szótárakat. Az első probléma a hozzáférés linkekre épülő természete. Számos probléma vetődik fel, amikor a kapcsolódó források „egy-az-egyben" megfeleltetést alkalmaznak a szótárak között. Ha ki akarnánk számolni a létező szótárak elemeiből összesen létrehozható sorozatok számát, elképesztően nagy – *(n-2)! – mennyiséget kapnánk, ahol „n" a létező szótárak száma. A *W3C Library Linked Data Incubator Group* nagy erőfeszítéseket tesz a szótárak összegyűjtése és értékelése érdekében, hogy ezeket a számokat csökkentse, de ez egy hosszabb folyamat. Egy másik lehetőség lenne egyetlen, központi forrás használata, mely az összes többi szótárhoz kapcsolódna; ekkor n-1 megfeleltetés jöhetne létre az összes lehetséges fogalom között. De a megoldásra váró probléma még akkor is az lenne, hogy a megbízhatóság érdekében milyen hatásoknak kell érvényesülniük a szótárak frissítése, javítása során.

Végigkövetve a vizsgálatot, amellyel a szemantikus visszakereső rendszerek fejlesztéseinek legfőbb követelményeit – mint például használhatóság (használó-központúság?) – azonosították, a kritikus pont ezen eszközök, források népszerűségének biztosítása lesz. A minél nagyobb számú felhasználó bevonása elemi feltétel ahhoz, hogy a szemantikus web és a linked data-technológiák elérhessék saját lehetőségeik legnagyobb kihasználtságát.

Következtetések

A tanulmány a szemantikus dokumentumok visszakeresésére szolgáló módszerek vizsgálatát és értékelését mutatta be. Az eredmények jelen pillanatban azt mutatják, hogy több, a szemantikus források visszakeresésére szolgáló rendszer hiányosságokat mutat azokban a minimális követelményekben is, amelyek pedig növelhetnék népszerűségüket. Ezek közül csak néhányat említve, a problémák a megbízhatóság hiányában; a források leírásának kevésbé szabályozott voltában; a használó-központúság alacsony szintjében; valamint az elemek elszigeteltségében keresendők, mely utóbbi nehézkessé teszi a keresést, és bonyolulttá a fogalmi navigációt.

Ahogy az elemzés kimutatta, egyedül az ontológiakeresők (NEM AZ ONTOLÓGIATÁRAK!) érték el az alig 50 százalékot a minősítés során. A szerzők értékelése szerint fejleszteni kell azokat a tulajdonságokat, amelyek elsődlegesek a szemantikus dokumentumok kezelésében; ilyenek például a szabályozottság, megbízhatóság, többnyelvűség és a keresés szemantikus jellegének kiterjesztése.

A folyamatban lévő kutatások jelenleg arra irányulnak, hogy hogyan lehet meghatározni a megbízhatóság és a minőségi linkek követelményeit. Noha néhány kereső eszköz statisztikák közzétételével segíti a felhasználókat a megfelelő szótárak kiválasztásában, az eddigi vizsgálatok hiányosságaként lehet említeni, hogy ezeknek az adatoknak a felhasználói viselkedésben és a kiválasztásban játszott szerepével még nem foglalkoztak kellőképpen.

Hivatkozás

[1] KITCHENHAM, Barbara: DESMET: A method for evaluating Software Engineering methods and tools: Technical Report TR96-09. University of Keele, Department of Computer Science, 1996.

/MORATO, Jorge Luis – SANCHES-CUADRADO, Sonja – DIMOU, Christos: Evaluation of semantic retrieval systems on the semantic web. = Library Hi Tech, vol. 31, no. 4 (2013) p. 638−656./

(Vass Johanna)

A *linked data* és a *big data* találkozása – a tudásszervezési rendszerek szempontjából

Big data

A változatos, komplex és hatalmas méretű digitális adathalmazok megjelenése a weben életre hívta a *big data* nevű jelenséget. Ilyen „nagy adatokat" termel például a közösségi média, az elektronikus kereskedelem, a kormányzat, a tudományos kutatás... A fogalomnak sokféle meghatározása született; másként közelítik meg ezt a témát a tudósok, a számítástechnikával és az információtudománnyal foglalkozó szakemberek, a tudománypolitikát irányítók és a finanszírozó szervek vezetői. Vannak, akik azt a technológiai kihívást hangsúlyozzák, amit az igen nagy méretű adatállományok –

beleértve természetesen a számadatokon kívül a szöveges és a multimédia tartalmakat is – okoznak, és ezért a *big data* kifejezés alá veszik mindazokat az eszközöket és eljárásokat is, amelyekkel létrehozni, kezelni, feldolgozni és tárolni lehet ezeket az óriási adathalmazokat. Olyan korszerű technológiákról van itt szó, mint például a hagyományos relációs adatmodellt meghaladó noSQL adatbázisrendszerek, a klaszterbe vagy gridbe kötött gépeken való párhuzamos adatfeldolgozásra kitalált MapReduce programozási elv, a természetes nyelvfeldolgozás, a gépi tanulás, vagy az újfajta vizualizációs megoldások.

A Digging into Data nevű kezdeményezés (*diggingintodata.org*), melyet több ország kutatásfejlesztési alapjai közösen indítottak, a humán és a társadalomtudományok területén szerveződő *big data* projekteket támogatja. Mivel a világunk egyre digitálisabbá válik és egyre nagyobbra nőnek az adatbázisok (a digitalizált könyvektől, újságoktól és zenéktől kezdve az olyan tranzakciós adatokig, mint a webes keresések naplói, a szenzorok mérései, vagy a mobil hálózatok cellainformációi), ezért új megoldásokra, újfajta kutatási infrastruktúrára van szükség a digitális adatok kereséséhez, elemzéséhez és megértéséhez.

Simon Hodson, a brit akadémiai szféra infokommunikációs infrastruktúráját működtető *JISC* szervezet kutatási igazgatója szerint számukra ezeken a területeken jelent kihívást a *big data* jelenség: webarchiválás, tanuláselemzés, felhasználói statisztikák és kutatási adatok. A JISC szponzorálja az *Oxford Internet Institute* egyik projektjét, amely a *British Library* kezelésében lévő webarchívum társadalomtudományi célú hasznosíthatóságát kívánja demonstrálni: az 1996 és 2010 között az *.uk* domén alá tartozó szerverekről begyűjtött mintegy 30 terabájtnyi weboldalból kivonják a hiperlink gráfokat, hogy statisztikai elemzéseket végezzenek rajtuk.

Az amerikai *National Science Foundation* és a *National Institutes of Health* BIGDATA nevű kutatási programjának felhívásában a *big data* fogalmát így határozták meg: nagy, változatos, komplex, longitudinális (vagyis idősorszerűen rögzített) és/vagy elosztott adathalmazok, melyek forrásai különféle eszközök, szenzorok, internetes tranzakciók, e-mail-, video- és kattintássorozatok, vagy egyéb digitális források. A program olyan tudományos és műszaki megoldásokat támogat, amelyekkel ezeket az adathalmazokat kezelni, elemezni, megjeleníteni lehet, és társadalmi, gazdasági, egészségügyi, életminőségi stb. célok érdekében hasznos információkat lehet kinyerni belőlük.

Bár maga a fogalom elég tisztázatlan még, abban egyetértés van a témában publikálók között, hogy az ún. *linked data* vagy *open data* kategóriába tartozó, vagyis a strukturált és többnyire nyilvánosan hozzáférhető adathalmazok is részei a *big data* világának, sőt: a szemantikus web alapját jelentő, automatizáltan értelmezhető és egymással kombinálható linkelt adatok ideális tesztkörnyezetet adhatnak a *big data* kutatások számára. Akár strukturált, akár strukturálatlan halmazokról van szó, hasonló műszaki kihívásokkal kell megbirkózni ezek rendszerezése, karbantartása, menedzselése, megőrzése, feltárása, vizualizációja, hozzáférhetővé tétele és használata során.

Linked data

A formalizált, strukturált és rendszerezett adatok a *big data* egyik típusát jelentik. A *linked data* és annak olyan speciális alkalmazásai, mint a linkelt kötött/korlátozott szótárak és a tudásszervezési rendszerek, szilárd szemantikus alapként szolgálhatnak a rendezetlen adatok osztályozásához és ábrázolásához. Felhasználhatók például automatikus vagy félautomata szövegelemzésekhez, tematikus metaadatoláshoz, és az adatok facettás, kategorizált vagy hierarchikus megjelenítéséhez.

Nagy tömegű szöveges adat értelmezése esetén olyan technikákra van szükség, mint amilyen a szemantikus szövegelemzés, a természetes nyelvi feldolgozás, az adatbányászat és az adatvizualizálás. A *W3 Konzorcium* által fejlesztett *SKOS* (*Simple Knowledge Organization System*) specifikáció egyfajta összekötő hídként szolgál a különféle tudásszervezési formák (tezauruszok, osztályozási rendszerek, tárgyszórendszerek, taxonómiák és folkszonómiák) valamint a *linked data* közösség között. A SKOS-alapú, linkelt, kötött szótárak megfelelő szemantikus keretet nyújtanak a *big data* halmazok elemzéséhez és ábrázolásához is. A SKOS-nak köszönhetően a különböző tezauruszok és egyéb szótárak leképezhetők egymásra és összekapcsolhatók, s így keresztül-kasul kereshetővé és böngészhetővé válnak a linkelt adatokat tartalmazó repozitóriumok, a nyílt archívumok, a digitális könyvtárak, valamint a különféle keresőrendszerek és szolgáltatások. Már vannak olyan nagy szótárak, amelyeket SKOS-formátumba is átkódoltak és linkelt adatforrásokként felhasználhatók. Ilyen például a környezettudományi GEMET,

az agráripari AGROVOC, az orvostudományi MESH, az interdiszciplináris szótárak építését segítő HIVE, a *Kongresszusi Könyvtár* LCSH tárgyszórendszere, a gazdaságtudományi STW Thesaurus for Economics és az oktatási területeken használatos ScOT.[1]

A *Bernard Vatant* és *Pierre-Yves Vandenbussche* által létrehozott Linked Open Vocabularies honlap (*lov.okfn.org*) a linkelt nyílt szótárak hasznos nyilvántartása. Ezek nemcsak a jól strukturált repozitóriumok és szemantikus webes alkalmazások számára érdekesek, hanem felhasználhatók rendszerezetlen szöveges adatok indexelésére, rendezésére és analizálásra is.

Olyan webes szolgáltatások is léteznek már, amelyek az RDF adatleíró keretrendszerre és a *linked data* szabványokra alapozva újszerű megoldásokat kínálnak a *big data* jelenséggel járó kihívásokra. A Zemanta (*zemanta.com*) böngészőkiegészítő például releváns címkéket, valamint külső forrásokból linkeket és képeket javasol automatikusan blogbejegyzések vagy cikkek, hírek írása közben. A Calais (*opencalais.com*) nevű szolgáltatással strukturálatlan szövegekből készíthetünk RDF formátumú kimenetet. Használható blogbejegyzések címkézéséhez, de akár múzeumi gyűjtemények kategorizálásához is. Jó példa a SKOS-alapú tezauruszokra a PoolParty (*poolparty.biz*), amely képes összekötni a különböző, linkelt adatokat tartalmazó repozitóriumokat, megkönnyítve így a keresést és az információgyűjtést.

Példák

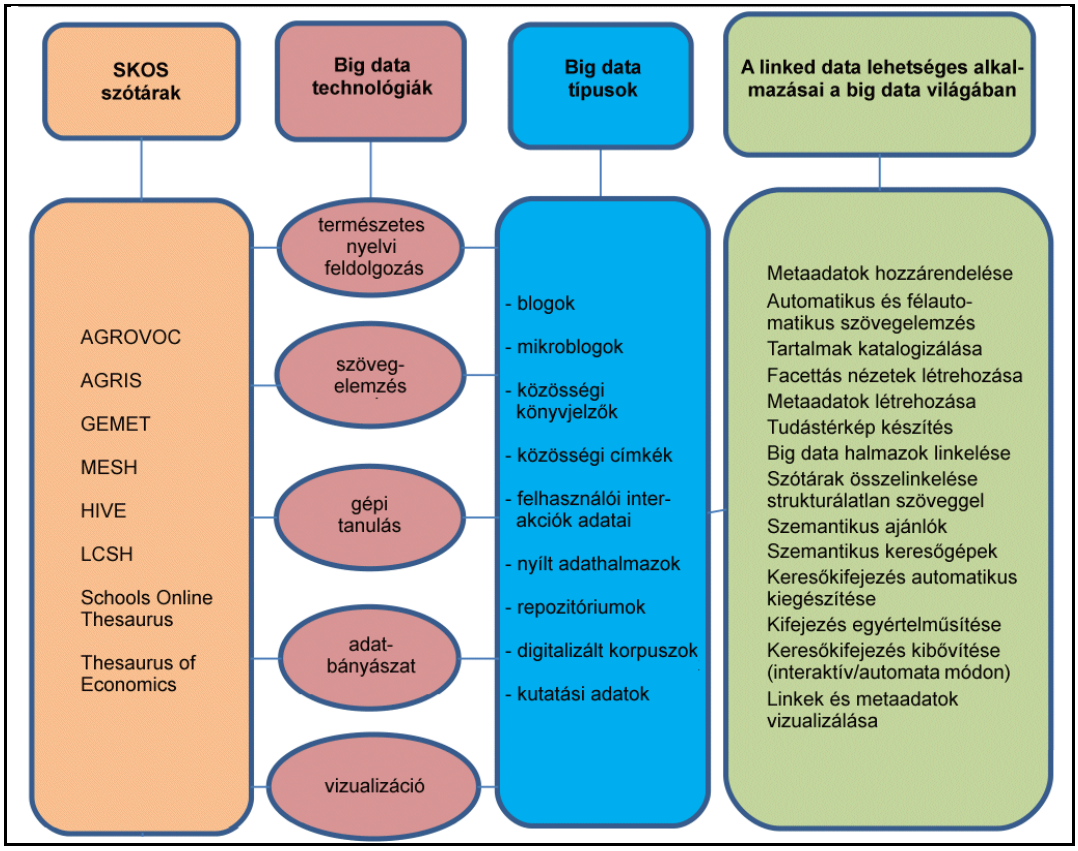
– Agráripari információk menedzselése és megosztása: A *FAO* (*Food and Agriculture Organization*) leíró metaadatok, tezauruszok, AGROVOC szótárak és ontológiák felhasználásával sokféle formátumban gyűjti, strukturálja és terjeszti az éhínség elleni harchoz szükséges táplálkozási, élelmiszerügyi és mezőgazdasági információkat.
– Személyre szabott egészségügy: Betegek adatait orvostudományi ontológiák segítségével nagy tömegben kielemezve és az adatok között mintázatokat keresve, az orvosok jobb döntéseket hozhatnak és személyre szabott betegségkockázati profilokat állíthatnak fel.
– Humán témák: A *DPLA* (*Digital Public Library of America*) és az Europeana egyaránt jó *big data* példa a művészetek és a bölcsésztudományok területén. Ezek a szervezetek egyrészt adatforrásként szolgálnak (az általuk biztosított API szolgáltatások révén), másrészt nagy adatfeldolgozó szervezetek is. Gondosan és folyamatosan fejlesztik az adatmodelljüket, és fontosnak tartják a szemantikailag gazdag metaadatokat, mert ezeknek köszönhetően a felhasználóik intuitívabb, intelligensebb módokon tudnak keresgélni a gyűjteményeikben.

Alkalmazási területek

A SKOS-alapú nyílt és linkelt szótárak többek között az alábbi területeken lehetnek hasznosak a *big data* típusú adathalmazok feldolgozása és használata során:
– Egyidejű keresés és böngészés nyílt adatrepozitóriumokban és *big data* forrásokban.
– Linkelt szótárak összekapcsolása strukturált, részben strukturált és strukturálatlan adatokkal.
– Nagy szöveghalmazok általános jellegű, illetve szakterületre specializált, természetes nyelvű feldolgozása.
– Tematikus metaadatok vizuális ábrázolása.
– Szemantikus keresőgépek és ajánlórendszerek kifejlesztése.
– Keresőfogalmak és -kifejezések automatikus kiegészítése.
– Keresőkérdések interaktív és automatikus kibővítése.
– Nézetek szerinti (facettás), illetve kategóriák alapján való böngészési lehetőség megteremtése.
– Digitális szövegek elemzése és feldolgozása humántudományi projektekben.

Néhány további lehetséges alkalmazás látható még az 1. ábrán.

[1] A *http://www.w3.org/2001/sw/wiki/SKOS/Datasets* oldalon felsorolt források közt ott találjuk az OSZK tezauruszait is. (A ref.)



1. ábra **Potenciális *linked data* alkalmazások a *big data* világban**

/SHIRI, Ali: Linked data meets big data: a knowledge organization systems perspective = Advances in Classification Research Online, 24. évf. 1. sz. 2014. p. 16-20/.

http://journals.lib.washington.edu/index.php/acro/article/view/14672

(Drótos László)

A speciális könyvtáros és a személyre szabott metaszolgáltatások: stratégiák a kutatók és a könyvtárosok újbóli összekapcsolására

A kutatók és a felsőoktatási könyvtárosok kapcsolata sokáig szoros volt. De a 90-es évek elején ez a kapcsolat lazult, ahogy az interneten át igénybe vehető véghasználói szolgáltatások erősödtek, ahogy a kutatók elkezdtek önkiszolgálókká válni. Másrészt a könyvtárosok figyelme is inkább a hallgatók, az online szolgáltatások felé fordult.

A cikk a speciális könyvtárosok[1] változó szerepét taglalja, különös tekintettel a kutatók és a könyvtárosok kapcsolatának megújítására: vagyis a speciális könyvtáros kívülállóból legyen bennfentes, a kutatócsoportok aktív tagja. A fő problémát a kutatók számára az információs túlterhelés jelenti mind a feladatok, mind a rendelkezésre álló eszközök vonatkozásában. A speciális könyvtárosok támogathatják a kutatókat feladataik megoldásában, a modern eszközök helyes megválasztásában.

A kutatók túl vannak terhelve, a speciális könyvtárosok által nyújtott személyre szabott metaszolgáltatások segíthetnek. Személyre szabott metaszolgáltatások alatt azt értve, hogy a speciális könyvtáros az egyéni kutatói gyakorlatokhoz illeszti