

ETL pro linked data

prof. Ing. Vojtěch Svátek, Dr.

4IZ440, Zimní semestr 2018

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

ETL a linked data

- ETL = „extract – transform – load“
 - pojem zavedený zejména v kontextu podnikových datových skladů a BI
- V kontextu propojených (otevřených) dat:
 - „extract“ může zahrnovat extrakci z ne-RDF i RDF zdrojů
 - „transform“ zahrnuje zejména čištění a propojování
 - „load“ má zpravidla charakter publikování na sémantickém webu spíše než „natažení“ do centralizovaného datového skladu

Odkud lze extrahovat data

- RDF data
 - soubory obsahující celé grafy (dumpy)
 - SPARQL endpointy
 - Různá jiná API, např. pro „triple pattern fragments“ (viz <http://linkeddatafragments.org/>)
 - stránky s RDFa (<http://www.w3.org/TR/xhtml-rdfa-primer/>)
- Tabulková data
 - RDBMS, CSV, HTML, XLS...
- Hierarchická data
 - XML...
- Síťová ne-RDF data
 - většinou ale uložena v tabulkách
- Volný text

Extrakce z relačních DB

- Mapování z RDB do RDF se definuje
 - pomocí jazyka R2RML
 - W3C doporučení ze září 2012, <http://www.w3.org/TR/r2rml/>
 - Umožňuje definovat mapovací pravidla, optimalizující strukturu dat pro využití jako LD
 - Podporuje různé způsoby vystavení RDB dat jako LD: endpoint, export dumpu, negociace obsahu
 - formou tzv. Direct Mapping
 - Struktura RDF doslova kopírující strukturu RDB
 - <http://www.w3.org/TR/rdb-direct-mapping/>
- Prototypové implementace: Oracle, PostgreSQL, MySQL...
 - viz <http://www.w3.org/TR/rdb2rdf-implementations/>

Extrakce z XML a CSV

- Pro XML jde zpravidla o XSLT šablony, někdy s nadstavbami
 - Např. Valiant, <https://github.com/bertvannuffelen/valiant>
- Pro CSV – řada nástrojů, např.
 - TARQL, <https://github.com/cygri/tarql>
 - Nadstavby obecného extraktoru OpenRefine:
 - LODRefine, <https://github.com/sparkica/LODRefine>
 - RDF Refine, <http://refine.deri.ie/>

Extrakce z volného textu

- Nejnáročnější a nejméně spolehlivá, využívá lingvistických technik
- Spíše identifikace entit a jednotlivých trojic než vyčerpávající převod obsahu dat
- Často zaměřeno na novinové zprávy nebo blogy, RDF jen jako jeden z výstupních formátů
 - <http://www.opencalais.com/>
 - <http://www.zemanta.com/>

Ruční extrakce z tabulek

- Tabulky (HTML, XLS, CSV apod.) menšího rozsahu lze převádět ručně
- Nejpopulárnější nástroj: OpenRefine
 - <http://openrefine.org/>
 - Původně GoogleRefine
- S pluginy pro LD je označován jako LODRefine
 - Vyvinuto v projektu LOD2, ve spolupráci slovinské firmy Zemanta, DERI, a dalších partnerů

OpenRefine

- Import dat ze souboru nebo přes clipboard
- Čištění a transformace dat v tabulkovém zobrazení
- Propojení dat na webové služby
- Linkování na entity z externích znalostních grafů
- Využívání skriptovacího jazyka GREL (podobného JavaScriptu)
 - Využívání regexp... analogicky k FILTER v SPARQL

LODRefine navíc nabízí

- Mapování na DBpedii, ev. další endpointy
 - Rekonciliace – přiřazení hodnot k URI
 - Augmentace – „dotažení“ dalších dat o entitě
- Extrakci entit z volného textu v polích tabulky
- Napojení na CrowdFlower – evaluace rekonciliace pomocí crowdsourcingu
- Export do RDF
 - Interaktivní návrh kostry RDF grafu
- Viz webinář na <https://www.youtube.com/watch?v=4Ve93C238gl>
- Aktuální funkční verze RDF rozšíření:
<https://github.com/stkenny/grefine-rdf-extension/>

Alternativní postupy pro extrakci RDF z tabulek

- „Tabulko-centrický“
 - Interaktivní předzpracování (čištění, linkování) tabulky
 - Např. LODRefine - využití regex v GREL, rekonciliace entit např. s DBpedií
 - Následné vytvoření struktury RDF
- „RDF-centrický“
 - Dávkové převedení tabulky do jednoduchého RDF
 - Např. tarql
 - Následné zpracování pomocí SPARQL
 - UPDATE, případně CONSTRUCT operace využívající mj. regex ve FILTER klauzuli
- Nebo něco mezi tím...

Nadstavby pro ETL proces

- Aktuálně ke špičce v této oblasti patří mj. dva „pražské“ nástroje pocházející ze společného základu
 - MFF UK Praha, lehce spolupracuje i VŠE
 - Oba používají pipelines složené z Data Processing Units (DPU), řeší i jejich ladění a rozvrhování
- Unified Views
 - <http://unifiedviews.eu/>, spolupráce MFF UK a firem Semantic Web company, EEA a TenForce
- LinkedPipes ETL
 - <http://etl.linkedpipes.com/>, aktuálně vyvíjené s podporou projektu EU OpenBudgets.eu; viz cvičení