Tudásgráfok és többnyelvű visszakereső felületek

A BIBFRAME mellett a kapcsoltadat-univerzum másik kiemelkedő, lassan megkerülhetetlenné váló pontja a Wikidata adatbázis, amely egyre inkább helyet kér és kap a bibliográfiai számbavétel és az authority kontroll területén.

A Wiki-szolgáltatási univerzum részeként 2012-ben meginduló *Wikidata* a szemantikus, vagyis az adatok értelmezett kapcsolataiból épülő világháló különleges szereplője. Az adatbázis, amely bárki számára térítésmentesen igénybe vehető (akárcsak a Wiki-univerzum többi tagja), a Wikipédiához hasonlóan tudást tárol, alapegységei azonban nem szócikkek, hanem minden esetben saját URI-val és egyedi Wikidata-azonosítóval rendelkező elemek, amelyek egymást követő, rövid és tömör kijelentésekből, metaadat-közlésekből épülnek fel. Ezek az adatok számos Wikimedia-szolgáltatásba, így a Wikipédiába is átemelhetők, így a Wikidata voltaképpen a Wiki-univerzum központi adattáraként funkcionál. Ugyanakkor külső szolgáltatásokba is képes adatokat továbbítani, amelyeket egyebek mellett a VIAF, illetve a Google tudásgráfja (Google Knowledge Graph) is felhasznál – ez utóbbinak markáns jele a személyek, testületek keresésénél a találati képernyő jobb oldalán megjelenő tudáspanel, infobox.

A Bibliographic Control konferencia idején, 2021 februárjában az adatbázis 91 millió elemet tárolt – hét-nyolc hónappal későbbre ez a szám már csaknem 96 millióra emelkedett. Ezek jelentős hányada (februári adatok alapján a teljes mennyiség körülbelül egyharmada) különféle tudományos publikációk adatait rögzíti, ezért a Wikidatát tudománymetriai szolgáltatások működtetésére is felhasználják, ilyen például a Scholia, amely a Wikidatából aktuálisan, a keresés pillanatában lekérdezett adatmennyiséget – szerzők, testületek, helyszínek, illetve publikációk legfontosabb adatait és azok sokszínű kapcsolatrendszerét – igen sok összefüggésben, számos megjelenítésmóddal (lista, grafikon, térkép) képes bemutatni a használónak.[51] Az elemek további egytizede személyek leírását adja.

[51] A szolgáltatás hozzáférhető a https://scholia.toolforge.org/ címen.

Egy elem mindig tartalmaz legalább egy címkét, tehát az elem megnevezését egy adott nyelven. Ilyen címkék felvételére korlátlan számú nyelvi változatban van lehetőség: a többnyelvűség a visszakeresés folyamatát támogatja, mivel az egyes visszakereső, megjelenítő felületek automatikusan válogathatnak a rögzített címkeváltozatok közül a felhasználó által megadott nyelvi beállításoknak megfelelően. A címkék mellett lehetőség van rövid értelmező leírást adni, ez az adott megnevezés egyértelműsítésére, az azonos alakú névformák egymástól való megkülönböztetésére szolgál, és magában a Wikidatában történő keresést is megkönnyíti, mivel rögzítése esetén részét képezi a keresőmezőbe gépelés közben érkező keresési javaslatoknak. Végül minden nyelvi címkeváltozathoz tartozhatnak névvariánsok is.

Ezt követi az elem tulajdonságainak (properties) megadása, amelyek közül kiemelkednek az úgynevezett azonosító típusú tulajdonságok (identifiers). Ezeken keresztül a Wikidatában előforduló entitás-előfordulás leírása összekapcsolható az azonos tartalmú, de más forrásban található adathalmazokkal. A hagyományos, közgyűjteményi authority állományok rekordjaira mutató azonosítók mellett szótárak, enciklopédiák, biográfiák és más források (ISNI-adatbázis, Discogs, Europeana, IMDb vagy akár az Instagram, Twitter, YouTube stb.) felé is létrehozhatók hivatkozások, így a keresett entitás-előfordulásról átfogó információ szerezhető; bizonyos elemek leírása akár 90–100 külső hivatkozást is tartalmaz.

A Wikidatára irányuló figyelem az utóbbi években a hagyományos technológiákon és módszertanon alapuló bibliográfiai forrásleírás krízise miatt erőteljesen megnőtt. Ez a krízis jelenti egyrészt az elektronikus bibliográfiai források nehézkes kezelését, ugyanakkor jelenti az RDA megjelenésével egyre nagyobb szerephez jutó entitásalapú feldolgozással kapcsolatban felmerülő kompatibilitási problémákat is – az ilyen leírások támogatására a hagyományos csereformátum nem, vagy csak nagyon korlátozottan alkalmas. A Wikidatában tárolt forrásleírások azonban már korszerűbb alapokra, a kapcsoltadat-technológiára, vagyis az adatok összekapcsolódására, az úgynevezett tripletekre és a belőlük épülő gráfokra épülnek, amelyek segítségével – legalábbis részben – megoldható az egyes entitásokra vonatkozó metaadatközlések elkülönítése. A külső azonosítók társításával, az egyes adatelemek minősítésével (kitüntetett, elavult stb.), valamint az adatelemek forrásainak rögzítésével átfogó, informatív leírások készíthetők az egyes forrásokról.[52]

[52] A leíráshoz felhasználható metaadatelemek, illetve külső azonosítók forrástípusok szerint csoportosított listáját lásd a https://www.wikidata.org/wiki/Template:Bibliographic_properties című oldalon, az egyes relációk entitások szerinti, alapszintű csoportosítását pedig a https://www.wikidata.org/wiki/Wikidata:WikiProject_Books című oldal közli.

A Wikidata adatbázis használata a bibliográfiai számbavétel szempontjából elméleti, illetve gyakorlati síkon egyaránt értékelhető. Az előadók szerint a számbavételt, illetve az azt támogató authority kontroll folyamatát úgy kell kezelni, mint az általános emberi tudás rögzítésének egy szeletét, s a Wikidata felületének újszerű megközelítése révén megvalósulhat ez az összeolvadás. A közösségi, megosztott/megosztható munkavégzésnek köszönhetően az identitáskezelés (identity management), azaz az entitás-előfordulások adatainak gondozása összhangban van a nemzetközi katalogizálási alapelvekben leírt két követelménnyel, amelyek a felhasználó kényelmére, illetve az általános szóhasználatra vonatkoznak.

Gyakorlati szempontból ki kell emelni a Wikidatában elkészíthető forrásleírások maximális testreszabhatóságát, az alkalmazási profilok elkészítésének tekintetében tapasztalható igen nagy rugalmasságot. Az előadók úgy vélik, az alulról felfelé történő építkezési metódusnak és a közösségi munkavégzésnek hála, megvalósulhatnak az UBC-program céljai, amelyeket a nemzeti bibliográfiai ügynökségek hosszú évtizedek óta tartó együttműködésével sem sikerült teljes mértékben megvalósítani.

A Wikidata, illetve az azt működtető Wikibase szoftver használata, valamint az egymással összekapcsolódó adatok előállítása, kezelése és megjelenítése az OCLC tevékenységében is fontos szerepet kapott. *John Chapman* számolt be arról, hogy a kapcsoltadat-technológiával összefüggő kísérletek, projektek már 2009 óta jelen vannak az OCLC tevékenységében, elég ha a FAST vagy a VIAF adathalmazára, vagy éppen a WorldCat találataihoz kapcsolódó,[53] a schema.org szótár segítségével megjelenített állításokra gondolunk. A 2010-es évek első felében *EntityJS* néven megjelenítő eszközt fejlesztettek ki, ahol a keresőkifejezésekre érkező találatok entitások – személyek, testületek, fogalmak, helyek, események, illetve művek – alapján válogatva jelentek meg. Egy entitás-előfordulást, például egy konkrét személyt kiválasztva a hozzá kapcsolódó más entitás-előfordulások listáját kapta a felhasználó. A szolgáltatásplatformot adatminőség-javító segédeszköz is kiegészítette, amely az OpenRefine-hoz hasonló adattisztítási lehetőségeket biztosított a felhasználóknak (CONTENTdm Metadata Refinery).

A Wiki-univerzum workflowba illesztésére már 2017–18-ban, a *Project Passage* néven ismert kezdeményezésben is sor került, amelynek célja közgyűjteményi kapcsoltadat-adathalmazok előállítása, illetve ezek külső azonosítókkal történő, széles körű adatgazdagítása volt a Wikibase felhasználásával. Az idén befejeződő, két évig futó Entity Management Infrastructure program célja az ekkor megalkotott adathalmaz további bővítése, és olyan szolgáltatássá terebélyesítése, mely más

[53] Az OCLC közlése alapján ez a szolgáltatás 2021 februárjáig állt fenn: https://help.oclc.org/Discovery_and_Reference/WorldCat-org/Troubleshooting/What_happened_to_the_linked_data_that_previously_displayed_on_records_in_WorldCat.org.

termékeket is támogathat a jövőben. A projekt keretében közösségi tudásgráf épül, amely közgyűjteményi authority állományok, a WorldCatben tárolt műinformációk és ellenőrzött szótárak adattartalmát gyűjti egybe, kiegészítve azok forrásaira vonatkozó adatokkal. A gráfot API-kon és lekérdezési végpontokon keresztül szolgáltatják.

Egy tudásgráf megalkotásának és karbantartásának folyamata alapvetően négy hívószó köré csoportosítható: *tudásalkotás,* amely az adatforrások integrálását, illetve a tripletek megalkotását (semantic lifting), a tudás kinyerését (extracting) jelenti; *a tudástárolás,* amely fázisban az adathalmazt valamilyen szemantikus repozitóriumban (gráfadatbázis, triplestore) helyezik el; *tudáskarbantartás,* amely az adatok helyessége és teljessége érdekében végrehajtott tevékenységeket összegzi, hiszen egy ilyen gráf soha nem egy statikus produktum; végül a tudásbevetés, amikor a gráf információtartalma alkalmazásokban hasznosul.

Az OCLC-n belül épített tudásgráf, ahogy már említettem, külső adatforrások (VIAF, WorldCat, sőt a Wikidata) anyagának importálásával és konverziójával bővül. A kinyert tudást nTriples-szintaxisban tárolják, majd entitásfelismerési–adatgazdagítási folyamatnak vetik alá, melynek során összekötik a forrásadathalmazokban található azonos entitás-előfordulások adatait. A kibővített tudásgráf már JSON-szintaxisban áll rendelkezésre, és fontos eleme az adatok többnyelvű tárolása. Az entitás-előfordulásokról rögzítendő adatokat a Minimum Viable Entity (MVE) dokumentáció, avagy adatminőségi modell tartalmazza – ez a Wikidata adatmodellje, illetve az LRM és a BIBFRAME relációinak tanulmányozása során készült. Az adatminőségnek a pontossághoz, az egyértelműséghez, a megbízhatósághoz, illetve a szerkezeti megfeleléshez kapcsolódó paraméterei egyaránt vannak.

A tudásgráf a legtöbb metaadatolási folyamatban felhasználható: ha a feldolgozási folyamat során autorizált adatokra van szükség, az entitáskezelő rendszerben kell a szükséges entitás-előfordulásra rákeresni (lesz felhasználói interfész, de alapvetően API közbeiktatásával), s az arra mutató azonosítót elhelyezni a metaadatközlések között. A WorldCat rekordjai például művekre, személyekre, illetve földrajzi helyekre mutató, az Entity Manager felé vezető hivatkozások sokaságát tartalmazhatják.

A szemantikus web technológiáján alapuló bibliográfiai adatfeldolgozás és -tárolás egyik legnagyobb lehetősége (és kihívása) a metaadatelemek és a hozzájuk rendelt értékek többnyelvűsítése. E kihívással kapcsolatos gondolatait *Pat Riva,* a könyvtári referenciamodell egyik kidolgozója osztotta meg előadásában. Gondolatmenetét a UBC-programban vállalt feladatoknál, illetve a hozzáférési pontok nemzetközi egységességének fejtegetésénél kezdte. Az egyes nemzetek bibliográfiai gyakorlata eltérően határozza meg ugyanazon névhordozó (pl. Platón) nevének autorizált alakját, így a globális egységesség nem valósul meg, helyette igen heterogén névváltozatok jelennek meg számos nyelven, amelyeket manapság

előszeretettel kapcsolnak össze és deklarálják azok egyenlőségét. Ennek a koncepciónak – tudniillik hogy minden, a világ közgyűjteményeiben kitüntetettnek minősített névformát össze kell gyűjteni – közismert példája a VIAF.

A bibliográfiai adatok több nyelven történő visszakereshetőségéről és megjelenítéséről azonban szükséges tovább gondolkodni. A felhasználók részéről már hosszú ideje jelen van a multilingvális hozzáférés igénye, mivel a közgyűjtemény használói populációja – egészét tekintve – többnyelvű, az egyes emberek is beszélhetnek több nyelven, nem is beszélve a gyűjtemények és az egyes bibliográfiai források nyelvi heterogenitásáról. Napjaink katalógusai azonban egy nyelv használatára épülnek, s jóllehet az RDA-szabályzat már demokratikusabban kezeli a leírás nyelvének kérdését, kevés útmutatást nyújt az egynél több kitüntetett nyelv (preferred language) használatára – így jó gyakorlatból is viszonylag kevés áll rendelkezésre.

A két nyelven történő bibliográfiai forrásleírás szép példája Kanada. Az ország könyvtárai a francia nyelvű forrásokat francia, az angol nyelvűeket angol nyelven írják le, a kétnyelvű forrásokról pedig két különálló, eltérő nyelvű rekord készül. Ezek a különböző nyelveken készülő forrásleírások extrém esetben két katalógusfelületen keresztül férhetők hozzá, silószerű architektúrát megvalósítva ezzel. A több nyelvet rendszeresen, napi szinten alkalmazó használóknak azonban hasznosabb, ha egyetlen kereséssel, egyetlen felületen kapják a releváns találatokat, függetlenül azok nyelvétől. Ezt az igényt igyekszik kiszolgálni a Quebec tizennyolc egyetemét tömörítő konzorcium 2020 nyarán indult közös katalógusrendszere, a Sofia.[54] Ebben a rendszerben egy rekord tárolódik, amelynek tartalma a felhasználói felület nyelvi beállításainak megváltoztatásával egyidőben átalakul. A helyes működéshez a metaadatelemek megnevezéseinek (a mezőcímkéknek), illetve az értékszótárak elemeinek fordítása szükséges, s a külső helyről beemelt autorizált névalakok is az adott nemzet gyakorlatának megfelelő formában jelenhetnek meg, hála az RDF-ben rögzíthető számos címkének. A többnyelvűség igénye a tárgyszavakra történő keresés során is jelentkezik, ekkor a tartalomjelölő kifejezések nyelvi változatait vagy maga az authority állomány tartalmazza, vagy valamilyen külső fordítószolgáltatást kell igénybe venni az információkeresési folyamat végrehajtása során. Ehhez arra is szükség van, hogy a felhasználó keresésének a nyelve egyértelműen megállapítható legyen, jó példa erre az *information* kifejezés, amely angolul és franciául egyaránt értelmezhető.

[54] Hozzáférhető a következő címen: https://sofia-biblios-uni-qc.org/en/.

A feldolgozó könyvtárosoknak régóta ismerősek a párhuzamos adatokkal vagy többes címoldallal rendelkező kiadványok, ezekben az esetekben a főcím kiválasztásakor alkalmazási sorrend tekintetében egymással fel nem cserélhető szabályokat kell alkalmazniuk. Vannak azonban olyan kiadványok is – az úgynevezett tête-bêche elrendezésűek – amelyek egy művet két fordításban tartalmaznak, s ezek egymáshoz képest fordított állásban (fejjel lefelé) helyezkednek el. Ilyenkor az elsődleges nyelv kiválasztása önkényes, ugyanakkor felmerül a kérdés, hogy egy vagy két leírás készüljön, s amennyiben a feldolgozó az utóbbi mellett dönt, hogyan lehet egymással összekötni ezeket a leírásokat? – tette fel a kérdést Pat Riva.

Hubay Miklós

Röviden a könyvtárak webes láthatóságáról és a schema.org-ról[55]

Richard Wallis kapcsoltadat-szakértő a Bibliographic Control konferencián a *schema.org*[56] nyílt szótár lehetséges könyvtári felhasználását ismertette. A strukturált adatok hasznossága elsősorban a keresőmotorok tükrében nyilvánul meg, hiszen részletes (és szabványos) leírásokkal szolgál a keresett dolgokról. Számtalan korai kezdeményezést követően a 2011-ben színre lépett schema.org egyre inkább a keresőmotorokat támogató, a webes láthatóságot elősegítő strukturált adatok szótárává válik a weblapok HTML-kódjába beágyazott formátumban, ami Microdata, RDFa vagy JSON-LD lehet. A schema.org jelentős penetrációval rendelkezik, a weboldalak közel felénél már alkalmazásban van. Google, Bing, Yahoo!, Facebook, Apple – csupán néhány példa a szótárat felhasználó nagy szolgáltatókra.

[55] Richard Wallis *Follow me to the library! Bibliographic data in a discovery driven world* című előadása alapján. Elhangzott: Bibliographic Control in the Digital Ecosystem: International Conference, Firenze, 2021. február 11. https://youtu.be/EoCt3ZYBmWI?t=5667 (2021.10.20.)

[56] A Schema.org kezdőoldala: https://schema.org/ (2021.10.20.)