Adam Blaščík
20. listopadu 2016

LinkedSpending
Referát z článku

Autoři Höffner, Michael Martin a Jens Lehmann prezentují projekt transformace dat týkajících se veřejných rozpočtů na otevřená propojená data. Konkrétně se jedná o data o veřejných rozpočtech publikovaná v rámci projektu openspending.org.

Projekt OpenSpending si klade za cíl stát se centrálním a otevřeným portálem pro informace o veřejných rozpočtech. V současné době obsahuje 1104 datasetů ze 75 zemí světa. Pro našince je jistě zajímavé, že je obsaženo i 7 datasetů z České republiky. V rámci projektu jsou data ukládána v databázových tabulkách. Je možné data filtrovat, vizualizovat a vyhledávat. Nedostatky tohoto systému jsou především: i) chybějcí možnost tvorby složitějších dotazů, například takových, které obsahují odvozené indikátory. ii) absence možnosti porovnávat jednotlivé datasety mezi sebou.

Oba výše zmíněné nedostatky řeší transformace projektu OpenSpending do propojených dat ve formátu RDF. Ta s sebou přináší i další výhody jako napojení na další datasety (například DBpedii pro informace o měnách) či snadnou automatizaci analytických úloh nad daty. Obojí je velmi důležité, vzhledem k tomu, že se dá v budoucnu předpokládat značný nárůst veřejně dostupných finančních dat. Zveřejňování veřejných rozpočtů napomáhá zejména společenská poptávka po transparentnosti, ale také to, že v mnoha zemích je zveřejnění dat určeno zákonem. Například v rámci EU existuje projekt Financial Transparency System, který ale z povahy věci trpí stejnými nedostatky jako OpenSpending.

Z těchto důvodů přistoupili autoři k transformaci datasetů z OpenSpending do RDF. Vzhledem k tomu, že se jedná o multi-dimenzionální statistická data, přirozenou volbou byl slovník Data Cube, který je kompatibilní s dalšími projekty propojených veřejných dat (např. LOD2). Nebylo zapotřebí žádného významného předzpracování dat, ale přesto se autoři setkali s několika problémy: i) Frekventované přidávání nových datasetů (v průměru 50 měsíčně), či jejich modifikace. ii) Chybějící hodnoty. iii) V různých datasetech mají atributy stejný název, ale je nejasné, zda význam je skutečně stejný. iv) Datasety se ve své vnitřní

struktuře ne zcela přesně drží data cube modelu. v) Neexistující jazykové tagy. Vlastní transformace byla, vzhledem k značnému množství dat a jejich častému přidávání, provedena automatizovaně pomocí programu, kteří autoři naprogramovali v prostředí JDK a zveřejnili jako open source. K popisu dat byly použity již zavedené datové slovníky (RDF, RDFS, OWL, XSD a DCMI), což usnadňuje jak uživatelskou práci s daty, tak linkování na další datasety.

A právě odkazování k dalším zdrojům je dle mého názoru nejsilnější stránka propojených dat. V tomto případě nebyla zatím tato schopnost naplno využita, v současné době existuje propojení s DBpedií a projektem LinkedGeoData.

Projekt LinkedSpending je zveřejněn v rámci prostředí OntoWiki, které umožňuje snadný uživatelský přístup a práci s daty. Možné je také použít veřejný SPARQL endpoint. Existuje množství dalších podobných projektů zabívající se daty z veřejných rozpočtů, jako například americký Data-gov nebo britský COINS. Některé tyto databáze byly již také transformovány do formátu propojených dat. Největší výhodou takové transformace je určitě možnost vzájemného propojení databází v rámci celého světa, které umožní jak vzájemné srovnání data, tak pokročilé analytiké úlohy. Bohužel v současné době se jedná stále spíše o izovolané ostrůvky dat než o globální web veřejných rozpočtů.

odkaz na článek: http://www.semantic-web-journal.net/system/files/swj923.pdf