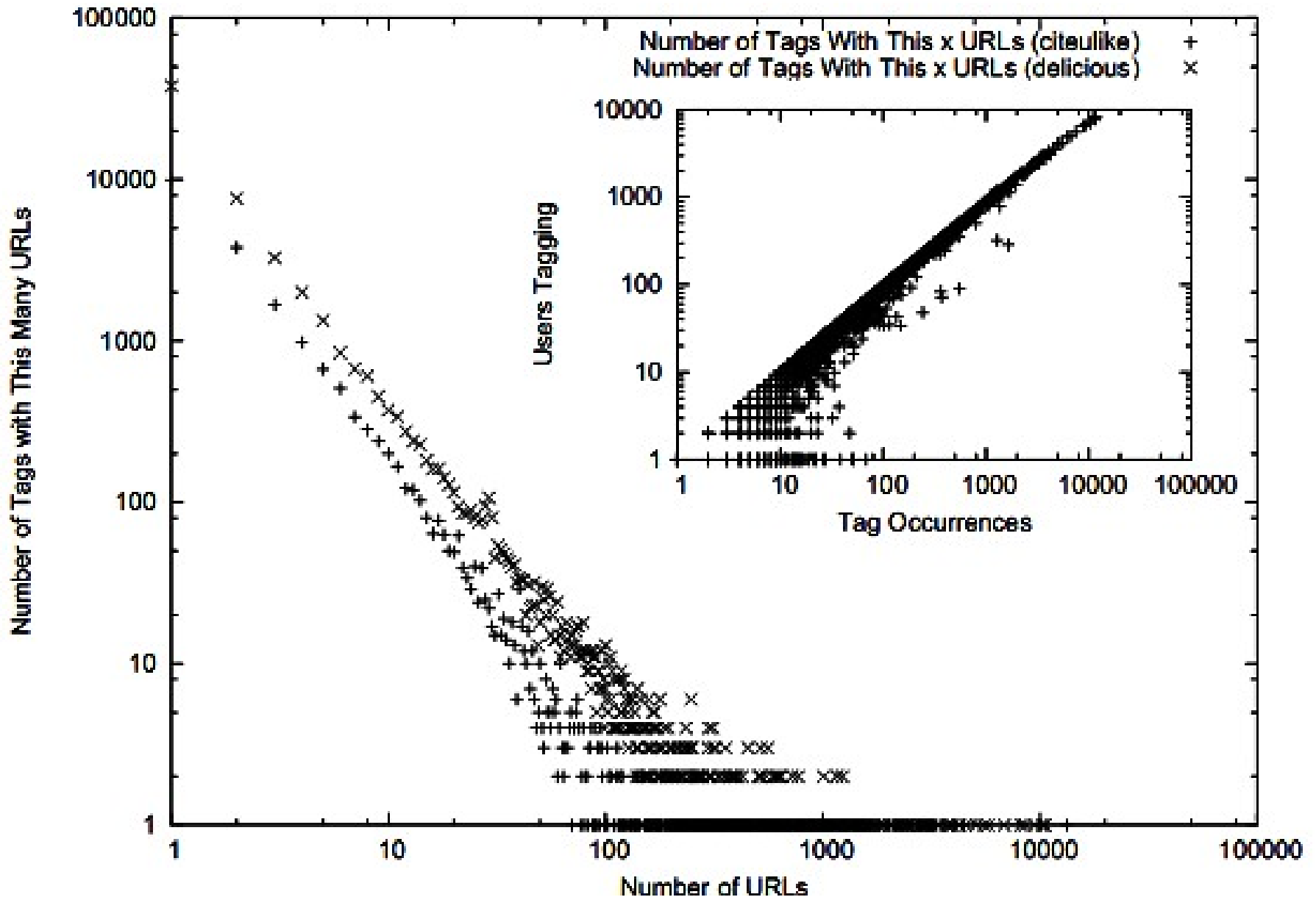
Webové anotácie

- Sociálne tagovanie
- „Znalosti“ z pohľadu používateľa
- Folksonómia – odvodenie konceptualizácie
- Vlastnosti:
 - 50% tagov nie je obsiahnutých v textovom obsahu
 - nízke pokrytie webu
 - relevantný podiel len v určitých doménach

Webové anotácie

- Typy tagov:
 - Opis zdroja
 - Typ zdroja
 - Autor/vlastník zdroja
 - Kvantifikátor
 - Osobná potreba, organizácia úloh
- Implicitné anotácie

(Heymann, Garcia-Molina, 2005)



Zdroje

- Cimiano, P. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag, ISBN: 978-0-387-30632-2. 347p (2006)
- Wong, W., Liu, W., Bennamoun, M. 2011. *Ontology learning from text: A look back and in the future*. *ACM Computing Surveys X*, (2011), 36p.