Bódog András

SWIB20 – Fókuszban a kapcsolt nyílt adatok könyvtári felhasználásának jó gyakorlatai

A koronavírus-járványra való tekintettel rendhagyó módon online került megrendezésre a *Szemantikus web a könyvtárakban* (Semantic Web in Libraries – SWIB) konferenciasorozat legújabb része 2020. november 23. és 27. között. Az esemény honlapján livestream formájában lehetett nyomon követni az előadásokat, amelyek többségét később a SWIB YouTube-csatornájára is feltöltötték.* Az online rendezvény középpontjában a kapcsolt nyílt adatok (Linked Open Data – LOD) és a szemantikus web könyvtári felhasználásának kurrens jó gyakorlatai álltak. A SWIB szervezését ezúttal is a ZBW – Leibniz Közgazdasági Információs Központ és az Észak-Rajna-Vesztfália Felsőoktatási Könyvtári Központja (hbz) látta el. A negyedórás prezentációkból álló előadások végén ötperces blokkok álltak rendelkezésre a felmerülő kérdések megválaszolására.

A konferencia első napja *Audrey Tang,* Tajvan első digitális miniszterének egyórás bevezető előadásával kezdődött, amely a nyílt adatok társadalmi hasznosulását mutatta be a szigetország példáján keresztül, elsősorban a koronavírus-járvány tükrében. Tajvan – főként a korábbi SARS-vírus tapasztalatai alapján, valamint a Vuhanban felbukkant vírusra adott villámgyors óvintézkedési reakcióval – példamutató módon kezelte a járványhelyzetet, minek köszönhetően komoly korlátozó

* A SWIB20 weboldalon (swib.org/swib20/) nyilvánosan elérhetőek az előadások összefoglalói, a hozzájuk tartozó prezentációk és a korábbi évek SWIB konferenciáihoz vezető csatolók. Az előadásokról készült videofelvételek a SWIB YouTube-csatorján tekinthetők meg. https://www.youtube.com/channel/UCfLnEmKaWqtR_4V1CNeK6rQ (2021.02.10.)

intézkedésekre sem volt szükség. Ennek egyik kulcstényezője volt a Tang által kollektív intelligenciának nevezett alapelv, amely a nyílt adatokon alapuló közhasznú információk gyors terjeszthetőségére épül. Az ázsiai ország demokráciafelfogásának lényege, hogy minél többen vegyenek részt a politikai döntéshozatalban, amihez eszközül a modern technológia szolgál. A lakosság lényegében valós idejű összeköttetésben áll a tajvani járványkezelő törzzsel, a díjmentesen hívható telefonszám és a nyílt adatokat felhasználó webes ökoszisztéma biztosítása révén. Ez az Ázsiában egyedülálló nyílt társadalmi forma a fenti eszközök segítségével félpercenként frissülő, pontos információkat tudott szolgáltatni a gyógyszertárakban országszerte – társadalombiztosítással ingyenesen – elérhető maszk- és gyógyszerkészletekről. Ezenfelül az állampolgárok szabadon kifejezhették véleményüket, ötleteiket a járványhelyzettel kapcsolatosan. Tajvanon a politikában online módon részt vevő állampolgárok alkotják az úgynevezett „árnyékkormányt", akik többféle online felületen keresztül kinyilváníthatják véleményüket, melyet figyelembe vesznek a politikai döntéshozók. Az egyik legfőbb ilyen fórum a *gov zero (G0v)* mozgalom oldala, amely a tajvani kormányzati domain (gov.tw) módosításával érhető el az o betű nullára cserélésével. Az állami költségvetést bemutató *budget.g0v.tw* oldal például interaktív és átlátható módon enged betekintést a polgároknak a közpénzek elköltésébe. A kormányoldalak ilyen átalakításával a civilek ugyanazon adatok alapján tevékenykedhetnek, mint a hivatalos szervek, egyedülálló transzparenciát hozva létre. A társadalmat a közszféra és a piaci szféra összefogásával is segítik. Az előadó által digitális társadalmi innovációnak nevezett szemlélet központi eleme a jelentős mértékű átláthatóság, az önkéntesség és a digitális tevékenységekből eredő helyfüggetlenség. Mindehhez központi helyszínül a *társadalmi innovációs laborok* szolgálnak, amelyek több száz társadalmi innovátor – köztük magánszemélyek és vállalatok – együttműködésében valósultak meg. Audrey Tang miniszter szerdánként napi 12 órában online felkereshető bármely polgár számára, bármilyen témát illetően. A laborok mesterséges „lakói" az *MIT Media Lab* által készített önjáró triciklik, amelyek szabadon és kreatív módon, egyedi igények szerint használhatók és programozhatók bárki által, akár személyszállításra, akár bevásárlást követő cipekedésre. A *mesterséges intelligencia* (artificial intelligence) Tang olvasatában a *segítő intelligenciá*t (assistive intelligence) jelenti. A társadalmi innováció e megközelítésében a kormányzás, a társadalom és a kereskedelem együtt keresi a közös nevezőt a problémák megoldására. A *sandbox.org.tw* oldalon bármely társadalmi innovátor tehet javaslatot egy egyéves kísérletre, legyen szó például önvezető járművekről, platformgazdaságról**,

**„A platformgazdaság lényege, hogy a szolgáltatók – de még a „kézzelfogható" termékek gyártói is – önálló ökoszisztémákban kapcsolódnak egymáshoz, aminek eredményeként több felhasználót (ügyfelet, fogyasztót) képesek elérni, utóbbiak pedig jobb minőséget, nagyobb értéket kapnak." Forrás: Varga Zsigmond: *Itthon még gyerekcipőben jár, de enélkül már nehéz lesz vállalkozni a jövőben.* = Forbes Magyarország, 2018. május 18. https://forbes.hu/uzlet/itthon-meg-gyerekcipoben-jar-de-enelkul-mar-nehez-lesz-vallalkozni-a-jovoben/ (2021.02.10.)

5G távközlésről vagy pénzügyi technológiáról (financial technology). A kísérletet követheti az adott újítást bevezető szabályozás, kevésbé látványos siker esetén pedig a társadalom egésze számára átlátható módon vonják le a tanulságokat. A társadalmi innovációs laborok regionális munkáját Tang miniszter online telekonferenciákkal koordinálja. Minden új innovációt intenzív, mesterséges intelligencia által moderált társadalmi párbeszéd követ az online térben. A szigorú gépi moderáció következtében jelentősen nő a konszenzussal záródó viták száma, átláthatóbb és gyorsabb információkat juttatva el a döntéshozókhoz. Tajvanon az egész társadalmat átjárja a nyílt adatokkal történő transzparens információközlés. Tang előadását az úttörőnek számító technológiai kifejezések társadalmi hasznot kifejező átalakításával, újrafogalmazásával zárta. – Így lett a *dolgok internetéből* (internet of things) *teremtmények internete* (internet of beings), a virtuális valóságból megosztott valóság, a gépi tanulásból kollektív tanulás, a felhasználói élményből emberi élmény, a szingularitásból pedig sokaság (plurality).

Az első nap könyvtárszakmai előadóinak sorát *Ricardo Eito-Brun,* a madridi III. Károly Egyetem oktatója indította. Előadásának témája az intézményi repozitórium automatikus indexelése volt a *SKOS* (Simple Knowledge Organization System) tudásszervezési rendszer felhasználásával. A legtöbb intézményi repozitórium esetében jelentős deficitet jelent a hatékony indexelési gyakorlat hiánya. A kutatók többsége a kulcsszavas keresést részesíti előnyben, így szolgáltatói oldalról nehéz előzetesen megfelelni a többnyire nem egyértelmű szavakat használó egyéni keresési preferenciáknak, amelyek ritkán mutatnak átfedést a tartalom leírására használt kontrollált szótárakból származó tárgyszavakkal. Ez a probléma még szembeötlőbb, amikor több repozitóriumból aggregálják a tartalmakat. A spanyol gyakorlat a deszkriptorok és a kulcsszavak tezauruszból történő automatikus kiosztását preferálja. Az UNESCO tezaurusza képezi az alapot, amelyhez a SKOS közvetítésével készítenek megfeleltetéseket. Szoftveres segédletként a *PoolParty*t használják. A kísérleti folyamat során automatikusan azonosítják a tezaurusz azon fogalmait, amelyekkel leírható a repozitóriumi tartalom. A dokumentumokhoz társított tezaurusz-deszkriptorok meghatározását követően maga a tezaurusz használható a lekérdezések bővítésére, továbbá segít a végfelhasználóknak a keresőkifejezések kiválasztásában. Ez az eljárás hatékonyabbá teszi a keresést, lehetővé téve a használók számára a kapcsolódó, általánosabb és szűkebb fogalmak kiválasztásával a keresőkérdés javítását.

A SWIB immár visszatérő előadója *Osma Suominen,* a Finn Nemzeti Könyvtár munkatársa, az *Annif* többnyelvű gépi tárgyszavazó eszköz fejlesztésében oroszlánrészt vállaló információs rendszerspecialista az Annif fejlesztéséről és aktualitásairól számolt be. Minden gépi tárgyszavazó rendszer fejlesztésének indoka az, hogy a manuális tartalmi feltárás rendkívül idő- és munkaigényes feladat, külö-

nösen, ha nagy méretű gyűjtemények feldolgozásáról van szó. A finn gyakorlat alapját a már létező metaadat-gyűjtemények gépi tanulással segített felhasználása képezi, a keresés és a felfedezés minőségének javítása érdekében. A metaadatok bázisát a *Finna* közgyűjteményi aggregátorportál nyújtja. A gépi rendszer neve is ebből ered (Annif), amely nem más, mint a közgyűjteményi portál neve (Finna) visszafelé olvasva. A rendszer korai prototípusa 2017-ben készült el, mára azonban túljutott a kísérleti fázison. A fejlesztők az implementációhoz több célt és alapelvet is megfogalmaztak, melyek a következők: többnyelvűség (finn, svéd és angol), független szótárindexelés, a különböző tárgyszóindexelő – elsősorban gépi tanuláson alapuló – algoritmusok támogatása, parancssoros felhasználói felület (Command Line Interface – CLI), más rendszerekbe integrálható webes felhasználói felület és a *REST API* (alkalmazásprogramozási interfész) kialakítása, valamint közösségorientált, nyílt forráskódú működés. Utóbbi elvnek megfelelően a projekt forráskódja elérhető a *GitHub*on, Python package-ként a *PyPI* és Docker images-ként a *Quay.io* felületén. A gépi tanuló algoritmusok képzésére és tesztelésére a Finna.fi metaadatait, online „Kérdezd a könyvtárost!" kérdés-válasz párokat, a Jyväskylä Egyetem szakdolgozatait és disszertációit, kiadói könyvleírásokat, valamint a nemzeti könyvtár e-könyveit használták fel. Ezt követően adott mintákban az automatikusan javasolt tárgyszavakat behasonlították a könyvtárosok által manuálisan hozzáadott tárgyszavakkal. Az Annif már gyakorlati alkalmazásban van a Jyväskylä Egyetem *JYX* repozitóriumában, ahol a rendszer a hallgatók által feltöltött szakdolgozatokhoz és disszertációkhoz javasol tárgyszavakat: üres mezők kitöltése helyett a gépi javaslat listájából választhatják ki a szerzők a megfelelőnek ítélt tárgyszavakat a saját kulcsszavaik mellé. Egyébiránt a Finn Nemzeti Könyvtár számos DSpace-alapú intézményi repozitórum hostja, melyek közül három (Osuva, Trepo, Theseus) 2020-tól kezdődően szintén megkezdte az Annif használatát.

A projekt újítása a 2020 májusában indított *Finto AI* gépi tárgyszavazó eszköz és API szolgáltatás. Ezen a webfelületen a rendszer automatikus szövegelemzést követően az oda beillesztett tetszőleges angol, finn és svéd nyelvű szöveg tárgyszavazását végzi el a finn nemzeti ontológia (YSO) tezauruszát használva. 2020 novemberétől a Finn Nemzeti Könyvtár a gyakorlatban is alkalmazza a Finto AI-t, az elektronikus tárhelyek építése során tárgyszójavaslatokat állítanak elő a segítségével. A könyvesboltokat és könyvtárakat ellátó *Kirjavälitys Oy* logisztikai vállalattal is együttműködnek: az Annif/Finto AI a kiadóktól kapott leírások alapján javasol tárgyszavakat. Az új könyvek így már a szakértők által véglegesített tárgyszavakkal kerülnek be a cég által is használt *Melinda* nevű közös katalógusba, amely a finn nemzeti bibliográfiát (Fennica) is magában foglalja.

Osma Suominen a fejlesztés során levont tanulságokat is ismertette. Alaptézis, hogy a tárgyszavazás nehézkes munka, amelyet befolyásol az emberi szubjektum, de az algoritmusok is gyakran hibáznak, és az emberi tévesztésekkel ellentétben

az emberi agy számára teljesen értelmetlen és ésszerűtlen megoldásokat kreálnak. Utóbbi hibalehetőség mérséklésére ajánlott párhuzamosan többféle algoritmust, illetve többféle módszert is alkalmazni. Összegezve elmondható, hogy a finn projekt példaértékű a maga nemében, ráadásul kezdettől fogva a nemzetközi együttműködésre gondolva fejlesztették, így a rendszer alkalmazhatósága szülőhazáján is messze túlmutathat a jövőben, melyre már akadnak példák a Holland Nemzeti Könyvtár és a konferencia-házigazda német ZBW Leibniz Közgazdasági Információs Központ részéről.

Ez utóbbi intézményt képviselte *Anna Kasprzik,* az első nap utolsó előadója, aki programkoordinátorként a német *AutoSE* gépi tárgyszavazó projektet ismertette. Többéves újratervezést követően a ZBW újrafogalmazta a gépi tárgyszavazással szemben támasztott követelményeit, és immár az elméleti kutatás helyett a gyakorlati használhatóságra helyezte a hangsúlyt. Ezt szem előtt tartva zajlott a 2014 és 2018 között „házon belül" fejlesztett, a *Közgazdasági Tezaurusz*t (Standard-Thesaurus Wirtschaft – STW) és a *SKOS*-t lexikai alapként használó, többféle gépi tanulási módszert kombináló *AutoIndex* projekt, amelynek tudományos alapjait egy doktoranduszhallgató kutatása fektette le. További kritérium volt a nyílt forráskódú szoftverek alkalmazása és előállítása. Ez a fejlesztés kapott új lendületet 2019-ben egy átszervezést követően, immáron AutoSE néven, alkalmi projekt helyett állandó egyetemi tevékenységként. A jelenlegi munka kétszer két pillérre osztható: egyrészt beszélhetünk produktív rendszerről, másrészt alkalmazott kutatásról és a módszerek tudományos fejlesztéséről. Mindkét megközelítés lebontható gépi tanulási folyamatokra és szoftvermenedzsmentre, illetve működésre.

A szoftverarchitektúra új módszerekkel történő bővítésénél és a rendszerfrissítésnél különböző mikroszolgáltatások és a *Docker*hez hasonló konténer-technológiák, illetve a *Kubernetes*hez hasonló ütemezőszoftverek alkalmazásával és folyamatos integrációval kerülik el a szolgáltatáskiesést. A rendszer metaadatbázisát egy közös katalógus nyújtja, az AutoSE maga pedig az *EconBiz* elnevezésű intézményi repozitórium adatbázisába és discovery rendszerébe illeszkedik. A közeljövő megoldandó feladata lesz a gépi feldolgozás tárgyszavainak visszaintegrálása a közös katalógusba, valamint a katalógussal összeköttetésben álló digitális aszszisztens fejlesztése.

Az előző előadásra reflektálva Kasprzik megosztotta a hallgatósággal, hogy az AutoSE fejlesztői rendszeresen ötleteket cserélnek az Annif csapatával, valamint a GitHubon keresztül szerepet vállalnak az Annif fejlesztésében is. Az Annif maga keretrendszerként képezi az AutoSE alapját. A ZBW szakértői a finnek forráskódját módosítják saját igényeiknek megfelelően, mivel a teljes körű integráció már a finn modell németországi használhatóságának a rovására menne. A németek *stwfsa* elnevezésű backendjét pedig hamarosan integrálják a finn rendszerbe annak egy újabb lexikai elemeként.

*

A SWIB20 második napja gyakorlati jellegű volt. A délután során előzetes jelentkezést igénylő online workshopokon lehetett részt venni. Elsőként *Osma Suominen* és csapata az *Annif* többnyelvű gépi tárgyszavazó eszközének használatába engedett betekintést. A résztvevők előzetesen oktatóvideókat és írásbeli feladatokat kaptak, ezekkel kapcsolatos tapasztalataikat vitathatták meg a fejlesztőkkel. (Maga az online gyakorló tananyag elérhető a GitHubon.)

A második workshopot *Jakob Voß* és *Stefan Peters* tartotta a Verbundzentrale des GBV-t képviselve. A műhelymunka középpontjában a könyvtári tudásszervező rendszerek közötti adatcserét elősegítő *coli-conc* projekt leglátványosabb eredménye, a *Cocoda* webes megfeleltető alkalmazás állt. A Cocoda egyszerűbbé teszi az eltérő osztályozási rendszerek, tezauruszok és kontrollált szótárak fogalmainak egymásnak való megfeleltetését. A fejlesztők iránymutatásait követve a workshop résztvevői egy egyszerű katalogizálási alkalmazást kódolhattak a források kontrollált szótárakból nyert fogalmakkal történő szemantikus címkézésére.

A kanadai LYRASIS-től *David Wilcox* a *Fedora* nyílt forráskódú repozitóriumplatform kapcsoltadat-kezelési és -megőrzési módszerét (RDF-alapú forrásmenedzsment, leíró és hozzáférést biztosító metaadatok létrehozásával és frissítésével) mutatta be a közös munka keretében. A résztvevők a felhőből hostolt Fedora-példányok segítségével tudták felfedezni a rendszer legújabb változatának (Fedora 6.0) összetevőit, amelynek középpontjában a hosszú távú megőrzés áll, az *Oxford Common File Layout* (OCFL) specifikációjával összhangban.

A nap utolsó műhelygyakorlata a *SkoHub* használatát mutatta meg a webalapú metaadat- és tartalomkezelés céljából. Az előadók és instruktorok a *hbz* munkatársai, *Adrian Pohl* és *Steffen Rörtgen* voltak. A *SKOS,* az *AktivityPub* és a *Linked Data Notifications* webes szabványokon alapuló, a tavalyi SWIB-konferencián bemutatott SkoHub kipróbálása során a workshop résztvevői megtapasztalhatták egy kontrollált szótár közzétételét SKOS-sémaként, Git-bázisú szerkesztési folyamatban, megismerhettek egy webkonfigurációt strukturált metaadatok webböngészős létrehozására, valamint egy adott témára történő feliratkozást, továbbá egy adott téma közzétételét a feliratkozó érdeklődőknek.

*

A konferencia harmadik napja a BIBFRAME (Bibliographic Framework) és az authority tematikája körül forgott. A *PCC* (Program for Cooperative Cataloging – Kooperatív Katalogizálási Program) részéről *Paloma Graciani-Picardo* és *Nancy Lorimer* tartottak előadást a katalogizáló közösségnek szánt BIBFRAME-

alkalmazásprofilok fejlesztéséről. A munka célja összhangban van az *LD4 Közösség* 2020-as küldetésnyilatkozatával, amely szerint *„a világot a könyvtári adatok, míg a könyvtárakat a világ adatai gazdagítják”*. Ezen elgondolás szerint a könyvtári kapcsolt adatok strukturált, minőségi, hasznos és kiszámítható adatkészletet képeznek. E könyvtári kapcsolt adatokat kívánja a PCC – a BIBFRAME-et ontológiaként felhasználva – egy kooperatív katalogizáló ökoszisztéma keretében összekötni mások kapcsolt adataival, legalább olyan szinten, amelynek során a könyvtárak képesek megosztani saját kapcsolt adataikat másokkal, valamint használni tudnak más forrásból származó adatokat. Jelenleg az *RDF* (Resource Description Framework) és a BIBFRAME még túl szabadon értelmezhetőek. Utóbbi esetében ugyanazon entitás többféleképpen is modellezhető, kibővített szótárak szükségesek a katalogizálás követelményeinek teljesítéséhez. További problémát jelent, hogy nem áll rendelkezésre hivatalos megfeleltetés az *RDA* és a BIBFRAME között, és jó gyakorlatokat sem ismerünk a könyvtári katalogizálásra kapcsoltadat-környezetben, ezért a katalogizálási területnek szüksége van egy tágabb értelemben vett ökoszisztémába történő integrációra. A PCC egyik munkacsoportja a *Sinopia* rendszerhez fejleszt és értékel ki létező, a szervezet standardjainak megfelelő alkalmazásprofilokat. Sablonok kialakításával kívánnak létrehozni szabványosított BIBFRAME-modelleket, RDA-BIBFRAME kapcsolatokat, továbbá szorgalmazzák a PCC-környezetben alkalmazott szótárak használatát.

A következő előadásban *Jeremy Nelson* a Stanford Egyetem könyvtárából az előző prezentációban is említett *Sinopia kapcsoltadat-szerkesztő*t ismertette. Az implementálást és a teszteket követően a gyakorlati alkalmazás (RDF-leírások formájában) 2019-ben kezdődött meg. Viszonylag új módosítás a felhasználói tapasztalatok alapján, hogy a szerkesztőalkalmazás átváltott a korábbi JSON-alapú (JavaScript Object Notation) forrássablonról RDF-alapúra, valamint dedikált profilszerkesztővel és új API-val is bővítették a repertoárját. Előbbi, magával a szerkesztővel egyetemben, a Kongresszusi Könyvtár által kifejlesztett megoldásokon alapul. A jelenlegi architektúra középpontjában a *Sinopia API* áll, ehhez kapcsolódik az *AWS* (Amazon Web Services – Amazon webszolgáltatások) rendszerén belül a *Lambda* RDF-MARC konverziós eszköz, a *DocumentDB* dokumentumorientált-adatbázisrendszer, a *Cognito* indexelő eszköz, valamint az AWS környezeten kívül a külső partnerekhez kapcsolódó authority-lekérdező szolgáltatás (Questioning Authority – QA). Az új, rugalmas RDF-sablonok használatával lehet elkészíteni a forrásleíró metaadatok rétegét. Ezekkel a sablonok egyéni szerkesztése és a katalogizálás egyaránt elvégezhető.

Az előadó a szerkesztőbe belépve, mindezeket a gyakorlatban mutatta meg. Hangsúlyozta, hogy a szerkesztőfelület szótára ugyan a BIBFRAME-ére épül, ám egyéb szótárakkal (például a Schema.org-gal) is használható. A Sinopia szerkesztője mindezeken felül lehetővé teszi még a QA-szolgáltatás révén külső

entitások hivatkozását, és külső források révén tripletek megalkotását. Összegezve elmondható, hogy a kapcsolt adatok szerkesztőjének funkcionalitásában a Sinopia-partnerek gyakorlata és igényei köszönnek vissza. Gondolati síkon a továbbfejlesztés következő fázisaként a gépi tanulás alkalmazását tervezik némely munkafolyamat gördülékenyebbé tétele érdekében.

A BIBFRAME-szekció zárásaként egy esettanulmányt ismerhettünk meg a Texasi Egyetem Harry Ransom Központjából *Brittney Washington* és *Paloma Graciani-Picardo* jóvoltából. Régi könyvek katalogizálását mutatták be az *LD4P2* (Linked Data for Production Phase 2) pilotprojekt eredményeként. A projekt három téma szerint szerveződik: MARC–BIBFRAME konverziók elemzése, kapcsolt adatok katalogizálását meghatározó helyi munkafolyamatok meghatározása, illetve alkalmazásprofilok kidolgozása különgyűjteményi anyagok számára. A munka alapját a *kongresszusi könyvtár*i régi anyagok BIBFRAME-profilja képezte. Ezeket példányszinten és az instance entitásra vonatkozó profilok meghatározására tovább kellett finomítani, így integrálták még az *RBMS* (Rare Books Manuscripts Section – az ALA régi könyvek és kéziratok szekciója), az *ARLIS/NA* (Art Libraries Society of North America – Észak-amerikai Művészeti Könyvtárak Egyesülete), illetve az *SAA* (Society of American Archivists – Amerikai Levéltárosok Társasága) BIBFRAME-kibővítéssel rendelkező ontológiáit. A kontrollált szótárak elemeit a *PeriodO* (művészettörténet), a *Kongresszusi Könyvtár* (név-authority), a *Getty Művészeti és Építészeti Tezaurusz,* a *Ligatus Language of Bindings* (könyvkötészet) és az *RBMS* szókészletéből integrálták a projekt által fejlesztett Sinopiába. A különgyűjteményi dokumentumok leírására a MARC nem ideális, mivel a ritka dokumentum egyediségét tükröző ismérvek nehezen fejezhetők ki ebben a formátumban. Ezen információk feltüntetésére a fizikai jellemzőket rögzítő 300-as mező mellett leginkább az 500-as „általános megjegyzés" mezőt szokás használni, de használatos még az 510-es „megjegyzés idézettségről, hivatkozásokról" mező is, ahol szabványosított szöveges karakterláncban rögzítik az erre vonatkozó egyedi információkat (pl. előző tulajdonosok [possessorok], hiszen nem egyszer ez az egyetlen ismérv, amely alapján visszakereshető egyes bibliográfiákban az adott dokumentum). Ugyanezeket és további, az adott ritkaságot azonosító információkat rögzítenek az 590-es (local note – helyi megjegyzés) és a 79X mezőkben (pl. származási hely). Az LD4P2 részeként a partnereknek lehetőségük van a tapasztalataik megosztására és ezek alapján a profilkészítés támogatására. Például ha egy PhD-hallgató díszkötéses könyveket keres a gyűjteményben, az ő funkcionális igénye az, hogy példát találjon a díszkötés minden típusára, és példányspecifikus kötésadatokat kapjon a keresés során. Ebben az esetben a releváns entitások a műfaj, a megjegyzések, a példányok és a kötés. Egy, a 17. századi magánkönyvtárakat kutató használó az annotációk példáit keresi, így neki ezekre az információkra van szüksége példányszinten. Ez esetben releváns a possessor és a kötet külső megjelenése.

Paloma Graciani-Picardo bemutatta a ritka könyvek leírásának modelljét, ahol egy adott példány tulajdonosainak sorát az 590-es és a 79X MARC 21 mezőkben egyaránt rögzített adatok alapján írják le, majd illesztik be a modellbe. Így egy példánynál láncszerűen prezentálható egy kötet „élettörténete", a hozzácsatolt tulajdonlást (ownership) és hozzájárulást (contribution) is feltüntetve. A formai feltárás során állandó kihívást jelent a régi rekordok konverziója. Régi könyveknél ez különösen nehézkes, mivel az 590 és a 79X mezőkhöz sincs elérhető megfeleltethetőség.

A következő szekció témáját az authority adatok képezték. Elsőként *Jim Hahn,* a Pennsylvaniai Egyetem metaadat-kutatásának vezetője ismertette a BIBFRAME-féle instance entitások „bányászatát" a *Share-VDE* projektben. A globális együttműködés keretében működő Share-VDE metaadatainak többségét konvertált MARC-rekordok képezik. A BIBFRAME adatmodellben ezek a leírások a művek szerint rendeződnek klaszterekbe. Kulcsfontosságú tényezők az *instance entitások* leírásai, mivel ezek kapcsolódnak más *művek*hez és *instance entitások*hoz. Az entitások közül az instance biztosítja a fizikai jellemzőkre utaló és a kiadásra vonatkozó leíró adatok megfelelő csoportosítását. Ez esetben visszatérő probléma, hogy a megjelenési adatok puszta karakterláncként vannak megadva, és nincs kontrollált azonosító hozzájuk társítva. E kihívásra az egyik válasz az OCLC kísérleti kiadói név-authority kutatási projektje, a *PNAF* (Publisher Name Authority File) volt, amelyet a rekordok adatbányászata (nyelvi kód, ISBN-prefixum, MARC 260) és a megjelenési adatok automatikus klaszterezésének útján építettek.

Kísérleti projektként a *Penn Könyvtárak* több mint 5 millió MARC-rekordjának metaadataival reprodukálták a PNAF-prototípusfolyamatát a Share-VDE környezetben. Ezt követően a VIAF (Virtual International Authority File – Virtuális Nemzetközi Authorityfájl) megfelelő adatelemeit alkalmazták a korpusz 260$b karakterláncain a kiadók azonosítása érdekében. Félautomata behasonlítással (vagyis emberi beavatkozással) a MARC megjelenési adatainak 30%-át sikerült VIAF-entitásokhoz kötni. Teljesen automatikus behasonlítás esetében a rekordok 16%-a érte el a 0,9–1,01 konfidenciaintervallumot, 4%-uk a 0,8–0,89, míg 3,4%-uk a 0,7–0,79 tartományt. A rekordok 59,6%-a nem rendelkezett ISBN-nel, ezeknek 9,8%-át sikerült félautomata módon VIAF-entitással párosítani. A sikeres behasonlítás (különösen ISBN-nel nem rendelkező dokumentumoknál) a hiteles BIBFRAME-entitások azonosításának a kulcsa.

A következő előadó *Charlie Harper,* a clevelandi *Case Western Reserve Egyetem* digitális kutatásokkal foglalkozó szakértője a metaadat-tárgyszócímkék *Doc2Vec* és *DBPedia* általi generálásáról számolt be. A projektet több tényező inspirálta, az egyik a témamodellezés mint az adatfelfedezés javításának egyik lehetsé-

ges módja. Visszatérő probléma a dokumentumok klaszterezése „bejáratott" klasztercímkék nélkül, ami interoperabilitási problémákat is magában hordoz. Az intézményi repozitóriumok nehéz kereshetősége és a szerzők által hozzárendelt kulcsszavak kezelése is problematikusnak bizonyul. Utóbbiaknál az általuk vizsgált 77 ezer ilyen típusú kulcsszó 80%-át az elmúlt öt évben csupán egyszer használták, míg a többi kulcsszó túlságosan gyakori és tág fogalmakat képezett le. E probléma megoldása érdekében a gépi tanuláshoz fordultak. A Word2Vec természetes nyelvfeldolgozó (Natural Language Processing – NLP) algoritmus a sokdimenziós vektortérbe integrált szavak között felismeri a fontos kapcsolatokat. A Doc2Vec már a teljes szövegre kiterjeszti ezt az eljárást. Az algoritmushoz a „képzési adatokat" a DBPedia előcímkézett szövegkorpusza szolgáltatja (5 millió weboldal). Ez több szempontból is megfelelő: nyílt hozzáférésű, kapcsolt adatokat tartalmaz és többnyelvű. Harperék hipotézise szerint a nyelvfeldolgozó folyamat vektortere túlterhelt DBPedia-oldalakkal, ezért a folyamatot a linkeken keresztülfuttatva magasabb szintű információszűrést valósítottak meg. Ennek eredményeképpen tárgyi és fogalmi címkéket tudtak létrehozni. A DBPediában minden oldal rendelkezett egy *dct:subject* címkével, amelyet tovább lehetett fogalmi síkon szélesíteni a *skos:broader* címkével. A kezdeti kísérleteket követően megkezdték a tárgyszócímkézés folyamatának finomítását az adatkészlet kibővítésével, valamint a címkék minőségének javításával, szakértőket is bevonva a munkába, még mindig túl sok volt ugyanis az irreleváns címke. Jelenleg a vizualizáció prototípusán és a keresőfelület felhasználói interfészén dolgoznak, amelynek jelenlegi változatát Harper is prezentálta néhány slide erejéig előadása zárásaként.

A következő előadást *Joeli Takala* tartotta a Finn Nemzeti Könyvtár *Finto* (Finnish thesaurus and ontology service – Finn tezaurusz- és ontológiaszolgáltatás) szótárainak közös hierarchiáját építő automata eszközökről. A 2003 óta fejlesztett holisztikus megközelítésű finn ontológia (lásd *KOKO*) központi eleme az *YSO* nevű központi finn ontológia és a hozzá kapcsolódó 15 további területspecifikus kontrollált szótár. Utóbbiak mindegyike kapcsolt nyílt adatokkal működik (kb. 7,6 millió skos:rdf triplet). A legnagyobb kihívást a különböző szótárak „együttműködésében" a szemantikus interoperabilitás jelenti. Az eltérő szótárak összehangolását a *TBC* (TopBraid Composer) és *Vocbench* szótárszerkesztők, a *SKOSIFY* minőségjavító eszköz és egyéb validátorok, a *MUTU* elemzőeszköz és végül a *KOKOAJA* fogalomösszevonást segítő eszköz szolgálják. Az előadó a rendszer folyamatábráján keresztül mutatta be az ontológiabővítés menetét. Ez egyaránt magában foglal automata eljárásokat, validálásokat és kézi beavatkozást. Rendező irányelv, hogy minden fogalomnak saját URI-val kell rendelkeznie, minden URI-nak egyedinek kell lennie, illetve az URI-kat nem távolíthatják el a webről, amikor egy fogalmat kivonnak a tezauruszból vagy mó-

dosítanak. További kihívás, hogy a szakterületi szótárakat a központi ontológián keresztül kell összekapcsolni, és előbbiek hierarchiája nem változtathatja meg a központi ontológia hierarchiáját. Itt kap szerepet a KOKOAJA összevonást segítő eszköz a SKOSIFY első minőségjavító alkalmazását követően. Az összevonást a validáció, újabb minőség-ellenőrzés, majd a fogalom elemzése követi, végül sor kerül a közzétételre. A Fintót széleskörűen használják Finnország-szerte a könyvtárakban, múzeumokban, levéltárakban, illetve közigazgatási és kormányzati területen.

*

A konferencia negyedik napja az azonosítók tematikájával folytatódott. Az első előadást a spanyolországi Salamancai Pápai Egyetem (UPSA) oktatója, *Ana Maria Fermoso García* tartotta, az *OpenUPSA*-nak nevezett kurrens kutatási információs rendszer (Current Research Information System – CRIS) tervezett szemantikustechnológia-integrációjáról és tudásszervezetéről. Az OpenUPSA nagy előnye az átláthatóság, ám az egyetem nagy hátránya az információközvetítés terén, hogy nem rendelkezik saját intézményi repozitóriummal, ezért olyan CRIS-megoldást keresnek, amely ezt a funkciót is képes betölteni. A leendő rendszert négy, egymást átfedő szempont alapján kívánják kialakítani: a kutatási információ visszakeresése és integrációja, vizualizációja, kezelése, illetve megosztása és újrafelhasználása.

A visszakeresési, integrációs, menedzsment- és adatvizualizációs funkciókat egy intézményi repozitóriumként is működő relációs adatbázissal tervezik megvalósítani. Az egyéni információk és dokumentumok, illetve a relációs adatbázis közötti integrációról PDF-, HTML-, Excel- és JSON-parserek gondoskodnak. Az információ lekérdezésére és újrafelhasználására egy, a *CERIF* (Common European Research Information Format – közös európai kutatási információs formátum) ontológiáján alapuló OpenUPSA ontológiát alkalmazó szemantikus repozitóriumot terveznek SPARQL végpontokkal. A javasolt szoftverarchitektúrában a relációs adatbáziskezelő és a szemantikus repozitórium között is egyirányú kapcsolat van a szemantikus összetevő irányába.

A következő előadást *Eva Seidlmayer,* a németországi ZB MED élettudományi információs központ munkatársa tartotta, a szerzők (illetve ORCID-azonosítóik) és publikációik Wikidatában történő összepárosításának munkafolyamatáról. A Wikipédián és társprojektjein alapuló Wikidata a szemantikus adatok hatalmas nyílt tudásbázisát képezi. 2020 elején mintegy 71 millió egyéni azonosítóval (Q-ID) rendelkező tételt tartalmazott. Ezek 31,5%-a (22,5 millió) tudományos cikk, 8,9%-uk (6,3 millió) pedig személyeket takar. A szerzők azonosításában fontos szerepet játszik a Q-Aktiv projekt a Kieli Egyetem, a ZBW Kiel és a kölni ZB MED együttműködésében. Ennek keretében bibliográfiai adatkészleteket gaz-

dagítottak a Wikidata API-ján keresztül. Az alacsony lefedettség (kevesebb mint 13%) oka a Wikidata API instabilitása tömeges lekérdezés esetén, a Wikidata hiányzó publikációs és szerzői Q-ID azonosítói, valamint a *has_author = P50* reláció hiánya a szerzői és megjelenési Wikidata-adatokban. Az egyedi azonosítók kulcsszerepet játszanak a digitális tudományos információközvetítésében, amelynek egyik fontos eleme az ORCID szerzői/közreműködői azonosító. A 2020. októberi ORCID statisztika szerint 9,8 millió kutató hozott létre ORCID azonosítót, és 62,8 millió publikációt regisztráltak. A projektben érintett adatpárok szavatolják a Wikidata fent felsorolt adathiányainak (szerző, publikáció, P50) pótlását. Az adatkészletek közzététele az ORCID-ból származó dokumentumazonosítók (PMID, PMC, DOI, EID, DNB, WOS) kinyerésével, majd ezek és a szerzők wikidatás behasonlításával, szükség esetében pótlásával történik. Az OrcBot a Wikidata-információ sűrítésére szolgáló bot, ez teszi lehetővé a szerzői és publikciós adatkészletek kombinálását a mindkét készletben megtalálható ORCID-azonosítók révén. Abban az esetben, ha egy szerző nem szerepelne a cikkek szerzői között, a Wikidata OrcBot létrehoz egy JSON-sablont a szerzői adatok rögzítésére. 2019-ben a Wikidata API alkalmazásával 948 szerző-publikáció adatpárt hasonlítottak már be, míg az új módszer szerinti tömeges adatátvitel (data dump) során 7,6 millió adatpárt ellenőriztek, és 33 ezer szerzőt rendeltek hozzá publikációikhoz, ellentétben az előző módszer 12 ezres adatával. A jelenlegi cél a meglévő adatok folyamatos javítása és kibővítése egyéb ismérvek (szervezeti hovatartozás, finanszírozás stb.) alapján.

Az azonosítókkal foglalkozó szekció utolsó előadójaként *Matt Miller* ismertette a Kongresszusi Könyvtár *ID.LOC.GOV – Linked Data Service* és *Wikidata* projektjének aktuális újdonságait. Az id.loc.gov a Kongresszusi Könyvtár kapcsoltadat-platformja authority adatokkal, kontrollált szótárral és egyéb szolgáltatásokkal. 2019-től kezdődött meg a Wikidata-rekordok integrálása a szolgáltatásba, felhasználva a Wikidata publikus SPARQL végpontjait. A Wikimedia-környezet a közintézmények számára világszerte igen hatékony szolgáltatási felület, mivel általa igen könnyen tehetnek közzé a weben kulturális örökséget megőrző tartalmakat. Jelenleg a több mint tízmillió Kongresszusi Könyvtár-authority közül csupán 1,2 millió (többségében névadat) van a Wikidatához kapcsolva. Miller az előadás előtt pár nappal ellenőrizte a Wikidata aktuális méretét, amely immáron 90 milliósra nőtt, tíz hónap alatt közel 20 milliós növekedést produkálva a Wikiadatok strukturálása terén, amely teljesítmény önmagáért beszél. A Kongresszusi Könyvtárhoz kapcsolt adatok több mint fele (53%) az angol nyelvű Wikipédiáról származik, 33%-a a németről, 28%-a a franciáról, 23%-a pedig a Wikimedia Commonsról. A Wikidatába integrált Kongresszusi Könyvtár-azonosítókhoz kapcsolt külső azonosítók 96%-át a VIAF, 90%-át a WorldCat Identities, 68%-át az ISNI, 40%-át a GND teszi ki.

A SWIB20 utolsó előadói blokkjában gyakorlati alkalmazásokat tekinthettünk meg. Elsőként *David Seubert, Shawn Averkamp* és *Michael Lashutka* ismertették az *Amerikai Történeti Felvételek Diszkográfiájához* (Discography of American Historical Recordings – DAHR) köthető kapcsoltadat-projektet. Az „audioenciklopédia" jobb kereshetősége érdekében a DAHR-szerkesztők több mint 20 ezer olyan nevet azonosítottak az adatbázisban szereplő nagyjából 60 ezerből, amely már rendelkezik Kongresszusi Könyvtári név-authority fájllal (Library of Congress Name Authority File – LCNAF). Adatkészletének bővítésével az a távlati cél, hogy a DAHR a szakterülete authority szolgáltatásává váljon. A projekt során a meglévő LCNAF-azonosítók alkalmazásával egyéb „aratható" vagy nyílt adatkészleteket (pl. VIAF, MusicBrainz, Wikidata, AllMusic stb.) kerestek, majd hasonlítottak be saját adatbázisukkal, a rekordokhoz authority fájlt és egyéb webhelyet (pl. Spotify vagy iTunes azonosítót) kapcsolva, valamint azokat további adatelemekkel gazdagítva. Például a Wikipedia API alkalmazásával szabad felhasználású képeket, valamint leírásokat nyertek így a rekordokhoz. A több mint egymillió hangfelvétel kiegészítésére learatott rekordok karbantartására a Kaliforniai Egyetem munkatársai a *Claris FileMaker* nevű szoftverplatformot választották, melyet feladatorientáltan testre szabtak, ezzel minimalizálni tudták a rekordok behasonlításához szükséges emberi erőforrást. A munka 2020 februárjában indult, majd a COVID-19-es karantén miatt márciustól a munkatársak távmunkában folytatták. A DAHR adatbázisában több mint 20 ezer rekordot kapcsoltak össze LCNAF- és VIAF ID-rekordokkal, több mint 8 ezret Wikidatával és MusicBrainzzel, kisebb számú egyéb forrásból származó behasonlítással egyetemben. A kapcsolt adatok valós idejű aratása során a szerkesztők havonta közel 300 új nevet adnak hozzá a DAHR-hoz. A következő lépésként a DAHR művészazonosítók URI-jait kívánják összekötni Wikidata-oldalakkal, majd folytatni az aratást a VIAF-ban, valamint a DAHR URI-kat hozzárendelni a MusicBrain, Discogs és egyéb adatbázisok rekordjaihoz.

A második gyakorlati példát *Huda Khan* mutatta be a Cornell Egyetem munkatársaként. Az egyetemi könyvtár kapcsolt adatokkal növelte a könyvtári discovery szolgáltatás hatékonyságát az LD4P2 program keretében. A hagyományos, „ismert példányon" alapuló könyvtári katalogizálást felváltja az *open-ended discovery* (tág körűen értelmezhető felfedezés) koncepciója. Ebben az esetben a *dokumentumközpontúság*ot felváltja az *entitásközpontúság,* amely magában foglalja a releváns kapcsolatokat és a katalógusra visszautaló külső adatforrásból származó kapcsolt adatokat egyaránt. Ez a megoldás az eddigieknél jóval relevánsabb találatokkal kecsegteti a használót. Az entitások központi elemei a név-authority és az autoritatív tárgyszavak. A külső adatforrások lényegében szerzői és tárgyi authoritykhez tartozó URI-k olyan szolgáltatásoktól, mint például az OCLC és a Wikidata. Az LD4P2 – és egyúttal a Cornell Egyetem megoldásának – fókuszában a tervezés, a fejlesz-

tés és a prototípusok kiértékelésének hármasa áll. Az entitásközpontúság lehetővé teszi a különböző formátumú művek tartalmi összekötését. A Stanford Egyetem könyvtárának katalógusában például Beethoven esetében a zeneszerző híres művei lejátszhatók a katalógusban. Az entitások és kapcsolatok mentén rendszerezett katalógus a böngészést is egyszerűsíti a kategóriák és kapcsolataik szerinti navigáció révén. A Cornell Egyetem könyvtárának katalógusa támogatja a tárgyszavak, a szerzői idővonal, a régió és a tárgykör szerinti böngészést is. A vonatkozó ismérveket vizuálisan megjelenítő idővonalas böngészés érdekessé teszi a könyvtári tartalmak felfedezését.

Az entitások efféle ábrázolását a Wikidata születési és halálozási adatai, illetve a tevékenység kezdetére és végére vonatkozó adatai teszik lehetővé. További előrelépés az újfajta katalogizálási megközelítésben, hogy a rendszer javaslatokat tesz a használók számára a vonatkozó keresési eredmények alapján. Ezt az *Annif Rest API* teszi lehetővé, amelyet a Kongresszusi Könyvtár tárgyszavai alapján hoztak létre. A kulcsszavas keresést authoritylekérdező szolgáltatás köti össze a Kongresszusi Könyvtár forrásaival. További funkció a személyek, helyek, tárgyszavak és műfajok keresésére szolgáló automatikusan kitöltődő javaslatok lehetősége. Ez a többi gépi javaslathoz hasonlóan a katalógusfelület bal oldalán lévő, erre szolgáló ablakban változik dinamikusan, a felhasználók aktuális keresésének függvényében. Ennek adatforrását szintén a Kongresszusi Könyvtár és a Wikidata biztosítja. A használói visszajelzések nyomán kijelenthető, hogy ez kedvelt és könnyen érthető szolgáltatás. A felhasználói felület fejlesztésébe laikusokat (egyetemi hallgatókat, fiatal kutatókat) is bevonnak a minél gazdagabb végfelhasználói élmény érdekében. A tárgyszavazásért felelős szakértők számára specializált munkafolyamatokat alakítottak ki, amelyek keretében külön segítség nélkül áttekinthetik az entitások közötti kapcsolatokat, ugyanis az interdiszciplináris kutatás jellemzője, hogy a kutató legalább az egyik területen tapasztalatlannak számít. A cornelli példában is nagy kihívást és egyben fejlesztési lehetőséget jelentenek az adatok teljességét és interoperabilitását, az adatbázis terhelhetőségét, az aggregációt, a rendszerfrissítést, illetve az akadálymentességet érintő kérdések. Az URI-k feltüntetése a katalógusban szintén remek lehetőség az entitásközpontú katalogizálás és a kapcsolt adatok előnyeinek kiaknázására, például a *https://schema.org/* adaptálásával. A Cornell Egyetem Könyvtára jelenleg is résztvevője a most folyó *LD4P3* kezdeményezésnek, amely az *LD4P2* implementációjának új megoldásait vagy kibővítését, illetve a projekt eddig keletkezett és kurrens ötleteinek gyakorlati megvalósítását tűzte ki célul. A munka alapját a kapcsolt adatok előállításának ciklusa képezi, amelynek összetevői a prototípusok, a discovery rendszerek, a Blacklight-környezet és azok közössége, valamint a kapcsoltadat-források. Jelenleg egy entitás-specifikus dashboardon, illetve a katalógusokban fellelhető források adatközi megtekinthetőségén (cross-data source view) dolgoznak. Ez például a tárgyszavak szűkítését, bővítését, illetve az időbeli és földrajzi információk láthatóságát foglalja magában.

Utolsó előadóként *Jeff Keith Mixter,* az OCLC munkatársa ismertette projektjüket, amelynek keretében az *IIIF* (International Image Interoperability Framework – Nemzetközi Kép-interoperabilitási Keretrendszer) és a Wikibase (utóbbi alapeleme a Wikidata infrastruktúrájának is) félig strukturált adatainak használatával gondoznak és osztanak meg a kulturális örökség tárgyában anyagokat a *CONTENTdm*-en, az OCLC digitális repozitórium-szolgáltatásán keresztül. Mindez tömeges aggregációval valósul meg. Az OCLC munkatársai a prototípust az IIIF *Change Discovery API* alkalmazásával fejlesztették, és mintegy 13 millió CONTENTdm-képelemet gondoznak vele. Utóbbiak metaadatait learatták, majd egy behasonlítást követően a karakterláncokban reprezentált tárgyszavakat a kapcsolt adatok URI-jaihoz rendelték (kizárólag Dublin Core-t használva). Jeff Keith Mixter a bevezetőt követően bemutatta az IIIF *Explorer* felületét, amely tárgyi facetták szerint rendszerezi a learatott képi tartalmat, feltüntetve a tárgyi megjelölést, a közreműködőket, a helyeket, a létrehozókat és hasonló entitásokat. Az online képnéző a nagy felbontású képeket formátumtól függetlenül azonos módon jeleníti meg a hozzájuk társított metaadatokkal és kapcsolódó tartalmakkal egyetemben. Mixter szerint a szolgáltatás legnagyobb előnye, hogy a segítségével váratlan dolgokat lehet felfedezni váratlan helyeken. Egy másik OCLC-projektben öt CONTENTdm-felhasználó könyvtár intézményenként három gyűjtemény metaadatait manuális munkával nézte át, feleltette meg és hasonlította be, majd ezeket az adatokat importálták Wikibase-példányként a menedzsment-munkacsoport számára. Ezt követően az OCLC készített egy prototípusként szolgáló discovery eszközt ezen példányok keresésére és felfedezésére; a felület azonos az IIIF Explorerével, ám jóval strukturáltabb metaadatok társulnak a repozitóriumban tárolt képek mellé. Mixter *Louis Armstrong* fotóján keresztül szemléltette mindkét projektet: míg az aggregált metaadatoknál csupán néhány tárgyszó kapcsolódott a fotóhoz, a könyvtári adatfeldolgozás eredményeképpen rendkívül részletes információkat kaphatunk ugyanazon dokumentum mellé, kitérve annak formátumára, tartalmára, példának okáért Armstrong rövid életrajzára is. Ezt a Wikibase hatalmas adatmennyisége teszi lehetővé, így egy adott kép bizonyos részletei is kiemelhetők és metaadatolhatók, amennyiben lehetséges e részinformációknak a kapcsolt adatokkal rendelkező tárgyszavak alapján való importálása.

Azonban hiába a technológia biztosította tömeges aggregáció, a minőségi metaadatolás továbbra is nélkülözhetetlenné teszi az emberi beavatkozást. Ugyanakkor a Wikibase hatalmas és rugalmas infrastruktúrája megkönnyíti a strukturált adatok létrehozását, kezelését és gondozását, és számos további, még kiaknázatlan lehetőség is kínálkozik a kulturális örökséget megőrző források metaadatolásában. A jövőbeli kutatás kiterjedhet a tárgyi szaktudás mentén történő algoritmikus rekordkonverzió kiegyensúlyozásának vizsgálatára, a kontextuális és leíró metaadatok szétválasztásának módszereire, valamint arra, hogy a végfelhasználói alkalmazások esetén hogyan lehetne új kontextuális metaadatokat hasznosítani.

*

A konferencia utolsó napján rövid, pár perces villámelőadások keretében mutattak be további témába vágó projekteket és jó gyakorlatokat. A bécsi KDZ-től (KDZ Zentrum für Verwaltungs Forschung – KDZ Közigazgatási Kutatóközpont) *Bernard Krabina* mutatta be az intézmény *Semantic MediaWiki* projektjét, amely egy könnyebben – programozási tudás nélkül is – használható „svájci bicskaként" foglal magába különféle nyílt forráskódú szemantikus webalkalmazásokat, segédeszközöket.

Anna Lionetti a már érintőlegesen említett *Share-VDE* (Virtual Discovery Environment) kezdeményezést ismertette. A Share-VDE keretében MARC-rekordokat alakítanak át BIBFRAME-alapú leírásokká, majd kapcsolt adatokká, amelyeket végül a közös webes discovery felületen tesznek közzé. A partnerkönyvtárak a platformon saját nevükkel és arculati jegyeikkel jelenhetnek meg. Mindezek mellett a kapcsolt adatok közös szerkesztésére is lehetőség kínálkozik, valamint megoldott az entitásmodellek interoperabilitása is.

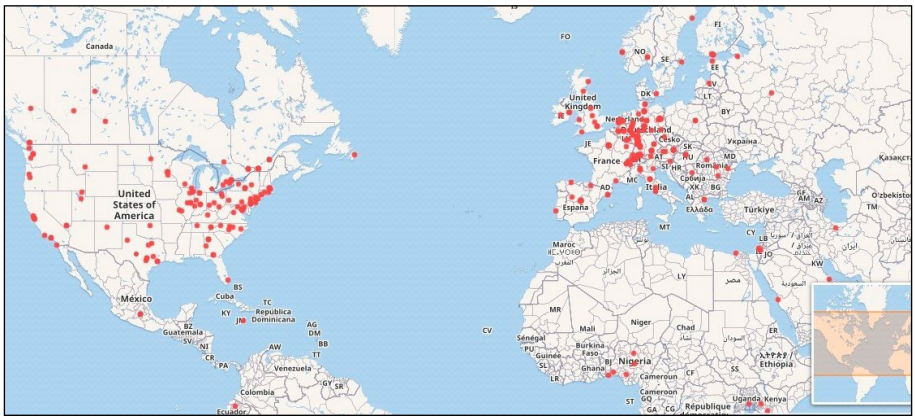
Anchalee (Joy) Panigabutra-Robert a Tennessee Egyetem könyvtárának egyik Wikidata-projektjét ismertette, melynek lényege a 13. századi arab feltalálóról, *Izmail al-Dzsazari*ról szóló szócikk szemantikus összekötése az automaták témakörével.

Fabian Steeg a *W3C entitás-adatgazdagítás közösségi csoportjá*ról (W3C Entity Reconciliation CG) tartott rövid ismertetőt. Alapvető eszközük az *OpenRefine,* amely remekül bevált adatkezelői munkafolyamatokra, köztük az adatgazdagításra is. A munkafolyamat lényege a saját adatok ellátása az ugyanazon adatoknak/entitásoknak az adatszolgáltatók által nyújtott azonosítóival, authority fájljaival. A munka során adatmezők is importálhatók az authority adatokból, így a saját adatok szabványos módon gazdagíthatók, bővíthetők. A közösségi csoport célja, hogy standardként szolgáljanak a webes adatgazdagítás, illetve behasonlítás témakörében, és ehhez célirányos eszközöket készítsenek.

Jakob Voß a német *K10plus* közös katalógusban fellelhető állományadatok Wikidatával való összekapcsolását mutatta be. A katalógusrekordok megfeleltethetők a vonatkozó Wikidata-bejegyzésekkel, így a közös katalógusban szereplő dokumentumoknak szemantikus réteget nyújtanak, valamint további információkat szolgáltatnak. A Wikidata-elemek nagyszámú learatását csak gépi úton, programozott botokkal lehet hatékonyan elvégezni, így a németek is ezt az utat választották.

A „lightning talk" szekció zárásaként *Joachim Neubert* prezentálta a SWIB20 online megjelent résztvevőinek regisztrációs adatai alapján készített *Wikidata-térképet.* Az adatgazdagítás szintén az OpenRefine szoftver használatával történt, elsősorban az országok és az intézménynevek szerint.

Összegzésként elmondható, hogy a SWIB20 konferencia pontos nemzetközi pillanatképet szolgáltatott a könyvtárak és a kapcsolt adatok jelenlegi helyzetéről. Az előadások mellett a konferencia szervezői által biztosított online fórumfelület alkalmat adott a tapasztalatok informális megosztására is.



Joachim Neubert Wikidata-térképének részlete a SWIB20 részvevőiről
(Forrás: https://zbw.eu/labs/en/blog/building-the-swib20-participants-map)