Hagyományok és kihívások VIII. Országos Könyvtárszakmai Nap, 2019.
Szerk. Szabó Panna, Székelyné Török Tünde.
Budapest, ELTE Egyetemi Könyvtár és Levéltár. 2020. 195–201.

Az ETO-n át az ügyfélhez

A könyvtári tartalmi feltárás üzleti hasznosulásának egy szép példája

Bevezetés

A 19. század végén megálmodott világkatalógus nyelvfüggetlen tartalmi visszakereső eszközeként létrehozott Egyetemes Tizedes Osztályozás (ETO) a 20. és a 21. századi technológiáknak hála új alkalmazási területeket hódíthat meg, túllépve a könyvtári felhasználás bennfentes körein. A webáruházak katalógusaiban alkalmazott elnagyolt tematikus besorolásokkal szemben a visszakeresési potenciál jelentős emelkedéséhez vezethet az ETO-jelzetek megfeleltetése a nemzetközi online könyvkereskedelemben elterjedt hierarchikus tárgyszórendszer (Thema[1]) fogalmaival. A gazdag és mély tematikai feltártság, a tárgyszóhierarchiában való közlekedés (a szűkebb fogalomtól a tágabbig, illetve a tágabbtól a szűkebbig terjedő tematikus böngészés) lehetővé teszi a speciális keresési igényeknek megfelelő, avagy az érdeklődés szűkebb mezsgyéjén található műveknek a piacra való visszatalálását.

Kiadó és könyvtár mint együttműködő partnerek

A 2010-es évek elején az Országos Széchényi Könyvtár (OSzK) két olyan nemzetközi projektben is részt vett, amelyeknek a célkitűzései között szerepelt a kereskedelmi (kiadói, jogkezelői, forgalmazói) és könyvtári informatikai rendszerek összeillesztése, adatcsere kialakítása a rendszerek között a könyvipari adatfolyamok optimalizálása érdekében. A projektpartnerek között külön említést érdemel az EdItEUR[2] nemzetközi cég, amely komoly szerepet játszik a könyvipari adatcsereszabványok és a nemzetközi permanens azonosító rendszerek fejlesztése, elterjesztése, gondozása terén. A *Linked Heritage*[3] és *ARROW*[4] projektekben a szerepük a könyvtári és könyvpiaci rendszerek együttműködésének előmozdítása volt, mindenekelőtt a cég által fejlesztett ONIX metaadatséma-szabványra alapozva.

A könyvipari értéklánc piaci és nonprofit (állami költségvetési) szereplői közötti együttműködésnek régi hagyománya van a nyugati világban. Néhány példa a könyvkereskedelmet a könyvtári metaadatszolgáltatással összekötő, piaci alapon működő *bibliográfiai központ*okra: *Bibliographic Data Services*

[1] Thema v1.3: https://www.editeur.org/151/Thema (2019. szeptember 25.)

[2] EdItEUR. https://www.editeur.org (2019. szeptember 25.)

[3] Linked Heritage. http://linkedheritage.eu (2019. szeptember 25.)

[4] Az ARROW projektről bővebben l. Dancs Szabolcs: „A digitális felvilágosodás felé": az ARROW program és az OSZK – digitalizálás, szerzői jog, innovatív megoldások. = Könyvtári Figyelő, 59. évf. 2013. 3. sz. 465–484. p.

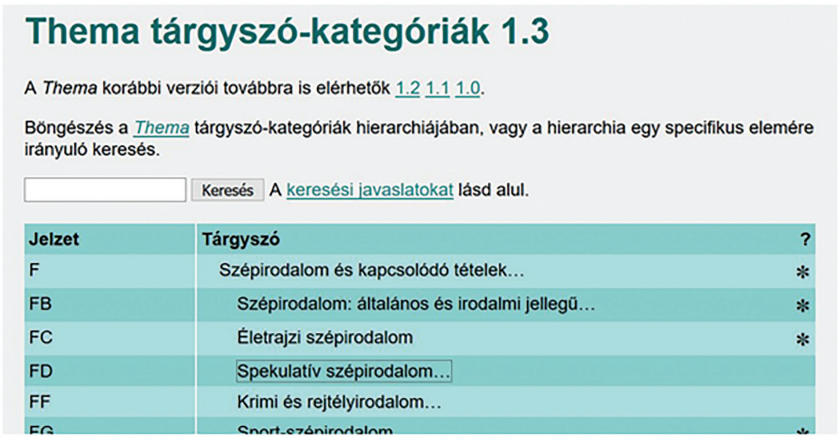
https://doi.org/10.21862/HagyKihiv_2019.195

(Egyesült Királyság), *Casalini Libri* (Olaszország), *Dansk BiblioteksCenter* (Dánia). A nemzeti International Standard Book Number (ISBN) ügynökségek is gyakorta működnek piaci alapon, és töltenek be hasonló hub-szerepet. Az említett EDItEUR cég régi jelenlétére és az ISBN-nel való komoly összefonódására utal, hogy maga a nemzetközi szabvány *(MSZ ISO 2108:2019 Információ és dokumentáció. Nemzetközi szabványos könyvazonosító szám (ISBN))* is tartalmazza a nemzetközi ügynökségnek való adatátadást támogató ONIX-metaadat-elemeket.

Az együttműködés előnye könyvtári oldalon többek között a piacra való nagyobb rálátásban mutatkozik meg, ami magában foglalja a nemzeti könyvtárba be nem érkezett kötelespéldányok könnyebb beazonosíthatóságát. Az elérhetőségi információk egyrészt a jogtisztázáshoz, így az árva művek és a kereskedelmi forgalomból kikerült művek azonosításához, a digitális tartalmak szolgáltatásához szükséges adatokat jelentik, másrészt a kiadványok beszerzéséhez vagy az e-kiadványok licenceléséhez szükséges információkat. A piaci szféra mindenekelőtt a visszakeresési potenciál látványos növekedése révén profitál a kooperációból, ezt pedig a könyvtárak által készített minőségi metaadatok teszik lehetővé, beleértve a tartalmi feltárás elemeit. És itt térhetünk rá közelebbi témánkra, az ETO-alapú feltárás üzleti hasznosulására.

ETO-ból Thema

A Thema (korábbi nevén BIC) a nemzetközi online könyvkereskedelemben elterjedt hierarchikus tárgyszórendszer, amit ugyancsak az EDItEUR gondoz. A magyar változat *Bánhegyi Zsolt* munkája, és online elérhető. Az ETO → Thema konverzió gondolata a már említett ARROW projekt során merült fel a Magyar BiP Kft. részéről.



1. ábra. Részlet a Thema tárgyszó-szótár online böngészhető verziójának magyar nyelvű felületéről

ETO-Thema megfeleltetés

A könyvtári katalógusokból származó ETO-jelzetek megfeleltetése a Thema tárgyszókategóriáknak, ha nem is könnyen, de automatizálható. Számos könyvtár használta, és használja integrált könyvtári rendszereket építve a tartalmi feltárás során ma is az ETO-számokat. A 080-as MARC tagbe kerülő ETO-szám természetesen más módon van jelen a különféle könyvtári katalógusokban, és az azokra épülő nyilvános webes keresőkben. Egyes rendszerekben csak rövid ETO-jelzeteket használnak, és ezek a jelzetek úgy viselkednek, mint egy tezaurusz tárgyszavai, tehát egy-egy jelzethez bibliográfiai rekordok százai tartoznak. Más esetekben a részletesen kibontott hosszú ETO-jelzetek szinte minden mű esetében különböznek. Az előbbi esetben az ETO-jelzet kiosztást általában érdemes értéklistával támogatni. Az ETO-jelzetek könyvtári katalóguson belüli jellege jelentősen különbözik az általános gyűjtőkörű és a szakkönyvtárak esetében is, illetve csekélyebb az ETO-jelzet kiosztás szerepe a szépirodalmi művek esetében. Mindezeket figyelembe véve az első, az ETO-jelzet és a Thema tárgyszókategória automatikus megfeleltetésének kiépítésére szolgáló gyakorlótanuló adatbázis alapjául mindenképpen hosszú, részletes ETO-jelzeteket használó, általános gyűjtőkörű, és nem túlnyomóan szépirodalmat tartalmazó könyvtári számítógépes katalógust volt érdemes választani. Emellett – mint minden tanulni képes és tanítható alkalmazás esetében – itt is fontos volt, hogy az adatbázis nagy legyen, bibliográfiai rekordok százezreit tartalmazza.

A megfeleltetést végző programot a Magyar BiP Kft. partnereként a Monguz Információtechnológiai Kft. készítette el. A tanuló adatbázis, amely a Magyar Országos Közös Katalógus (MOKKA) számítógépes katalógusából származott, a fejlesztés kezdeténél mintegy félmillió, összesen 230 000 féle különböző ETO-jelzettel ellátott, döntően monografikus jellegű, különféle nyelveken, de elsősorban magyarul íródott kiadványokat leíró bibliográfiai rekordot tartalmazott. Ezt a kétszázezernél többféle ETO-jelzetet kellett megfeleltetni a Thema tárgyszórendszer magyar változatának, illetve a magyar változat, az ugyancsak a változatot gondozó Bánhegyi Zsolt által előkészített ETO-sémákat is tartalmazó adatbázisának. A táblázat közel 5000 Thema-kód – ETO-séma párt tartalmazott. Egy Thema-kód több ETO–sémához, és hasonlóképpen egy ETO-séma több Thema-kódhoz is kapcsolódhatott. Az alábbi ábra a táblázat fejlécét és egy sorát mutatja be:

Code	English Heading	Notes	Related (see also)	Hungarian Language Heading	ETO
YNGL	Children's / Teenage general interest: Libraries, museums, schools			Gyermek- / serdülőkorúak: általános érdeklődés: könyvtárak, múzeumok, iskolák	02... (02.053.2)… 069… (02.053.2)… 3 7 3 … / 3 7 8 . . . (02.053.2)…

1. táblázat. A Thema-kód és az ETO-séma megfeleltetése

Az első oszlopban látható Thema-kódot kellett tudnia a Monguz Kft. által fejlesztett programnak automatikusan hozzárendelnie egy bibliográfiai rekordhoz, annak 080-as MARC mezője, vagyis az ETO-jelzete alapján. Tehát minden olyan bibliográfiai rekordnak, amelyre igaz, hogy a 080-as mezőjének tartalma 02-vel kezdődik, és van benne 02.053.2 érték, meg kellett kapnia a YNGL kódot. Ugyanezt a kódot megkaphatták más ETO-számmal rendelkező bibliográfiai rekordok, például a 373…/378... adatrészt tartalmazó ETO-számmal rendelkező bibliográfiai rekordok is, és a 373…/378... séma, ha más sorában is szerepelt a táblázatnak, akkor a bibliográfiai rekord más Thema-kódokat is kaphatott. (A ... jel a megfeleltető táblázatban a csonkolás jele.) Tehát valójában nem az ETO-jelzetek kaptak Thema-kódot, hanem a bibliográfiai rekordok, ETO-jelzetüknek megfelelően lettek egy vagy több Thema-kategóriához rendelve. A munka eredményeképpen nem maradhatott a tanuló adatbázisban olyan bibliográfiai rekord, amely nem kapott egyetlen Thema-kódot sem.

Elsőként tehát az ETO-sémákat kellett olyan formára átalakítani, hogy a Postgres SQL adatbáziskezelővel kezelt adatbázisban az adatbázis kezelő eszközeivel minden bibliográfiai rekord ETO-jelzetéhez legyen találat. Ehhez a táblázatban levő sémákból az SQL parancsnyelv által értelmezhető maszkokat kellett létrehozni. Ezt a táblázatban levő sémák szintaktikai átalakításával, gépi eljárással néhány száz esetet leszámítva, el lehetett érni. A bibliográfiai rekordok által reprezentált művek ETO-jelzete, a hozzájuk rendelt séma és Thema-kódok táblázata ezután ellenőrzésre került. Az ellenőrzést szúrópróba-szerűen végezték el, először a Monguz Kft. munkatársai, majd a partnerek által felkért tesztelők. Miután az ellenőrzéseken talált hibák alapján a maszkok kialakításának módszere tökéletesedett, a maszkokból már létre lehetett hozni a JSON programnyelv (melyen a Monguz Kft. által fejlesztett alkalmazás íródott) számára értelmezhető reguláris kifejezéseket.

Kód	Géppel formázott maszk	Reguláris kifejezés alap	Reguláris kifejezés
YNGL	02... (02.053.2)…	02.* \ (02\.053\.2\).*	{"bicCode":"YNGL", "eto":"02.*\\(02\\.053\\.2\\).*" },
YNGL	069… (02.053.2)…	069.* \ (02\.053\.2\).*	{"bicCode":"YNGL", "eto":"069.*\\(02\\.053\\.2\\).*" },
YNGL	373…/378... (02.053.2)…	37[3-8].* \ (02\.053\.2\).*	{"bicCode":"YNGL", "eto":"37[3-8].*\\(02\\.053\\.2\\).*"}

2. táblázat. Gépi /kézi átalakítás eredménye

A tesztelés következő lépéseként a program az újabb Thema-kódokat még nem tartalmazó adatbázist a reguláris kifejezések alapján átalakította, és a bibliográfiai rekordokhoz Thema-kódokat rendelt. Miután a szúrópróba-szerű ellenőrzés a program működését megfelelőnek találta, újabb és újabb rekordokkal bővítették az immár félmilliósnál nagyobbra duzzadt adatbázist, melyet a program emberi beavatkozás nélkül Thema-kódokkal lát el. Amennyiben a program egy 080-as mezőhöz nem találna Thema-kódot, azt ellenőrzésre kiadja. Ekkor egy embernek kell megállapítania a hiány okát, és ha szükséges, a kódok listáját ki kell bővítenie, vagy új reguláris kifejezést kell hozzáadnia egy már meglévő Thema-kódhoz. A kézi beavatkozásra a munka kezdetén is szükség volt, mert a túl komplex ETO-sémákból nem lehetett automatikusan géppel értelmezhető maszkot készíteni, illetve a katalógusban tárolt ETO-jelzeteket is le kellett egyszerűsíteni a tanuló környezet kialakításakor. Az előbbire, tehát az automatikusan géppel olvasható maszkká nem átalakítható sémára jó példa az alábbi eset: „66…/68…kiv.681.3… „

Itt a maszkot és abból a reguláris kifejezést kézzel kellett létrehozni. Amennyiben a Thema-kategóriák táblázata vagy a Thema-kategóriák ETO-sémákkal való megfeleltetése változik, a változások között lehetnek hasonló, géppel nem egyértelműen átalakítható sémák, tehát az új kód- és sémaértékeket mindig embernek kell ellenőriznie. Természetesen lehetne olyan algoritmust alkotni, amely ilyen esetekben is kiváltja az emberi ellenőrzést, ennek megtervezése azonban túl hosszadalmas, és ezért az alkalmazás ilyen szintű automatizálása túl költséges lenne. Tekintettel arra, hogy sem a Thema-kategóriák, sem az ETO-jelzetek nem változnak nagy ütemben, és a változások jó része is az alapvető gépi eljárással kezelhető, ezért ilyen fajta automatizálás a projektben egyelőre nem indokolt.

Ugyancsak emberi ellenőrzést igényelt az, hogy nem maradhatott Thema-kód ETO-maszk, illetve főként ETO-maszk Thema-kategória nélkül, és nem volt szerencsés, ha egy Thema-kategóriához túl sok ETO-maszk került. Az óriási, több százezer soros megfeleltető adatbázist ebben a tekintetben az SQL adatbáziskezelő eszközeivel ellenőrizni kellett. Tekintettel arra, hogy az ETO-maszk Thema-kód

párosítás a program működése során kiállta a próbát, efféle ellenőrzésekre már csak akkor lesz szükség, ha a program által már Thema-kódokkal ellátott adatbázis jelentősen megnövekszik, vagy ha a Thema-kategóriák táblázata, illetve a Thema ETO-séma megfeleltetést tartalmazó táblázat jelentősen átalakul. E tanulmány szerzői mindamellett bíznak benne, hogy a létrehozott módszertan és a kidolgozott algoritmus kiállja az idők próbáját, és tartósan képes lesz közvetíteni a könyvkereskedelem számára a könyvtári tartalmi és formai feltárás eredményeit.

Összegzés

Számítástechnikai eszközökkel körülvett világunkról is elmondható, hogy az általunk létrehozott információtömeg legnagyobb és legmaradandóbb értéke változatlanul az emberi értelem, és különösen a nagy munkával létrejött hozzáadott szellemi érték. Igaz ez a tartalmi és formai feltárás során felépített számítógépes könyvtári katalógusokra is, ahol az évtizedek alatt létrehozott sok ETO-jelzet éppen a benne levő szellemi többletnek köszönhetően a jelen projektben ismét hasznosításra kerülhet.

Rezümé

Az Egyetemes Tizedes Osztályozás a 20. és a 21. századi technológiáknak hála, új alkalmazási területeket hódíthat meg, túllépve a könyvtári felhasználás bennfentes körein. A webáruházak katalógusaiban alkalmazott elnagyolt tematikus besorolásokkal szemben a visszakeresési potenciál jelentős emelkedéséhez vezethet az ETO-jelzetek megfeleltetése a nemzetközi online könyvkereskedelemben elterjedt hierarchikus tárgyszórendszer (Thema) fogalmaival.

A Thema-kódokat és az ETO-jelzeteket más-más szervezet más-más igények mentén hozta létre, de a két kategóriarendszer összhangba hozható. A megfeleltetést tartalmazó táblázat alapján számítógéppel futtatható algoritmus jött létre. Ez a kiinduló adatbázist képes volt ellátni Thema-kódokkal, és remélhető, hogy az adatbázis bővülése során alapvetően automatizáltan történhet majd az új bibliográfiai rekordok Thema-kódokkal való kiegészítése.

Via UDC to the Client

A Nice Example for Utilizing Libraries' Subject Descriptions in the Business Area

Due to technologies developed during the last two centuries Universal Decimal Classification (UDC) can be applied in new fields and go beyond usefulness limited to libraries. The retrieval potential of webshop catalogues can be raised through mapping UDC numbers to terms of the international hierarchical thesaurus Thema, which has been spread

in the book industry, instead of applying the not-too-detailed topical classifications used nowadays in the area for retrieval.

The lists of Thema codes and UDC numbers were created by various corporates, according to different demands, but from the two structures a coherent one can be built. The list of matched Thema codes and UDC numbers can be prepared for use in computer algorithms. The application, based on this table, could add Thema codes to the bibliographic records of the initial database. Hopefully the application can do this work in the future with the new records of the database almost completely in an automatized way.

DANCS SZABOLCS
mb. igazgató
OSZK Országos Könyvtári Szolgáltatások Igazgatósága

SIMON ANDRÁS
ügyfélmenedzser
Monguz Információtechnológiai Kft.