**Hubay Miklós**

# Adatgazdagítás, adatszolgáltatás és discovery hagyományos és szemantikus metaadat-környezetben

***A szemantikusweb-technológia közgyűjteményi alkalmazásának területén sokak számára a legfontosabb kérdés, hogy mihez lehet kezdeni azokkal az adatokkal, amelyeket fáradságos munkával, szótárak és elemkészletek körültekintő kiválasztása után MARC-ból valamilyen más adatformátumra konvertáltunk. Érdeklődő, türelmetlen, vagy egyszerűen gyakorlatias közgyűjteményi dolgozók egyre gyakrabban vetik fel: mi hasznát fogják látni ennek a használók és a könyvtárosok, mikor és hogyan jön el (ha eljön egyáltalán) az ígért Google-kánaán? Ténylegesen szükség van új adatformátum alkalmazására? Melyek azok a lépések, amelyeket már ma megtehetünk a haladás útján, és melyek azok, amelyek még csak vízióként lebegnek a szemünk előtt? Ebben a tanulmányban a linked datán alapuló adatszolgáltatás és információkeresés területén elért eredményeket vesszük szemügyre, s közben azt is megvizsgáljuk, hol áll napjainkban a hagyományos, illetve az új technológiákat előnyben részesítők igen szoros és izgalmas versenye.***

*Tárgyszavak:* szemantikus web; bibframe; bibliográfia; katalogizálás; szabályzat; adatszerkezet; adatszolgáltatás; információkeresés

### A linked data és az entitások

A könyvtári szakirodalom nagyon gyakran együtt tárgyalja az új informatikai technológiák használatát az entitásalapú katalogizálási paradigma térhódításával, pedig az entitások fogalmának megjelenése a forrásleírás területén csaknem egy évtizeddel megelőzi a szemantikus web koncepcióját. Hogy ennek mégsem lett komolyabb gyakorlati jelentősége a könyvtári munkában (kivéve a később tárgyalandó, utólagos FRBR-esítési törekvéseket), arra a közismert MARC-formátum alapstruktúrája a válasz. E sajátos szerkezet miatt a világ könyvtáraiban több mint ötven éve alkalmazott MARC egyszerűen nem alkalmas arra, hogy az entitásokat leíró fogalmi modelleknek (FRBR, újabban az LRM) megfelelően, egymástól elkülönítve láthassunk el metaadatokkal műveket, illetve kifejezési és megjelenési formákat. Mivel a MARC szerkezete a katalóguscédulán alapul, e három entitás adatait összesítve, egyetlen bibliográfiai rekordban teszi leírhatóvá, és csupán a példányadatok (item) elkülönült rögzítését biztosítja − önálló példányrekord felvételével −, amely jelentős hatással van az információkeresésre is. Az RDA forrásleírási szabályzat kidolgozásakor a szakemberek világosan látták, hogy az entitások egymástól elkülönülő metaadatolására építő rendelkezések csak akkor lesznek teljes mértékben és teljes eredményességgel alkalmazhatók, ha a MARC-ot valamilyen más hordozóformátummal helyettesítik[1], mivel annak struktúrájába nem illeszthető bele a számos entitás között szőtt, sűrű kapcsolati háló. A katalogizálási munkafolyamat teljes áthangolása az új paradigmára tehát egészen addig nem fog bekövetkezni, amíg az integrált gyűjteménykezelő rendszerek a jelenlegi metaadat-formátumban „gondolkodnak". Új hívójelek / indikátorok / almezők definiálásával, vagy régi elemek funkcióbővítésével némiképp (de csak ideiglenesen!) hozzáigazítható a MARC az RDA-ban foglaltakhoz, továbbá lehetőség nyílik arra is, hogy a későbbi konverziós folyamatot megkönnyítendő, egységes forrásazonosítókat (URI-kat) tároljunk a rekordokban szereplő entitásokhoz (személyekhez, testületekhez, fogalmakhoz stb.) kapcsolva.[2] Az FRBR első entitáscsoportjában szereplő négy entitás (angol nyelvű megnevezéseik rövidítése alapján: WEMI = Work, Expression, Manifestation Item) egymástól elkülönülő metaadatolása azonban továbbra sem megoldott. Ugyanakkor ez nem jelenti azt, hogy a MARC-rekordok *utólagos* vizsgálatával és feldolgozásával ne lehetne már ma is az

entitásalapú információkeresés előnyeit biztosítani a felhasználók számára.

### Az FRBR-szemlélet érvényesítése MARC-keretek között

Az ún. FRBR-izált (avagy förbörizált) OPAC vagy discovery-felület használatának tehát egyáltalán nem feltétele a szemantikus adatformátumok alkalmazása − az egyetlen követelmény a rekordok bizonyos adatmezőinek megléte, azaz a kielégítő adatgazdagság. A hagyományos, MARC-formátumban tárolt bibliográfiai rekordok FRBR-izált megjelenítésének technikáit egy 2015-ben született összefoglaló tanulmányban[3] három alapvető szempont mentén csoportosították a szerzők, amelyekből közelebbről az elsőt vesszük szemügyre. Ez a megkülönböztetés az *entitások kialakításának logikája mentén* történik: e folyamat végbemehet az eredeti rekordok csoportosításának, illetve az adatmezők gondosan kidolgozott szétválogatásának útján. Az első esetben a rekordokban előforduló egyes MARC-mezők, leggyakrabban természetesen a szerzők/közreműködők, illetve a címek egyezése alapján igyekeznek műveket és kifejezési formákat jelentő csoportokat alkotni. Ezt a módszert alkalmazza pl. az Ex Libris által fejlesztett Primo discovery-felület. A MARC-rekordból a szoftver első lépésben egy normalizált, ún. PNX-formátumot állít elő (Primo Normalized XML). A PNX-címkekészlet több eleme (az ún. FRBR-kulcs) a rekordokban eredetileg tárolt adattartalom normalizálási lépéseken átesett alakját jelöli, az FRBR-izálás pedig ezek egyezésére alapozva megy végbe.[4] Például a szerzői/közreműködői névalakok (MARC: 100/110/111 illetve 700/710/711) átalakított[5] változatait a Primo a PNX előállításakor a “k1” címkével látja el − címek esetében a 130-as, illetve a 240/242/245/246/247/740 mezők adattartalmával dolgozik.[6] Hasonló, összevetésen alapuló algoritmust dolgoztak ki az OCLC szakemberei, amely Work-Set Algorithm néven vált ismertté, és művek elkülönítését teszi lehetővé. E munkafolyamat alapdokumentuma 2005-ben, második változata 2009-ben készült el.[7]

FRBR-izálni azonban lehetséges más módon is – ez az ún. *szabályalapú* megközelítés –, ezt a módszert alkalmazta pl. a Kongresszusi Könyvtár (Library of Congress) a 2000-es évek második felében, amellyel egy FRBR Display Tool nevű megjelenítőeszközt „hajtottak meg”. A MARC-rekordokból ebben az esetben először MARCXML-t készítettek, amelyet különböző átalakítási szabályokat tartalmazó stíluslapok (ld. később) segítségével feleltettek meg a MODS (Metadata Object Description Scheme) metaadat-formátum elemeinek, valamint ilyen stíluslapok segítségével végezték az eredeti rekordok információtartalmának entitások mentén történő csoportosítását is.[8]

### MARC-rekordok gazdagítása azonosítókkal

Szintén nem igényel semmilyen különleges konverziós folyamatot, illetve új adatformátum használatát az ún. *adatgazdagítási* folyamat (data refining, data reconciliation). Ennek során más névterekből, elemkészletekből származó azonosítókat szúrunk be a gyűjteménykezelő rendszerben tárolt hagyományos rekordokba − mint láttuk, a MARC21-ben már szabványos helye van az entitásokhoz kapcsolódó, külső adatforrásokból származó egységes azonosítóknak. Napjainkban egyre több és több külföldi és hazai intézmény párosítja össze a rekordokban azonosítható entitásokat azok VIAF-ban, Geonames-ben, Getty-ben, vagy éppen a Wikipédiában, Wikidatában tárolt megfelelőivel, oly módon, hogy − lehetőségei függvényében szabványos vagy nem szabványos (de következetes!) módon − beszúrja azok URI-jait a MARC-rekord egy meghatározott helyére, ezzel deklarálja a rekordban, illetve a külső adatforrásban leírt entitás azonosságát, s nem utolsósorban ugrópontot biztosít a felhasználónak, hogy az még több információt szerezhessen a kérdéses entitásról, pl. egy személyről. Az ilyen párosítási folyamatok manuálisan − egyenkénti rákereséssel −, vagy informatikai eszközök (API-k és más szolgáltatások, OpenRefine stb.) igénybevételével csoportosan, igen nagy tételszámú halmazokon is végrehajthatók. A későbbi adatkonverziós folyamatot jelentős mértékben megkönnyíti és felgyorsítja, és nem utolsósorban pontosabbá teszi, ha már a kezdeti pillanatban rendelkezésre állnak az entitásokat azonosító, illetve egymástól megkülönböztető URI-k.[9]

Természetesen − ahogyan egy korábbi írásban már szó esett róla[10] − lehetőség van konverzió közben is adatgazdagítást végezni (ez az ún. *aktív konverzió)*, azaz a MARC-rekordokból elkülönített entitásokhoz más elemkészletekből vett, de ugyanazt az entitást leíró URI-kat a szemantikus formátumra történő átalakítással egyidőben beszúrni a gráfszerkezetbe. Tehát ha a MARC-rekord 100-as mezőjében Jókai Mórt találjuk mint szerzőt, az automatikus konverziós folyamat az előzetesen meghatározott helyeken keresést végez Jókai Mórra, s siker esetén begyűjti a talált URI-t, ezáltal hivatkozást hozva létre a két adathalmaz között.

Az efféle adatgazdagításnak a fentiek mellett fontos keresőoptimalizálási jelentősége is van: elérhetjük általa, hogy a gyűjteményi katalógusra hivatkozások mutassanak, ezáltal megnövekedjék a keresőmotor általi indexelés valószínűsége, s ezzel − számos egyéb feltétel teljesülése mellett − megjelenhessünk, vagy épp előkelőbb pozíciót szerezhessünk az online keresőszolgáltatások találati listáiban (erről bővebben ld. később, a *„Hogyan juthatunk a Google-ba?”* cím alatt).

**Az adatpublikáció legújabb módszere**

Az integrált gyűjteménykezelő rendszerek segítségével épített bibliográfiai és authority-adatbázisok különféle adatátviteli protokollokon alapuló, intézmények közötti megosztása már régóta bevett gyakorlat – sőt az online katalógusfelületek és discovery-szolgáltatások is ilyen protokollok segítségével tartanak kapcsolatot az intézmények által kezelt adatbázissal, követve annak bővülését, módosulásait. A MARC-rekordok átvitelét a Z39.50 teszi lehetővé már évtizedek óta, az újabb megoldások pedig már XML-ek mozgatásán alapulnak; ilyen adatszerkezetek átvitelére az OAI-, illetve az SRU/SRW-protokoll egyaránt alkalmas.

Amikor arról beszélünk, hogy egy intézmény által épített gyűjteményi adatbázist publikálunk, az csupán annyit jelent, hogy nem kizárólag a katalógusfelületen keresztül biztosítjuk annak használatát, hanem az adatok közvetlen eléréséhez szükséges informatikai paramétereket (Z39.50 szervercímet, az OAI-csatorna adatait stb.) is közzétesszük, azaz engedélyezzük, hogy bárki tetszőleges célra használhassa a bibliográfiai, vagy épp authority-állományunkat – mégpedig a neki megfelelő formátumban –, legyen az akár profi közgyűjteményi szolgáltatásfejlesztő, vagy éppen az adott intézményben felgyűlt tudásvagyon iránt érdeklődő, az informatikai technológiákban kisebb-nagyobb mértékben járatos magánszemély. Napjainkban a legtöbb webes szolgáltatás megelégszik azzal, ha bináris MARC-ban vagy MARCXML-ben jut adatokhoz, így tehát a legtöbb könyvtár és más közgyűjtemény ezeket a formátumokat részesíti előnyben adatai nyilvánosságra hozatalakor. A korszerűbb, innovatívabb megoldások azonban már valamilyen szemantikus, azaz értelmezett adatokat közlő formátum használatát igénylik; érdemes tehát elgondolkodni, hogy az intézmény szolgáltatási profilját ilyennel is kiegészítjük − ami tehát nem jelenti azt, hogy a hagyományos formátumok szolgáltatását be kellene szüntetni.

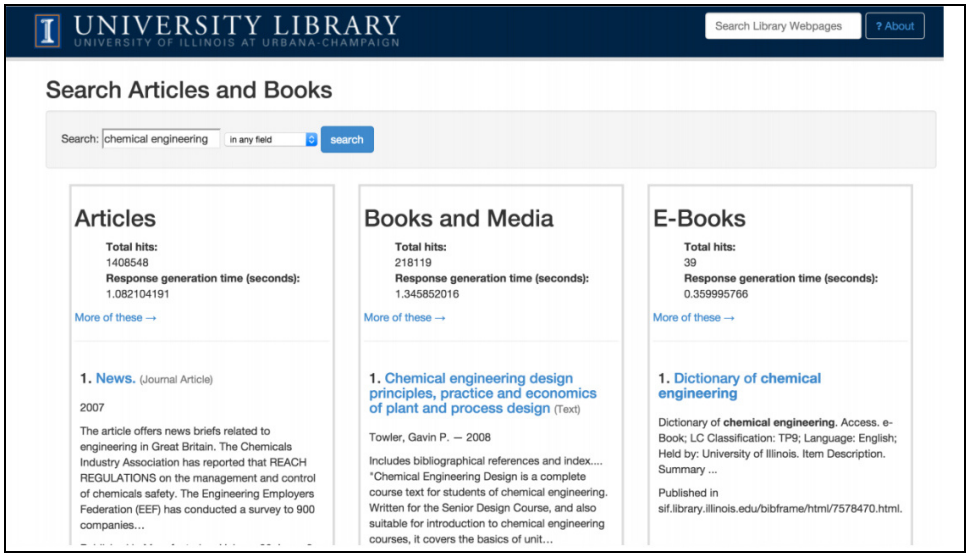
Amennyiben például MARC-rekordokat szeretnénk szemantikus formátumban közreadni − azaz nem elégszünk meg a MARCXML „adatértelmezésével”, amely csak annyit közöl a használóval, hogy hívójelet, almező- vagy indikátorértéket lát-e −, tehát magukat az adatelemeket is értelmezni kívánjuk az olyan felhasználók számára, akik nem ismerik a MARC-hívójeleket (szerző neve, megjelenés éve, sorozatcím stb.), valamilyen értelemgazdag formátumra kell konvertálnunk az adatainkat (ilyenek pl. az RDF/XML vagy a JSON-LD.) Ennek az egyik célravezető, igen elterjedt módja az XSLT-stíluslap alkalmazása. Mielőtt egy ilyet kialakítanánk, egy alapos, körültekintő mappinget kell készíteni, azaz olyan megfeleltetést, amely megmutatja, hogy melyik MARC-almezőt milyen relációnak kell megfeleltetni, és ez a reláció milyen szótárból származik. A MARC-almezők jó része megfeleltethető − egyebek mellett − a BIBFRAME-szótárban közölt relációkkal, más esetekben (tehát pl. a speciális adattartalmat hordozó 900-as mezők vonatkozásában) más szótárak alkalmazása is szükségessé válhat. A mapping, avagy más szavakkal, az intézmény linked data-alkalmazási profiljának kidolgozása után kezdődhet a konverziós segédeszköznek, azaz magának a stíluslapnak az elkészítése. Az eszköz működési elve, hogy a kiinduló XML-állomány faszerkezetében azonos elemeket keres (tehát például olyan „c” subfield-értékeket, amelyek 245-ös „datafield” szülőelem gyermekei), majd ezek mindegyikét ellátja az előzetesen megjelölt relációt azonosító URI-val, valamint elhelyezi az adatelemet az adott szótárban definiált modellbe − azaz megmondja például, hogy az adatelem művet, kifejezési formát vagy példányt, netán közreműködőt, vagy bármely más entitást ír-e le. Így áll össze a háromelemű állítás (triplet), amelynek tehát első eleme lesz a leírt entitás megnevezése, ezt követi a reláció megnevezése, végül az XML-ből kiolvasható, MARC-almezőkben tárolt értékek. A keletkező állításhalmazt a konverzió lefutását − és az esetleges adathibák, adatvesztések kijavítását követően − valamilyen adatbázisba szervezzük (relációs vagy gráfadatbázis is elképzelhető), és ezen adatbázis elérhetőségét (például az ún. SPARQL-endpointját) ugyanúgy közzétehetjük, ahogyan a Z- vagy OAI-protokollok esetében láttuk. Szabad a vásár!

**A szemantikus formátumú adatok felhasználási lehetőségei**

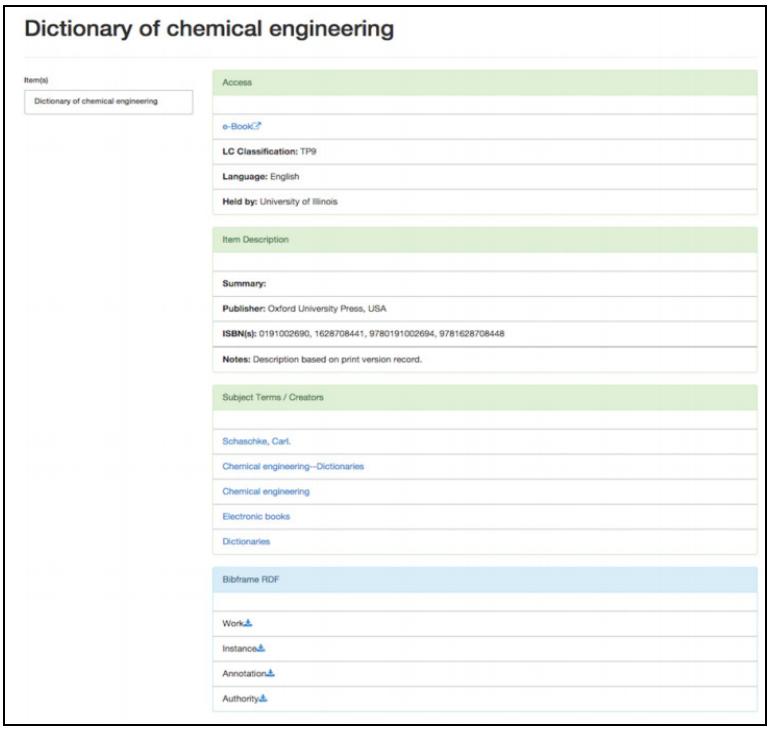
A konverzió egyik lehetséges kimenete az ún. RDF/XML-állomány, amelynek általános szerkezete számos formázási lehetőséget kínál: ezeket

ugyancsak stíluslapok írják le, amelyek a konverzionál leírt módon működnek, a megtalált azonos gyermekelemeken azonban nem átalakítást, hanem minden esetben azonos megjelenítési utasításokat hajtanak végre. Ilyen, RDF/XML-állományokon és stíluslapon alapuló discovery-eszköz működött az Illinois-i Egyetem könyvtárának honlapján (*1. ábra*). Az intézmény egy ún. Bento-stílusú keresőszolgáltatást üzemeltetett, amelynek lényege, hogy a keresett kifejezést egyidejűleg több adatbázisban kutatta fel, majd az eredményeket egymástól elkülönítve, többnyire oszlopokban közölte.

A harmadik oszlop (e-könyvek) találatai olyan RDF/XML-fájlokból származtak, amelyeket még a BIBFRAME 1.0 adatmodellje szerint konvertáltak, a stíluslappal szabályozott megjelenítés pedig a 2. ábrán látható kinézettel ruházta fel az XML-állományt[11].




1. ábra **Az Illinois-i Egyetem könyvtárának honlapja**




2.ábra **XML állomány**

Lehetőség van arra is, hogy a hagyományos OPAC-felületet összekapcsoljuk valamilyen triplestore-ral, például a Virtuosoval − magától értetődően ehhez már szükség van az adatok szemantikus formátumban tárolt változatára is − s a tripletekben tárolt, külső adathalmazok felé mutató hivatkozásokat a MARC-rekordhoz hozzárendelve jelenítsük meg a használók számára; ezt láthatjuk a Magyar Nemzeti Múzeum központi könyvtárának katalógusfelületén. Ebben az esetben tehát a külső névterekre, gyűjteményi katalógusokra mutató linkek nem előzetes, MARC-ban végrehajtott adatgazdagítás eredményei, sőt ezek az adatok utólag sem lettek a rekordok részei, hanem két külön forrásból érkeznek, egy relációs adatbázisból, valamint a triplestore-ból (*3. ábra*).

*Rob Sanderson*, a Getty munkatársa ugyanakkor éles kritikát fogalmazott meg az RDF/XML-formátum felhasználásával, SPARQL-végpontokon keresztül publikált adathalmazokkal szemben. Állítása szerint az adatokat igénybevevő partner – webes szolgáltatásfejlesztő vagy egyéb informatikai szakember – számára jóval egyszerűbb és jövedelmezőbb, ha valamilyen alkalmazásprogramozási interfészen (API) keresztül, az általa elküldött kérésekre érkező válaszok formájában fér hozzá az adatszolgáltató adataihoz.[12] Sőt − teszi hozzá − az összekapcsolt adatok átadását lehetővé tévő szerializációs szintaxisok közül a JSON-LD-t érdemes választani az RDF/XML helyett; azt a módosított JSON-struktúrát, amely az egyes objektumtulajdonságokat megfeleltethetővé teszi egy tetszőlegesen választott ontológia formalizált relációival.[13]

## Könyvtárasított Google − vagy Google-szerű könyvtár?

A szemantikus web fejlődésének kezdeti időszakában egyre-másra olvashattuk a könyvtári katalógus, a könyvtári feldolgozó felület hatalmas átalakulásának ígéretét, illetve hogy az új technológia forradalmi változásokat hoz majd a használók információkeresésében, s hogy a könyvtári-közgyűjteményi tudásvagyonok, amelyek mindeddig elszigetelt silókban léteztek, egy csapásra kiszabadulnak és a keresőszolgáltatások találati listáinak integráns részét képezik majd. Az idő múlásával már jóval tisztábban, reálisabban látjuk azokat a lehetőségeket, amelyek a szemantikus formátumban történő adatközzététel, a keresőoptimalizálás, valamint a discovery-felületek szemantikus adatformátumokra alapozott átalakítása területén állnak előttünk. A nyilvános SPARQL-endpointokon, valamint az API-n keresztüli publikációról már esett szó, a tanulmány utolsó részében a másik két területet vesszük szemügyre.

### *Hogyan juthatunk a Google-ba?*

A közgyűjteményi keresőoptimalizálás építőköveit és feladatait *Horváth Ádám*, a Magyar Nemzeti Múzeum könyvtárának vezetője gyűjtötte csokorba egy 2018-ban elhangzott előadásában. Az itt elhangzottak alapján a teljes katalógus indexeltetésének egyik legfontosabb művelete az ún. linkfeloldás, azaz amikor az online katalógusfelület által az egyes leírásokhoz tartozó − esetenként igen


3 ábra **Bibliográfiai rekord teljes nézetének részlete, a személynévhez kapcsolódó más információforrások bemutatásával (képernyőkép, saját szerkesztés)**

hosszú, változó részeket tartalmazó − URL-eket valamilyen állandó szerkezetű, ún. tetszetős (cool) URI-nak feleltetjük meg, s a webszerver ún. újraírási szabályainak *(rewriting rules)* konfigurálásával biztosítjuk a rövidített és az eredeti webcímek közötti kapcsolatot. Ezt követően a rövidített URI-k teljes jegyzékét át kell adni az indexelést végző keresőszolgáltatásnak: ezt egy ún. oldaltérkép segítségével tesszük meg, ez az állomány XML-formátumban tartalmazza a gyűjteményben hozzáférhető valamennyi bibliográfiai leírás (esetenként több százezer) saját weboldalának címét.

Egy másik fontos feladat az oldalak tartalmának érthetővé tétele a keresőszolgáltatás számára: egy megfelelően konfigurált megjelenítési sablon segítségével a weboldal kódjába olyan jelölőket helyezünk el, amelyek a MARC-rekordokból származó értékekhez (pl. „598 p.") egy-egy meghatározott relációtípust rendelnek (pl. numberOfPages), azaz értelmezik a rekordban tárolt adatelemet. Korábban már láttuk, hogy egy bizonyos relációt több szótár is leírhat; amikor azonban a keresőoptimalizálás és nem az adatok konverziója a cél, csak egy választásunk van: a nagy szolgáltatók (Google, Yahoo, Yandex, Bing) által épített ontológiát, a schema.org-ot kell használnunk, s az ebben rögzített relációkkal kell megfeleltetnünk a saját adatainkat. Számos OPAC- és discovery-felület már eleve rendelkezik a MARC-mezők és a schema.org-relációk megfeleltetésén alapuló sablonnal; függetlenül attól, hogy milyen leírást jelenítünk meg, az azonos MARC-mezőben lévő adattartalmak minden esetben ugyanazt a relációcímkét kapják majd.[14]

### *Hogyan válhatunk mi magunk Google-lá?*

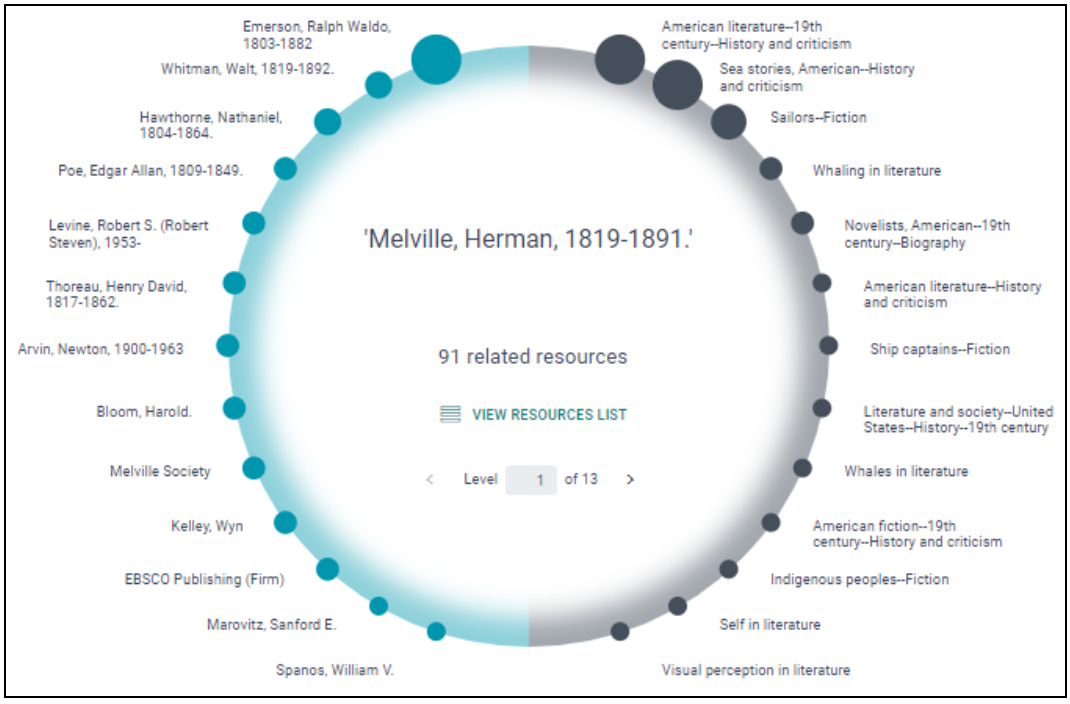
Az Innovative informatikai rendszereket fejlesztő cég 2019 áprilisában vizsgálatot végzett a felsőoktatási könyvtárak gyűjteményeit használó oktatók, kutatók körében, amely tulajdonképpen csak megerősítette a több éve köztudomású tényt: a közgyűjtemények által szolgáltatott adatokat a használók igen megbízhatónak gondolják, azonban az információkereső felületek használatát túlságosan összetettnek, bonyolultnak látják − ezért fordulnak a nagy, webes keresőszolgáltatásokhoz, jóllehet, az azok által szolgáltatott információkat már jóval kevésbé vélik hitelesnek. Logikusnak és érthetőnek tűnik a következtetés, hogy a felhasználószám csökkenésének megállításához vagy visszafordításához a könyvtári tudásvagyont minél nagyobb arányban reprezentálni kell például a Google találatai között, ugyanakkor valós alternatíva, hogy *mi magunk váljunk Google-lé*: azaz egyszerűsítsük le, közelítsük a webes keresőkhöz a közgyűjtemény keresőfelületét, hogy az minél egyszerűbben, informatívabban biztosítsa a bibliográfiai források megtalálhatóságát. A korszerű discovery-szolgáltatásokban ezért az egypontos keresés mellett – ismét csak a Google-hoz hasonlóan − tudáspaneleket jelenítenek meg a találati listákhoz kapcsolva, vagy épp olyan, újszerű keresési lehetőségeket teremtenek a használók számára, amelyeknek elengedhetetlen feltétele az entitások és tripletek mentén felépülő adatbázis.

Az *Inspire* discovery-felületet 2019 áprilisában mutatta be az Innovative Solutions, amelynek első implementációja − a pennsylvaniai Cairn Egyetem könyvtárában − szabadon hozzáférhető, kipróbálható.[15] Működésében nagy szerepet kap a keresés eredményéül kapott információk vizualizációja, s ez a vizualizáció már nem MARC-rekordokra, hanem a BIBFRAME-szótár használatával tripletekre konvertált adatállományra épül. A felületet megnyitva egyetlen keresőmező áll rendelkezésünkre, amelynek segítségével szerzőkre, címekre és fogalmakra, azaz tárgyszavakra kereshetünk. A találati halmaz három részre tagolódik (*4. ábra*): az első rész (Resources) a keresett szerzőhöz kapcsolódó források listáját tartalmazza − amelyek létrejöttében az illető részt vállalt, vagy amelyeket róla írtak. A második rész (Related People) a keresett szerzőhöz kapcsolódó más személyeket és testületeket, míg a harmadik szakasz (Topics) a vele összefüggésbe hozható LCSH-deszkriptorokat tartalmaz.

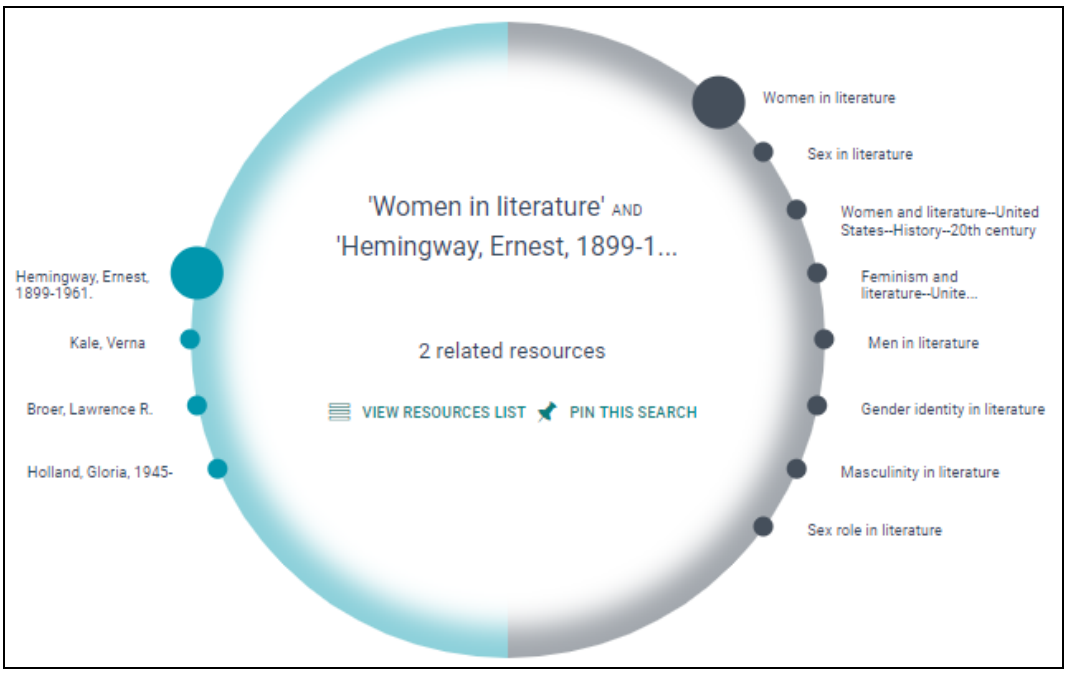
A keresett szerzőhöz kapcsolódó más szerzők, valamint deszkriptorok a listás megjelenítés mellett egy könnyen áttekinthető vizuális eszközre, az ún. *context wheel-re* vetíthetők, így pl. *Herman Melville* kapcsolati hálója az 5. ábrán látható formát ölti, ahol az egyes csomópontokra kattintva további kapcsolatokat tekinthetünk meg.

A visszakeresési szempontok ugyanakkor természetesen kombinálhatók is, az ún. Workbench eszköz segítségével, majd a végeredmény ugyancsak megjeleníthető (*6. ábra*).


4. ábra **A tárgyszavakra történő keresés eredményhalmazát a megtalált források borítóképei teszik élénkebbé**




5. ábra **Herman Melville kapcsolati hálója**




6. ábra **Visszakeresési lehetőségek**

### *A Stanford Egyetem discovery-fejlesztései*

A Stanford Egyetemen évek óta nagy volumenű fejlesztési projekt zajlik, amely a linked data-technológia vívmányainak alkalmazására irányul. Mivel a kapcsoltadat-technológia a felhasználók (és nem utolsósorban a források felhasználhatóságáról határozó döntéshozók!) szempontjából legígéretesebb, legjobban bemutatható eredményeket az információk visszakeresési metódusának megújításában hozhatja, a Linked Data for Production (LD4P) projekt második fázisának kiemelt területe egy, az összekapcsolt adatokra épülő discovery-felület kifejlesztése, egészen pontosan egy már létező open-source megoldás továbbgondolása. A projektben várhatóan az alábbi fejlesztések valósulnak meg:

- tudáspanel a keresési eredmények mellett például személynevekre és földrajzi helyekre történő keresés esetében, amely a Wikidata, illetve a Who's On First földrajzi névtér kapcsolódó adatait emeli be a keresőfelületre;
- gazdagított adatokra alapuló, kibővített fazettás keresési lehetőségek (pl. olyan szerzők műveinek megtekintése a katalógusfelületen, akik a Stanford Egyetemen végeztek);[16]
- alternatív keresési javaslatok közlése a "nincs találat"-oldalon, kapcsolódó értelmű kifejezések felajánlása a keresőkérdés gépelése közben.[17]

### A Wiki-univerzum

Igen sokat tanulhatunk a megjelenítés lehetőségeiről a Wikidata adatbázisát vizsgálva. A Wikidata a Wikipédia-szócikkeket kiszolgáló, de attól teljesen függetlenül épülő tudásbázis, amely a leírt entitásokra vonatkozó tudást nem teljes szövegű szócikkek, hanem strukturált adatok formájában tárolja. Az entitás (pl. egy személy) adatait tartalmazó beviteli űrlapot is maga a felhasználó, adatrögzítő állíthatja össze, a megjelenítendő ismérvek (pl. a személy neve, születési-halálozási helye, foglalkozása, elnyert díjai stb.) listáját egy nagyméretű szókészletből kiválasztva. Az így előálló adatlap − a szakterminológia szerint *Wikidata-elem* − több komponensen keresztül beemelhető a különféle nyelvű Wikipédia-szócikkekbe: így működnek az ún. infoboxok, és a külső azonosítókat tartalmazó Nemzetközi katalógusok sablon is a Wikidatában rögzített tudást veszi alapul.

Az adatbázis a query.wikidata.org oldalon, a SPARQL-lekérdező nyelv ismeretében könnyedén kereshető, de van lehetőség grafikus keresésszerkesztő használatára is. A lekérdezés eredményeként visszakapott információk természetétől függően az adathalmaz több formátumban (táblázatosan, csomópontok és élek segítségével, fastruktúrában, vagy éppen térképre vetítve) is megjeleníthető.

## Záró gondolatok

A szemantikus web – ez a korábban kissé misztikusan hangzó kifejezés – egyre nagyobb gyakorlati jelentőséggel bír a közgyűjtemények adatrögzítési és adatszolgáltatási feladatainak területén. Gombamódra szaporodnak a különféle entitástípusok köré csoportosuló szemantikus elemkészletek, SKOS-sal konvertált különféle tudásszervező rendszerek, valamint a rendelkezésünkre áll ezerféle, a közgyűjteményi szakma területén előforduló relációkat összegző, azokat valamilyen adatmodell szerint csoportosító szótár (ontológia). Az adatok transzformációjához szükséges technikai feltételek tehát adottak, a szaporodó implementációk és új megoldások azt mutatják, hogy a „hogyan”-ra adandó válasz sokak számára egyre egyértelműbbé válik. A nagyobb problémát a „miért” jelenti – ugyanakkor bízunk abban, hogy erre a kérdésre jelen írásban foglaltak áttekintése után mind több és több közgyűjtemény tud nem egy, hanem máris két választ adni. Először: könyvtárspecifikus csereformátumok helyett sokkal szélesebb kör által értelmezhető adatszerkezetekkel lehetővé tesszük, hogy az általunk kezelt adatok kreatív, innovatív szoftveres megoldások – például részterületi aggregációs szolgáltatások – segítségével minél több használóhoz jussanak el. S másodszor: az információkeresési folyamat immanens logikájára való tudatos építkezéssel – amelyet megtalálhatunk az FRBR funkcionális szemléletében – vélhetően sokkal könnyebben kezelhető, felfedezésre inspiráló felületet biztosíthatunk használóink számára, s így a közgyűjtemények használata számukra porszagú kötelességből végre valódi élménnyé válik.

## Irodalom


[1] Report and Recommendations of the U.S. RDA Test Coordinating Committee [elektronikus dok.] http://www.loc.gov/bibliographic-future/rda/source/rdatesting-finalreport-20june2011.pdf [hozzáférés: 2015.10.12.] p. 8

[2] A HUNMARC-ban a fent említett, az RDA implementálását legalább részben segítő módosítások nem jelentek meg. Az új katalogizálási szabályzat bevezetéséről hozott döntés ezért logikusan a MARC21-re való áttérés szükségességének kérdését is felveti.

[3] DECOURSELLE, Joffrey – DUCHATEAU, Fabien – LUMINEAU, Nicholas: A Survey of FRBRization Techniques = TPDL 2015. Lecture Notes in Computer Science, vol 9316. Cham, Springer, 2015., p.185-196.

[4] HAÁSZ Antal: A Primo használata a Magyar Tudományos Akadémia Könyvtár és Információs Központban = Hagyományok és kihívások V. Múlt és jövő. Országos Könyvtárszakmai Nap, 2016. Budapest, ELTE, 2017., p. 167-178.

[5] Egyes karakterek törlése, számos írásjel szóközzel helyettesítése, a diakritikus jelek kezelése, valamint kisbetűs írásmódra történő alakítás zajlik az átalakítás közben.

[6] The FRBR Vector [elektronikus dok.] https://knowledge.exlibrisgroup.com/Primo/Product_Documentation/Primo/Technical_Guide/040FRBRization/010The_FRBR_Vector [hozzáférés: 2020.04.08.]

[7] HICKEY, Thomas B. – TOVES, Jenny: FRBR Work-Set Algorithm : version 2.0 [elektronikus dok.] https://www.oclc.org/content/dam/research/activities/frbralgorithm/2009-08.pdf [hozzáférés: 2020.04.08.]

[8] http://jodischneider.com/pubs/2008may_frbr.html

[9] Eltérő MARC-rekordszerkezetre írt konverter konfigurációjánál nagyon körültekintően kell eljárni, hogy a kimenetben ne történjen adatvesztés, azaz hogy a konverziós folyamat minden adatelemet ott keressen, ahol az a rekordokban ténylegesen előfordul. Ellenkező esetben több triplet nem kap értéket, vagy még rosszabb, téves, az entitások azonosságát nem egyértelműsítő azonosítók generálódnak, és a halmazon futtatott keresések pontatlan eredményeket hoznak.

[10] BAY Miklós: Tények, mítoszok és lehetőségek a szemantikus web világában = Könyvtári Figyelő, 65. évf. 2. sz. 2019. p. 245-253.

[11] N, Quiang – HAHN, Jim – CROLL, Gretchen: BIBFRAME Transformation for Enhanced Discovery = Library Resources & Technical Services, 60. évf. 4. sz. 2016. p. 223-235.

[12] SANDERSON, Rob: Shout it out: LOUD [elektronikus dok.] https://www.slideshare.net/azaroth42/europeanatech-keynote-shout-it-out-loud [hozzáférés: 2020.04.08.]

[13] JSON-LD [elektronikus dok.] https://en.wikipedia.org/wiki/JSON-LD [hozzáférés: 2020.04.08.]

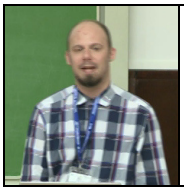
[14] HORVÁTH Ádám: Online katalógusok felhozása a felszíni webre [videofelvétel] https://kifu.videotorium.hu/hu/recordings/24689/online-katalogusok-felhozasa-a-felszini-webre [hozzáférés: 2020.04.08.]

[15] Elérhető a https://pbibu.na.innovativeinspire.com/ címen.

[16] LD4P2 Knowledge Panels Work Cycle Demo [video-felvétel] URL: https://www.youtube.com/watch?v=XjBFmBE3Pck [letöltés: 2020.04.08.]

[17] Discovery (WP4) [elektronikus dok.] https://wiki.lyrasis.org/pages/viewpage.action?pageId=101783940 [hozzáférés: 2020.04.08.]

Beérkezett: 2020. IV. 9-én.



***Hubay Miklós***
Petőfi Irodalmi Múzeum
múzeumi szakinformatikus.
E-mail: hubaym@pim.hu