A könyvtári szemantikus webes fejlesztések világa. Gondolatok az SWIB konferenciasorozat kapcsán

Bevezetés

Az utóbbi években a könyvtárakat érintő szemantikus webes fejlesztések egyre látványosabb formát öltenek. E terület talán legfontosabb fóruma hosszú évekre visszanyúlóan az SWIB (Semanthic Web in Bibliotheken) konferenciasorozat. Fő érdeme, hogy minden évben workshopok és előadások formájában megvilágítja a kevésbé informatikus lelkületű könyvtáros szakemberek számára is, hogy mit jelent az információkeresés és -szervezés átalakulása, az RDF alapú adatmodellre épülő nyílt kapcsolt adatok által kínált új utak megjelenése a saját munkájuk jövőjére nézve, illetve miként fejlődik teljesen a szemünk előtt ez a terület. A konferenciák teljes programja, beleértve a legutóbbi 2018-as rendezvényét visszanézhető videofelvételeken keresztül a weboldalról kiindulva (http://swib.org/swib18). Néhány előadásról szegedi kollégáim készítettek összefoglalót. Ebben a cikkben inkább a konferencia workshopjai által tárgyalt témakörökről, a szemantikus webes könyvtári fejlesztések mögött rejtőző dilemmákról adok számot.

A nyílt kapcsolt adatok világával kapcsolatos workshopok

Bevezetés a nyílt kapcsolt adatok világába

Ha áttekintjük az SWIB hosszú idő óta meglehetősen állandó szerkezetét, feltűnik, hogy az előadások mellett meghatározó súlya van a workshopoknak is. Az itthoni Networkshophoz hasonlóan ezek a workhopok-tutoriálok lehetővé teszik a gyakorlottabb szakembereknek az egyes szemantikus webes részterületekben való fokozott elmélyülést. Ám ami még fontosabb, hogy minden évben van olyan workshop is, ahol a szemantikus webes információkeresés, a nyílt kapcsolt adatok világával meg lehet ismerkedni. S természetesen, habár ezek a gyakorlati jellegű foglalkozások az előadásoktól eltérően nem kerülnek képi rögzítésre, a prezentációk széleskörűen felhasználhatók akár belső könyvtáros továbbképzések segédanyagaiként is.

2018-ban *Christina Harlow* a Stanford, illetve *Simeon Warer* a Stanford Egyetemről vállalták a nyílt kapcsolt adatok világát feltáró bevezető workshop megtartását. Felvázolták, hogy miként néz ki az RDF gráf alapú adatmodellje igen egyszerű köznapi példák ábrázolásán keresztül (Mutasd be magad rdf leírás formájában, s ábrázold a kapcsolatod másokkal az ő bemutatkozó dokumentumtartalmaikon keresztül.) Nemcsak azt próbálhatták ki így a résztvevők, hogy miként lehet RDF dokumentumokat létrehozni, hanem rögtön azt is, hogy miképpen tudják megosztani egymással azokat, mini tematikus felhőket építve fel a kapcsolt adatokból. Így a területtel csak most ismerkedő felhasználók számára is nyílvánvalóbbá válik az adatleírás logikája, illetve a gráfhoz kötődő összekapcsolhatóságban rejlő előnyök is. Az alapok elsajátítása után lehet áttérni arra, hogy a különféle adatelemeket, adatkészleteket magában foglaló gráf miként ölt globális formát. S itt ismertették azokat az alapelveket, melyek kapcsán *Tim Berners Lee* a szemantikus web felépülését elképzelte: 1. Nevezd meg a dolgokat URI azonosítók révén, 2. tedd azokat a http protokollhoz kapcsolódva kereshetővé, 3. Ha valaki rákeres egy URI-ra, szolgáltass érdemi információt szabványos formában az azonosító mögött rejlő tartalomról (ebben segít az RDF adatszerkezet, s a SPARQL visszakereső nyelv), 4. Helyezz el linkeket más tartalmakra, hogy így könnyebben visszakereshetők legyenek a saját tartalmakhoz kapcsolódó egyéb dokumentumok is. Az alapelvek tisztázása után, miközben mindenki leírta RDF nyelven a saját magáról szóló alapinformációkkal, benne a munkára, hobbyra utaló adatokkal, neki lehetett állni kapcsolódó tartalmakat keresni nyílt kapcsolat adat alapú adatbázisokból (wikidata, DBPedia) s URI azonosítók segítségével hozzákötni a saját

tartalmainkhoz. A következő lépés annak áttekintése, hogy milyen adatmodellek, adatkészletek, szótárak keretében írják le a szemantikus adatokat, miképpen lehet átkonvertálni már meglévő, például Dublin Core alapú adatkészleteket RDF alapú állítások formájában megosztható nyílt kapcsolt adattá.

A workshop második része pedig azt járta körül, hogy mit jelent az adatok nyíltsága. Milyen jogi keretei vannak az adatkészletek licencelésének, miként lehet biztosítani az adatok újrafelhasználhatóságát. A következő részben pedig fejest ugorhattak a résztvevők abba, hogy miképpen épülnek fel az RDF alapú dokumentumokat tároló speciális adatbázisok (az ún. triplestore-ok), illetve hogyan lehet az azokban rejlő információkat visszakeresni a SPARQL nyelv parancsainak segítségével. A LODLIve alkalmazás bemutatásával pedig azt is áttekinthették a résztvevők, hogy miként lehet a nyílt kapcsolt adatkészleteket vizualizálni. Zárásként pedig be lett mutatva néhány már létező és dinamikusan fejlődő nyílt kapcsolt adatok felhasználására épülő projekt, ezt követően pedig a résztvevők megvitathatták, hogy milyen felhasználási lehetőségeket látnak a saját munkájukhoz, tapasztalataikhoz kapcsolódva e területen.

A workshop nagyszerűsége tehát abban állt, hogy néhány óra alatt a teljes ismeretlenségből kiindulva már a konferencia elején tisztába kerülhettek a résztvevők a szemantikus adatkezelés, adatmodellezés alapjaival, azzal a hozzáadott értékkel, amit a szabványos formában leírt adatok összekapcsolhatósága, kombinálhatósága, újrafelhasználhatósága jelent.

A nyílt kapcsolt adatok felhasználásának esettanulmányait ábrázoló workshopok

A nyílt kapcsolt adatokkal már hivatásszerűen foglalkozó német nyelvterületről érkezett szakemberek a szemantikus metaadatkezeléssel kapcsolatos speciális kihívásokat tekinthették át. Az Észak-Rajna-Vesztfáliai Hochschulbibliothekszentrum munkatársai *Fabian Steeg*, *Adrian Pohl* és *Pascal Christophe* a nyílt kapcsolt adatok publikálási módjaiba s a széleskörűen használható adatkészletek előállítási módjaiba avatták be a hallgatóságot.

Jakob Voß és *Joachim Neubert* a ZBW Leibniz Information Centre for Economics munkatársai a wikidata használatába nyújtottak betekintést. Áttekintést kaphattunk arról, miként lehet már meglévő adatkészleteket importálni, összekötni a már a rendszerben lévő többivel. A központi adatbázisból kiválasztott minta adatkészleteken keresztül pedig az összes kapcsolódó adatkezelési munkafolyamatot, eljárást is ki lehetett próbálni néhány adatkezelő eszközzel együtt.

Stacy Allison-Cassin és *Dan Scott* a kanadai York és Laurentian egyetemek munkatársai, az eredetileg a Wikidata projekt keretében fejlesztett Wikibase-t mutatták be, ami voltaképpen programkönyvtárak és alkalmazások gyűjteménye a strukturált adatok kezelésének céljából. Nagyon fontos az, hogy soknyelvű kezelőfelülettel bír, szerkesztőfelülete könnyen átlátható metaadatkezelő szakemberek számára. Olyan programkörnyezetet kínál SPARQL végponttal, amely lehetővé teszi az egyes intézményeknek, hogy ne csupán a Wikidata felületén publikálják adatkészleteiket, hanem saját, helyi környezetben működő szemantikus adatbázist is üzemeltessenek. A résztvevők feltelepíthettek egy virtuális gépet, rajta a Wikibase rendszerrel, s így első kézből ki tudták próbálni a szemantikus adatkészletek összeállításának és publikálásának lépéseit.

A kanadai Duraspace-t képviselő *David Wilcox* a Fedora repozitóriumi rendszert érintő szemantikus webes funkciókat, új fejlesztéseket mutatta be, illetve tette kipróbálhatóvá a résztvevők számára, a rendszer szemantikus webes adatbeviteli, adatmenedzselő és adatkereső moduljain át. Mindenki előretelepített virtuális gépet kapott a Fedora rendszerrel, a SolR keresőalkalmazással, s a szemantikus tripleteket tároló adatbázissal (triplestore) együtt. A workshop kitért arra is, hogy a nem nyilvános érzékeny nyílt kapcsolt adatok formájában tárolt adatkészletekhez kötődő webes hozzáférést miképpen lehet szabályozni. Fontos eleme az egész Fedora rendszernek a nyomon követhetőség, skálázhatóság, az adatkezelési műveletek visszakövethetősége, akár egyedi fájlok verzió kezelésének szintjén is. Az ehhez kötődő munkafolyamatok áttekintésére is lehetőség nyílt a workshop keretében.

A nyílt kapcsolt adatok világában a legnagyobb előzetes elmélyülést igénylő workshopot a Yale Egyetem, illetve partnerként fellépő kutatóintézetek munkatársai tartották az RDF alapú adatmodellek megosztásának módjairól, valamint arról, hogy miképpen lehet ezeket az RDF alapú adatkészleteket validálni. Erre a célra az RDF alapú adatbázisokhoz illeszkedő Shape Expressions formalizáló, modellező és validáló programnyelvet használ-

ták fel. Konkrét példákat mutattak be arra, hogy miként tudják a szemantikus adatkészleteket használó emberek, illetve gépek közötti kommunikációt megkönnyíteni a különféle adatkészletek közötti kommunikáció megkönnyítésével. Áttekintették az adatkészlet validálási munkafolyamatait. A bibliográfiai területről szolgáltak példákkal arra, hogy miképpen lehet modellezni az egyes adatkészleteket. Ahogy egyre nagyobb teret hódít a nyílt kapcsolt adatok használatára, s egyre több hangsúly kerül a problémamentes kommunikáció feltételeinek előmozdítására a megfelelő adatmodellezési háttérrel s adatkezelő munkafolyamatokkal együtt, úgy válik egyre inkább kulcsfontosságúvá az itt leírt feladatkör.

Szemantikus webes könyvtári fejlesztések főbb trendjei az SWIB konferencia sorozat fényében

James Hendler bevezető előadása érdekes összefüggésrendszerbe állította az előző nap a workshopokon tárgyalt témaköröket. A tudásgráf alapú információszervezés koncepcióját már a web születése kapcsán 1989 táján felvázolták a CERN-ben, Tim Berners Lee pedig 1994-ben állt elő a webet átfogó koncepciójával, melyben már utalt a szemantikus dimenzió szükségességére, habár konkrét elképzelésekkel 2001-ben állt elő ennek kapcsán. A Google 2012-től kezdte nagy erővel fejleszteni a tudásgráfját, abból a célból, hogy az addigiaknál is több terhet vegyen le a keresést végző felhasználók válláról. Rövid idő alatt látványos eredményeket értek el. Hendler is utalt arra előadásában, hogy egyes becslések szerint a Google keresések megválaszolásakor már mintegy 40%-ban támaszkodik szemantikus értelmezést lehetővé tévő metaadatokra, illetve az ebből felépülő tudásgráf szolgáltatásokra. Tehát ebből is látszik, hogy jóval korábbi ez a történet, mint ahogy elkezdett konkrét formát ölteni, habár az utóbbi években gyorsult fel igazán a fejlődés dinamikája. A hangsúly azonban véleményem szerint még mindig döntően a nyílt kapcsolt adatokon alapuló adatmodellek, szoftvereszközök, adatbázis-alkalmazások fejlesztésére helyeződik. Jónéhány olyan ország van Európában, ahol legalább a tudományos könyvtári szférában hosszú évek szorgos tudományos ismeretterjesztő munkájának köszönhetően meghonosodott az az újfajta szemléletmód, amely a hagyományos katalogizálási, információkeresési megoldásokon alapulva ugyan, de mégis meglehetősen új környezetet jelent. Egyre több adatforrást ültetnek át RDF alapú adatszerkezetbe, kiépül a SPARQL végpontok hálózata, melynek révén az összekapcsolhatóság is egyre határozottabb formát ölt. Sokkal elfeledkeznek azonban arról, hogy az emberi tényezőnek még mindig kritikus jelentősége van. A könyvtárak felelőssége hatalmas abban, hogy ellenőrzött, jó minőségű adatokat tegyenek nyílt kapcsolt adatokként közkinccsé, megfelelő modellekbe szervezve. Hendler találó megfogalmazása szerint a történet kritikus pontja nem az összekapcsoltság, hanem a megfelelő minőségű metaadatforrások összekapcsoltsága. Abban, hogy megtaláljuk a megfelelő tartalmakat, majd új formában feldolgozva integráljuk, interoperábilissá és újra felfedezhetővé tegyük őket, hatalmas az egész közintézmény felelőssége. Számos kihívást rejt magában a különféle adatmodellekben feldolgozott adatok közös platformon történő kezelhetősége, hiába vannak meg elvileg a közös keretek a nyílt kapcsolt adatok rendszerében. De még nagyobb kihívást jelent az adatok felfedezhetőségének, újrafelhasználhatóságának biztosítása a felhasználók szintjén. A Google itt látványos eredményeket mutat fel, de nagy kérdés, hogy mekkora kockázat lehet az, hogy egy kereskedelmi cég algoritmusai telepednek dominánsan rá erre az információszolgáltatási szegmensre is. A szemantikusan közzétett adatok és az értelmezési lehetőségek bemutatásában sajnos még sokkal kevesebb előrehaladás történt, mint az adatkészletek publikálása terén. Ami persze valahol természetes is, hiszen meg kell teremteni az alapot, melyet bemutatni, szolgáltatni lehet. Megítélésem szerint azonban egyre nagyobb a nyomás a közgyűjteményi szférán, hogy ebbe az irányba látványosan elmozduljanak, mielőtt más teszi meg helyettük (pl. a Google), dominánsan uralva a nyílt formában közzétett adatok felhasználói szintű reprezentációját is. Hendler előadásában megemlít olyan egyelőre béta állapotú alkalmazásokat, melyek olyan információ-visszakereső szoftveres ügynököket alkalmaznak, amelyek elemzik az információentitásokat, eseményeket azok kapcsolataival együtt s kontextusfüggően nyújtanak információkat a felhasználók számára. Ezek öntanuló eszközök, tehát minél inkább elterjed a használatuk, annál inkább tudják azonosítani a felhasználói igényeket, illetve lehetőség van például kutatási, oktatási célú célzott beállítások alkalmazására is. Hendler arra is utal, hogy a szemantikus web segítségével fel tudunk címkézni, azonosítani tudunk és össze tudunk kötni tartalmakat, de amiről a fentiekben szó van, az már átcsúszik a kognitív alapú informatika tartományába, amelyben újfajta keresési eszközöket, heurisztikákat fejlesztenek ki és kínálnak a felhasználók felé. Gyorsan, dinamikusan generálódó adatszolgáltatásról beszélünk a meglévő szemantikus adatokra építve a keresés

alapjául szolgáló webes entitásokból kiindulva, feltárva az adott entitás minél tágabb értelmezési lehetőségét, illetve kontextusát. De Hendler utal arra is, hogy ez még mindig nem elég. Olyan új digitális narratívák megalkotását segítő technológiai háttérre is szükség van, mely egyaránt támaszkodik az eddig felsorolt két tényezőre, a szemantikus webre, illetve a kognitív informatikára. Választ ad arra a kérdésre, hogy miként vegyünk ki információkat a tudásgráfból különféle általunk meghatározott értelmezési utak, lehetőségek mentén, amelyek érdekes és értelmes történetté állnak össze. Hozzáadott értéket jelentenek az eddigi tudásunkhoz hozzáadva, illetve ki tudnak indulni a felhasználók előzetes ismereteiből, vagy akár érzelmi reakcióiból is. A szemantikus metaadatgazdagítás tehát egyaránt szolgálja a különféle nyílt kapcsolt adatforrások integrációját, illetve segíti azt, hogy különféle értelmes narratívákat is fel lehessen azokból tárni. Világosan látszik, hogy könyvtárosoknak, információtudományi szakembereknek, informatikusoknak, matematikusoknak egymással szoros szövetségben kell dolgozniuk e célok megvalósításának érdekében. A digitális bölcsészetek fejlődése az ilyenfajta partnerségi formákon alapulva járulhat hozzá pont ahhoz, hogy a nagy mennyiségű, összekapcsolt adattömegeket értelmes módon meg is tudjuk szólítani majd, minél színesebb, minél többféle célhoz kötött módszerekkel. S akkor meg lesz a remény is arra, hogy ezeknek az újfajta információszolgáltatásoknak egy széleskörű eleven térképe alakuljon ki, amelyet nem egyetlen piaci szereplő fog szinte kizárólagosságra törekedve uralni

Németh Márton
(Országos Széchényi Könyvtár
Elektronikus Könyvtári Szolgáltatások osztálya)