Linked Open Data: konverzió és vizualizáció

Micsik András, Turbucz Sándor, Tóth Zoltán
MTA SZTAKI, 1111 Budapest, Lágymányosi utca 11.
{micsik.andras, turbucz.sandor, toth.zoltan}@sztaki.mta.hu

Sziládi Zoltán
Nokia Solutions and Networks, 1092 Budapest, Köztelek u. 6.
zoltan.sziladi85@gmail.com

Kivonat: *A Linked Open Data (LOD) mozgalomhoz kapcsolódva a SZTAKI 2011 óta működteti a lod.sztaki.hu szolgáltatást, amely magyar kulturális adatokat tartalmaz a korábbi HEKTÁR és NDA projektek gyűjtött adatai alapján. Az XML formátumú adatok RDF-re konvertálása és "lodifikálása" után felmerült az igény arra, hogy a felhasználók is kényelmesen tudják az adatokat böngészni és áttekinteni. Erre készült a LODmilla nevű, Javascript alapú grafikus böngésző, amely a LOD gyűjteményeket mint gráf és szöveges rekord egyszerre tudja megjeleníteni. A szoftver alapfunkciója a megjelenített gráfrészlet kezelése, mint például nagyítás, csúcsok mozgatása, új csúcs megnyitása, de tartalmaz közösségi funkciókat is: a gráf nézetet el lehet menteni, és megosztani másokkal. Példaképpen néhány bonyolultabb funkciót is megvalósítottunk: útvonal keresése csúcsok között, adott típusú csúcsok és élek keresése. Mindemellett a csúcsokhoz tartozó lokális adatok (data properties) kényelmes böngészése is megoldott egy szöveges lista formájában. A LODmilla szolgáltatás és forráskódja is nyílt hozzáférésű, valamint nem csak a lod.sztaki.hu adataival hanem tetszőleges más LOD tárolóval is kipróbálható.*

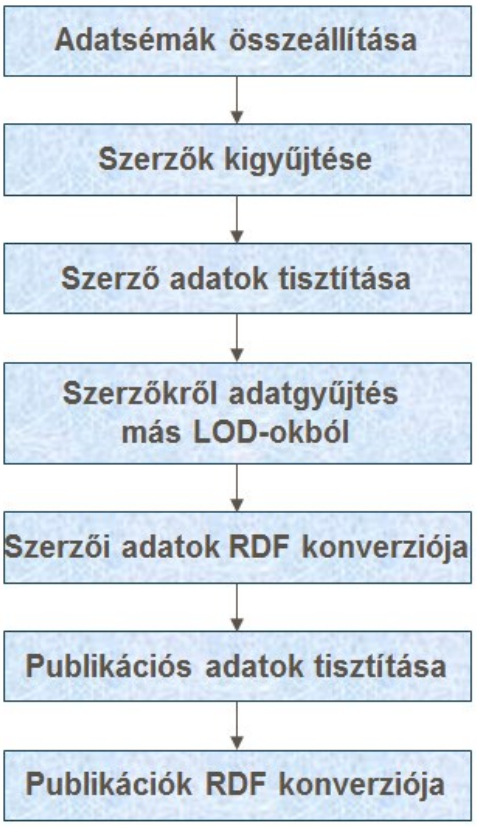
Bevezetés

A Szemantikus Web terve szerint a világhálón gépileg feldolgozható adatokat közzétéve, azokat összekapcsolva elosztott tudásbázisok alakíthatók ki. Bár a Szemantikus Web alapkövei (RDF, SPARQL, OWL) közül számos több, mint tíz éve a WWW Consortium (W3C) ajánlása, ennek ellenére nem tekinthetők széles körben elterjedtnek. Az eredeti elképzelés egyszerűsítve és kicsit újrafogalmazva Linked (Open) Data néven viszont ígéretesebbnek tűnő jövő előtt áll. Tim Berners Lee 2006-ban fogalmazta meg először az összekapcsolt adatok alapelveit []. A LOD Cloud projekt kezdte gyűjteni az elérhető LOD adatokról a statisztikákat, és 2011-re ezek szerint a világon 295 adathalmaz és 31 milliárd elemi tény (triple) állt rendelkezésre LOD formátumban. Azóta ezek a számok szinte követhetetlen gyorsasággal tovább nőttek.

A LOD hazai története ennél sokkal rövidebb. [http://nektar.oszk.hu/wiki/Semantic_web, Networkshop?] 2010-ben az OSZK publikálta teljes katalógusát (szerzői és kiadvány adatok) LOD formátumban. Az MTA SZTAKI 2011 folyamán konvertálta a korábbi HEKTÁR és NDA projektjei során gyűjtött metaadat rekordokat Linked Data formátumra, és ezzel a tartalommal elindította a lod.sztaki.hu szolgáltatást. Az MTA SZTAKI Elosztott Rendszerek Osztályán azóta is folynak kutatások a Szemantikus Web és a LOD érdekes alkalmazási lehetőségeiről; ezekről fogunk részlegesen itt beszámolni.

LOD konverzió

Az NDA projekt OAI-PMH begyűjtései során körülbelül 800.000 metadata rekord gyűlt össze 16 adatszolgáltatótól (pl. MEK, Magyar Filmunió, MATARKA, A38 hajó, radio.sztaki.hu). A rekordok többek között leírnak könyveket, folyóiratcikkeket, rádióműsorokat, filmeket, festményeket, fotókat az egységes Dublin Core rendszer szerint.



1. ábra: Konverziós lépések

A „lodosítás" folyamatának első lépéseként (1. ábra) meg kellett határozni, hogy milyen entitásokat hogyan akarunk leírni, vagyis milyen RDF sémákat használunk majd. Két fő típust különböztettünk meg, a szerzőt és a művet, ezeket több séma szerint is osztályba soroltuk (pl. schema.org:CreativeWork, dbpedia-owl:Work, foaf:Person, schema.org:Person). A legfontosabb lépés annak a meghatározása volt, hogy milyen tulajdonságokat milyen sémaelemekkel fogunk leírni. Természetesen adta magát a Dublin Core RDF sémája, a dcterms, ezzel a művek legtöbb tulajdonságát le lehetett írni, mivel az adatok amúgy is Dublin Core formában érkeztek.

TULAJDONSÁG	TARTALOM
rdf:type	foaf:Person, schema:Person, stb.
foaf:name	Név
rdfs:label	Név
dcterms:alternative	Eredeti (nem feldolgozott) név
owl:sameAs	VIAF, OSZK, Dbpedia kapcsolatok
dbpedia:birthYear	Születési év
dbpedia:deathYear	Elhalálozási év

1. táblázat: RDF sémák használata a szerzők leírására

E gyakorlattól a szerzők esetében kellett eltérnünk, mivel ott a FOAF séma sokkal praktikusabban fogalmazza meg a személyek tulajdonságait. Egy másik szempont a LOD saját hagyományainak követése volt, ezért az rdfs:label, az rdf:type, illetve néhány elterjedt Dbpedia-ban használt tulajdonságot is bevettünk a listába (1. táblázat).

A legtöbb gondot a szerzők külön feldolgozása jelentette, mivel a szerzőket különálló entitásként akartuk ábrázolni. Ehhez ki kellett gyűjteni az összes szerzői nevet az összes tételből, majd ezeket valahogyan egyesíteni. A feldolgozás nem lett és nem is lehetett tökéletes, de nagy százalékban jól sikerült a szerzőket és műveiket összekapcsolni, ami végül a Linked Data lényege.

A szerzői nevekhez a katalogizálás során mindenféle egyéb adat rakódott hozzá, pl. „Gárdonyi Géza szerk. 1863-1922”. Ezeket egyrészt le kellett vágni a névről, másrészt segítséget jelentettek mondjuk a szerző születési évének kivonásához. További segítéget nyújtottak más szerzői nyilvántartások, például az OSZK korábban említett LOD szolgáltatása, a VIAF, és néha a Dbpedia is. Ezeken a helyeken megpróbáltuk a mi szerzőinket beazonosítani, amely kettős nyereséggel járt: egyrészt pontosítani tudtuk saját adatainkat, másrészt kapcsolatokat létesítettünk külső LOD adatokkal. A LOD felhő annál jobb, minél több az összeköttetés a különböző adathalmazok között. A 375,000 talált szerzői előfordulásból végül 130,000 szerzőt sikerült legalább egy külső adathalmazhoz kapcsolnunk (a maximum három külső kapcsolt volt, híres személyek esetén).

Az előkészítő folyamat után a művek feldolgozása következett, amely során azokat a megfelelő szerző egyedekhez kapcsoltuk, tisztításokat és javításokat végeztünk egyes tulajdonság adatokban, és a végül Turtle formátumban kiírtuk a kapott mű leírását. Ezeket feladatokat számos PHP szkript írásával lehetett megoldani.

Végül, a szolgáltatást a Virtuoso szerver telepítése és konfigurálása után az adatok betöltésével el lehetett indítani. A Virtuoso beépítve tartalmaz egy SPARQL végpontot, egy egyszerű táblázatos adatböngészőt, valamint szükséges adatlekérési mechanizmusokat is támogatja (dereferencing URIs). Ez utóbbi arra szolgál, hogy automatikusan és deklaratívan is meg lehessen határozni, hogy a lekért adatokat milyen formátumban kapja meg a kliens. Gépi kliens automatikusan RDF-et kap, míg webböngészők esetén az RDF egy HTML táblázattá konvertálva jelenik meg. Ha megadjuk a megfelelő kiterjesztést (pl .ttl), akkor pedig a kért Turtle formátumot kapjuk meg. Ezek összessége alkot egy teljes LOD szolgáltatást, és teszi lehetővé, hogy emberek és gépek egyaránt sokoldalúan felhasználhassák az adatokat.

A lendületünket felhasználva, egy sokkal kisebb, de más problémákat rejtő adatkészletet is LOD formára hoztunk: a SZTAKI publikációinak katalógusát. A szerző és mű megfeleltetés itt sokkal könnyebb volt a pontosabb belső ábrázolásnak köszönhetően. Viszont a szerzők sorrendjének megőrzése az előzőtől kissé különböző ábrázolási megoldást igényelt.

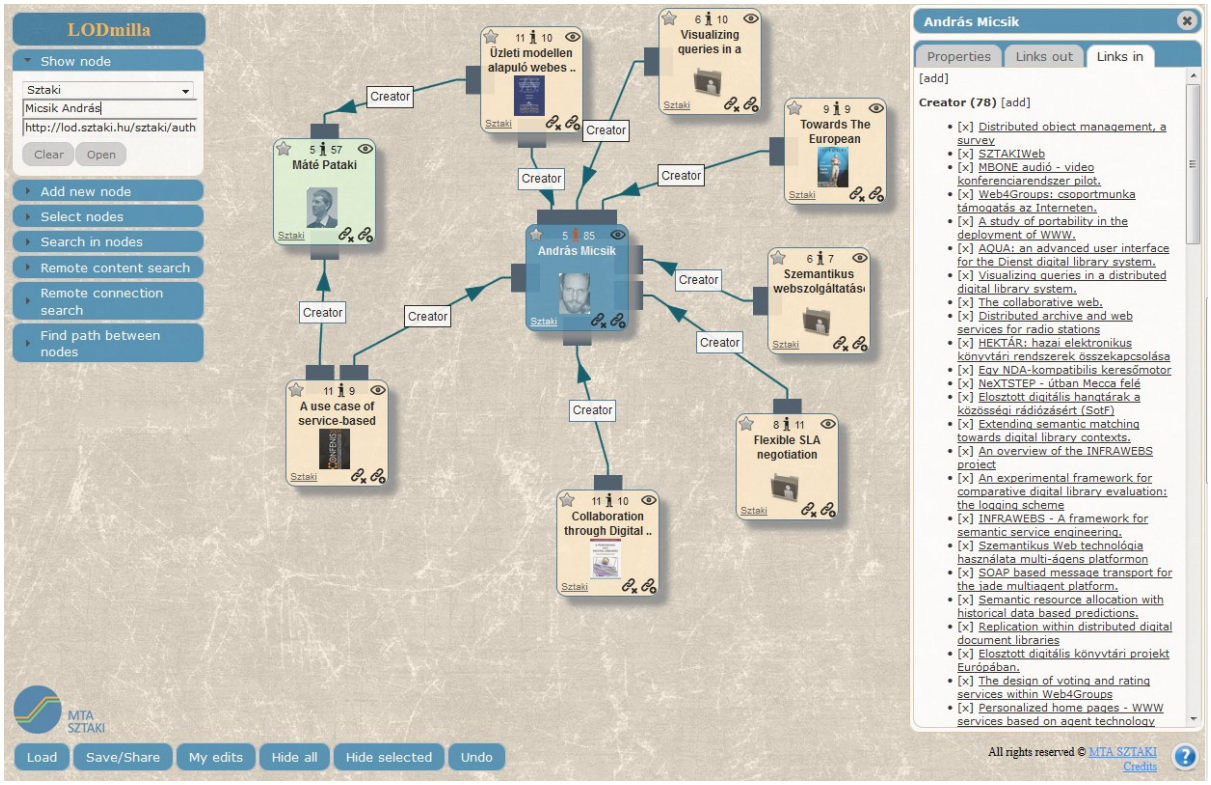
A LODmilla böngésző

A LOD szolgáltatás készítése és használata során tapasztaltuk, milyen nehézkes az RDF gráfokban tárolt adatokat áttekinteni, ellenőrizni. A Virtuoso ad egy táblázatos böngészési lehetőséget, de itt egyszerre csak egy entitáshoz (egy URI) kapcsolódó tulajdonságokat lehet látni, a korábban olvasottakat „fejben kell tartani”. Utánanézve, sok érdekes kezdeményt, de semmilyen használható vizualizációs megoldást nem találtunk RDF gráfokra. A megoldások között volt táblázatos (Tabulator) és gráf-alapú (oobian, LodLive, RelFinder), de vagy a

telepítés és konfigurálás nehézségein vagy használhatósági problémákon akadtunk el. Összefoglaltuk elvárásainkat egy LOD böngészővel szemben:

- Lehetőleg egyedi konfigurálás nélkül használható legyen minden „szabályos” LOD adatkészletre.
- Legyen gráfszerű ábrázolás, amelyben a kapcsolatok élekként látszanak.
- A gráfot lehessen kicsinyíteni, nagyítani, mozgatni.
- Legyen táblázatos ábrázolás is, amely az egyedek nagy számú tulajdonságának áttekintését segíti.
- Támogassa a speciális gráfkereséseket.
- Lehessen menteni, betölteni és megosztani gráf nézeteket.

A fenti követelményeknek eleget tevő saját fejlesztésű böngésző a LODmilla nevet kapta, és mind felhasználása, mind forráskódja nyílt, azzal a reménnyel, hogy a további bővítmények fejlesztéséhez külső segítséget is kapunk.



2. ábra: A LODmilla böngésző

A böngésző főbb elemeit a 2. ábrán mutatjuk be. Középen látható a gráf nézet, amely az általunk választott egyedeket és azok választott kapcsolatait jelenítik meg. Ezáltal ki tudjuk emelni a fontosnak tartott összefüggéseket, és csak a lényeget jelenítjük meg. A jobb oldalt látható info-panelben a választott egyed kimenő és bejövő éleinek teljes listáját kapjuk, tulajdonságtípus szerint csoportosítva. Ugyanitt olvashatjuk egy másik fülre kattintva az adattulajdonságok listáját is, például név, munkakör, születési év, stb.

Az egyedeket ábrázoló gráfcsúcsokban láthatjuk a tulajdonságok számát, és az i betűre kattintva az info-panelt aktiváljuk. A csillag ikonnal kiválasztani lehet egyedeket, a szem ikonnal eltüntetni. A bal alsó sarokban az adat forrása látható.

Az alsó sávban az általános műveletek kaptak helyet: névvel elmenthetjük az aktuális nézetet, és az erről kapott egyedi linket másokkal megoszthatjuk. Tiszta lapot is kérhetünk, vagy az utolsó művelet visszavonását.

A bal oldali harmonika-panelekben helyezkednek el az összetett műveletek kezelőgombjai. Mivel ezek többsége több csúcsponton működik, ezért a kiválasztás módosítása fontos műveletként helyet kapott. A különféle keresések közül az első a megjelenített egyedek tulajdonságaiban keres adott szövegrészlet előfordulásaira. Ezt a keresést a következő panelben kibővítetten a csomópontok szomszédságában hajthatjuk végre. Így például meg tudjuk keresni, hogy az adott személy környezetében előfordul-e a „szemantikus" szó az adattulajdonságokban. A harmadik keresési művelet a tulajdonságok neveiben keres ugyanígy, tehát ha a creator szóra keresünk, akkor a szerzőségi kapcsolatok kibontását kapjuk meg az egyed környezetében. Ezeknél a műveleteknél meg kell adnunk a hatókör sugarát, vagyis hogy hány élnyit távolodhat el a keresés a kezdőponttól. Az utolsó művelet útvonalakat keres két választott csúcs között. Szemantikailag ezzel a művelettel a „Mi a kapcsolat X és Y között?" kérdésre próbálunk választ találni.

Új fejlesztésként a böngészőben lehetőség nyílt az RDF gráf módosítására is. Az éleket törölhetjük vagy áthúzhatjuk más csúcsba. Új éleket húzhatunk be vagy új adatokat írhatunk be az info-panelbe. A változtatások SPARQL UPDATE parancsokként összegyűjtve a „My edits" gombbal hívhatók elő. A kapott SPARQL parancsokat már a mi feladatunk végrehajtani vagy továbbítani az adatok gazdájának, de mindenestre ezzel a módszerrel az RDF-hez nem értők is könnyedén javíthatják az RDF-ben tárolt adatokat.

A LODmilla felülete Javascriptben íródott, és a legtöbb modern webböngészőben használható. A tárolást és a keresési feladatokat egy szerveroldali Java komponens végzi.

Összefoglalás

Beszámoltunk a SZTAKI-ban a Linked Open Data-val kapcsolatos tevékenységeink egy részéről, melyek egyrészt a Linked Open Data előállításának gyakorlatát tanulmányozzák, másrészt az így előállt adatok felhasználását. Ennél az új technológiánál is fontos az, hogy a felhasználók számára barátságos kezelhetőséget biztosítsunk, melyre a LODmilla böngészővel keresünk megoldást. A LODmilla bebarangolhatóvá teszi az RDF gráfokat a laikusok számára is, valamint egyedi pillanatképek készíthetők a gráfról, amelyeket másokkal is meg tudunk osztani. Ezáltal az RDF gráfban tárolt tudásnak adott kivetüléseit tudjuk rögzíteni, egyedi tudástérképeket tudunk rajzolni. És nem csak megjeleníteni tudjuk a tudást, hanem javítani is különösebb képzettség nélkül, az LODmilla szerkesztési funkcióit alkalmazva.

Hivatkozások

[1] Berners-Lee, T. Linked-data design issues. W3C design issue document, June 2009. http://www.w3.org/DesignIssue/LinkedData.html

[2] Schenk, S., Gearon, P., and Passant, A. Sparql 1.1 update. Tech. rep., W3C, 2008. Published online on October 14th, 2010 at http://www.w3.org/TR/2010/WD-sparql11-update-20101014/

[3] András Micsik, Sándor Turbucz and Zoltán Tóth. Browsing and Traversing Linked Data with LODmilla. ERCIM News 96, Jan 2014, http://ercim-news.ercim.eu/en96/special/browsing-and-traversing-linked-data-with-lodmilla