Linked Data v praxi

Doc. Ing. Vojtěch Svátek, Dr.

Zimní semestr 2011

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

Struktura přednášky

- Encyklopedická LD
- LD pro vyhledávače
- LD pro sociální sítě
- LD pro elektronické obchodování
- LD pro knihovny a publikační zdroje
- LD pro média
- *LD pro veřejnou správu*
- *LD pro oblast biomedicíny*

Linked Data dnes

- Mnoho stovek propojených datasetů
  - nejvýznamnější viz http://richard.cyganiak.de/2007/10/lod/lod-datasets_2010-09-22_colored.html
- Různé metody (open) přístupu
  - URI-based (dereferencování)
  - RDFa (v HTML – následně destilace)
  - Endpoints (SPARQL) – viz http://labs.mondeca.com/sparqlEndpointsStatus/
  - RDF dumps
- Různá míra využívání v praxi
- Konsensuální vs. proprietární slovníky

Encyklopedické/všeobecné

- DBpedia http://dbpedia.org
  - facetový prohlížeč http://dbpedia.neofonie.de/browse/
- Freebase http://www.freebase.com/
  - aktuálně pod Google
  - crowdsourcing
- Uberblic http://platform.uberblic.org/
  - proxy pro další zdroje, deduplikace
- Yago http://mpii.de/yago
  - důraz na pokrytí časoprostorových atributů
- Wordnet – RDF verze na VUA
  - http://semanticweb.cs.vu.nl/lod/wn30/

Vyhledávače

- Zpracovávají mikroformáty, mikrodata a RDFa; výsledkem je zejména
  - Zobrazení strukturovaných informací ve výsledcích vyhledávání a v mapách
  - Upravený ranking
- Google
  - RDF jako alternativa k mikrodatům a mikroformátům
  - vlastní ontologie „pro všechno“: http://www.data-vocabulary.org/
  - z jiných RDF slovníků explicitně podporuje GoodRelations pro nabídky produktů a služeb
- Yahoo
  - obdobné, ale podpora více RDF slovníků (pův. SearchMonkey)
  - http://developer.yahoo.com/searchmonkey/smguide/profile_vocab.html
- Aktuálně aliance Google-Yahoo-Bing
  - http://schema.org – založeno na mikrodatech

Sociální sítě

- Facebook – protokol OpenGraph
  - http://developers.facebook.com/docs/opengraph/; http://ogp.me/
  - metadata v RDF umožní zapojení libovolné stránky do sítě Facebooku, když na ní někdo klikne tlačítko Like
  - povinné vlastnosti: "og:title", "og:type", "og:image", "og:url“
  - dále např. sitename, description; lokalita (geo, ulice), kontaktní info (email, phone); produkty - UPC, ISBN
- Zdroje v LD: Revyu, OHLOH

E-commerce

- Těžištěm jsou ontologie a nástroje vyvinuté na UBW týmem prof. Heppa
  - http://www.heppnetz.de/
- GoodRelations: generická ontologie „nabízení produktů a služeb“
- Rozsáhlé ontologie typů produktů
  - eClassOWL – podle standardu eCl@ss, 60 tisíc typů
  - The Product Ontology – z Wikipedie, 300 tisíc typů
- „Vertikální“ ontologie pro parametry komodit
  - vozidla, lístky, stavebnictví…

E-commerce

- Aplikace
  - Vyhledávání (aktuálně největší motivace)
  - Doporučování (gr:offers vs. gr:seeks), porovnávání nabídek
  - Předvýběr obchodních partnerů
- Tvorba RDFa
  - ruční anotační nástroje
  - pluginy do e-shopových nástrojů
  - middleware nad webovými API (Amazon, eBay…)
- Alternativy specifikace produktu:
  - jen instance třídy „produkt“, typ uveden textově v komentáři
  - odkaz na produktovou ontologii
  - odkaz na DBpedii (ale URI z DBpedie nemají sémantiku třídy…)
  - převod proprietární hierarchie na „pseudo-ontologii“

E-commerce

- http://productdb.org/
  - endpoint pro data o produktech
  - využívá GoodRelations, FOAF a OpenVocab (kolekce ad hoc entit bez dedikovaného slovníku)
- http://linkedopencommerce.com/sparql
  - endpoint pro data o nabídkách
  - využívá zejména GoodRelations

Knihovny

- Jako ontologie se standardně používá SKOS http://www.w3.org/TR/skos-reference/
  - běžné tezaurové asociace (broader, narrower, related, …), lexikální informace (prefLabel…), metainformace
  - těsnost shody (broadMatch, closeMatch, exactMatch)
- Library of Congress Subject Headings
- Polytematický strukturovaný heslář (PSH)
  - NTK Praha (převod do RDF: J. Mynarz)
  - propojení na LCSH a DBpedii

Publikace

- Řada světových bibliografických databází
  - DBLP http://dblp.l3s.de/d2r/
  - Citeseer, ePrints, ACM, IEEE, …
  - Semantic Web Dog Food (nejen publikace…)
  - PCVSE – ve vývoji (vč. propojení na DBLP)

Média

- BBC
  - /programmes, /music, /wildlifefinder
    - Využívají MusicBrainz, Wikipedii; model “web jako CMS”
  - Vlastní ontologie, např. http://www.bbc.co.uk/ontologies/wildlife/2010-02-22.shtml
  - Endpointy hostované Talis a OpenLink
- Guardian
  - prozatím: ContentAPI umožňuje vyhledávat podle externích identifikátorů z MusicBrainz
- Úspěšnost LD pro média závisí na věcné oblasti
  - potřeba předem známých identifikátorů
  - OK pro sport, přírodu, hudbu
  - obtížnější pro běžné zprávy (objevují se nepředvídané entity)

Veřejná správa, biomedicína

- Někdy příště, zasluhuje samostatně…