UTCA: egy fejlesztésben lévő „közösségi katalógus”

Jó tízéves múltra tekint vissza a hazai közös katalogizálás története. 2005-ben egy összefoglaló tanulmánykötet is megjelent a témáról „Közös katalogizálás Magyarországon” címmel (ISBN 963-9364-60-6) a soproni *Nyugat-magyarországi Egyetem* kiadásában. A kötetből – és saját kutakodásaimból is – az derül ki, hogy a közös katalogizálásnak a szakma által ismert előnyeit a gyakorlatban csak részlegesen sikerül kihasználni. Néhány éve, amikor mint „külsős” – olvasó és informatikus – megismerkedtem a *MOKKÁ*-val, a nagy adatbázisban rejlő tudás felhasználásának és megjelenítésének további lehetőségei foglalkoztattak – úgy gondoltam, ebből többet lehetne kihozni. A szakmai nehézségeket, és a több forrás által említett, e témában folyó vitákat akkoriban még nem ismertem eléggé. Időközben az *ELTE* könyvtár szakos hallgatójaként tájékozottabb és szakmabeli lettem. A lelkesedésem megmaradt, így kezdtem kollégámmal, *Csámer Iván*nal az „UTCA”, az „Univerzális Tartalomfeltáró és Csoportosító Alkalmazás” fejlesztésébe. A projekt célja, hogy a könyvtári katalogizálás és tartalomfeltárás, valamint más, hasonló jellegű tevékenységekkel épülő adatbázisokat minél egységesebb tudásbázissá alakítsa, s hogy azt – ismert és újszerű szolgáltatások keretében – minél hatékonyabban állítsa az olvasók és a könyvtárak szolgálatába. E beszámoló az előzmények, hasonló projektek eredményein és nehézségein keresztül mutatja be saját elképzeléseinket, újszerű eszközeinket.

Nem kétséges, a közös katalogizálás sok szempontból hasznos. Az olvasó egyszerre kereshet sok könyvtár katalógusában, sőt, új könyvtárakat is felfedezhet. A könyvtárakban a közös katalogizálás a katalogizáló munkát, az adatok egységesítését és a könyvtárközi kölcsönzést segíti – így fontos költségcsökkentő eszköz is lehet.

Az interneten ugyan sokféle információ elérhető, mégis csupán a könnyű elérhetőség az, amiben a „net” felülmúlja a könyvtárakat, a könyvekben felhalmozott tudásnak és szakértelemnek csak töredékét tartalmazza. Egy könnyen használható katalógus, amely hatékonyan segíti az olvasókat a számukra keresett művek megtalálásában, fontos ahhoz, hogy a felhalmozott tudás egészére ráirányítsuk a figyelmet. Szeretnénk a katalógusok tartalmát jobban kihasználni, és a katalógust egyúttal közösségi térré is tenni, ahol a könyvtárosok és az olvasók újszerű együttműködés keretében segíthetik egymást az eligazodásban. Mint látni fogjuk, ennek érdekében olyan technikai megoldásokat alkalmazunk, amelyek a jelenlegi közös katalógusokhoz képest jobban hasznosítják a könyvtárosok szakértelmét, s amelyekhez a nagyobb és a kisebb könyvtárak munkatársai egyaránt aktívan hozzájárulhatnak. A közös katalógus új funkciója lehet, hogy információkat adjon az új könyvekről és a fölöspéldányokról is. Gondoljuk csak meg, menynyivel egyszerűbb lenne, ha a fölöspéldány-jegyzékeket egy helyen kereshetnénk, ugyanolyan OPAC-ban, mint a kölcsönözhető könyveket, és ha automatikusan bekerülnének az adatbázisba a könyvtárunk „törlendő” státusú könyvei is!

A közös katalógus a tartalmi feltárás eszközeinek népszerűsítésére (ETO, tezaurusz), vagy a könyvtári adatok névterekkel, geográfiai adatbázisokkal való összekapcsolására is alkalmas lehet. Több információt adhatunk így a szerzőkről, vagy újszerű módon – például térképen – jeleníthetünk meg könyvtárakat és helyi jelentőségű kiadványokat. Az olvasói felület olyan új szolgáltatásoknak adhat otthont, melyeket egy-egy könyvtárban nem érdemes, nem logikus, vagy nem költséghatékony megvalósítani. Létrehozható például az olvasó saját könyveinek, kedvenc szerzőinek, korábbi kereséseinek jegyzéke, az olvasó személyes megjegyzést, véleményt írhat a könyvekhez, utalhat más hasonló könyvekre, ajánlót küldhet a barátainak. Az ilyen szolgáltatások segíthetik a személyre szabott, kényelmes és hatékony információkeresést, növelik a könyvtári információk használhatóságát. Sok hasonló tervünk van az UTCA projekttel, de ebben a beszámolóban inkább a technikai részletekről lesz szó.

A jelenlegi közös katalógusokat – *MOKKA, ODR, Szikla, Szirén, TextLib, HunKat és Theka* – megvizsgálva észrevehető, hogy jellemzően hasonló, a felhasználók életét igencsak megnehezítő problémákkal küzdenek. Így van ez annak ellenére, hogy az egyes katalógusok eltérő feltételek között alakultak ki. Véleményünk szerint a gondok egyik fő oka, hogy az informatikusok és a könyvtárosok nem mindig tudják a két szakterület követelményeit, lehetőségeit és erősségeit megfelelően egyeztetni. Megállapíthatjuk, hogy a közös katalógusok keresőfelületei túlságosan „technikaiak", nem eléggé segítőkészek. Több lekérdező felület érzékeny a nevek írásmódjára, és csak a szabályos formát fogadja el: *Robert Merle-t* csak *Merle, Robert*-ként találja meg a szegedi MOKKA-tükör, a Szirén, a TextLib és a Theka is. A MOKKA és az ODR *Kollár Éva*ra keresve olyan találatokat is megjelenít, amelyekhez neki semmi köze – csak egy másik *Kollár*nak, és egy másik *Éva*nak. Tárgyi melléktételként felvett nevekre nem mindenütt lehet célzottan keresni, így *Arany János* művei keverednek a róla írottakkal. Gondot jelent a címek keresése is: a MOKKA, az ODR, a HunKat és a Theka a cím bármelyik szavára tud keresni (pl. az *elmék*re az *Állati elmék*ből), a Szikla, a Szirén és a TextLib viszont csak akkor, ha a külön erre szolgáló „cím szava" mezőt használjuk a cím mező helyett – feleslegesen nehezítve a keresőkérdés összeállítását.

A kényelmetlenségek elidegenítik az olvasót a katalógustól, pedig a fenti gondok informatikailag viszonylag egyszerűen kiküszöbölhetőek lennének. A névindex természetes sorrendű névváltozatokkal való kiegészítésével, okosabb indexelő szoftver használatával, vagy egyszerű automatizmusok (automatikus újrakeresés, indexváltás stb.) beépítésével. Az ijesztő nevű „csonkolás" helyett nyelvészeti programok alkalmazásával segíthetjük a felhasználókat. A katalógusok a webes megjelenítés lehetőségeivel is csak szerényen élnek, például nem használnak betűstílusokat, színeket, táblázatos megjelenítést, amelyek könnyebben olvashatóvá tennék a találati listákat. Néhol az egyes adatelemeket a bibliográfiai leírás szabályainak megfelelő írásjelekkel (központozással) jelenítik meg, ami természetes a könyvtárosoknak, de felesleges az olvasóknak. Helyzetérzékeny segítséget egyik katalógus sem ad.

A legfőbb probléma a fentiek mellett az, hogy a jelenlegi közös katalógusok rengeteg duplumrekordot tartalmaznak. Zavaró, ha egy találati listában tucatnyi felesleges kattintásra van szükség az összes releváns tétel átnézéséhez. Az olvasónak az is duplum, ha egy könyv öt egymás utáni változatlan kiadása külön-külön jelenik meg, hiszen ez számára lényegtelen különbség. Emiatt még azok a keresők is nehézkessé válnak, amelyek egyazon integrált könyvtári rendszer alkalmazóit kötik össze, s ezért elvileg előnyösebb helyzetben vannak a duplumok ügyében.

Az említett nehézségek főként az olvasókat és a webes katalógust használó könyvtárosokat érintik, de emellett vannak gondok, amelyek a könyvtárakat intézményként sújtják. A közös katalogizálásban való részvétel nekik hasznos ugyan, de egyúttal terhet is jelent a szervezési és technikai kérdések megoldásában; a munkamegosztás a könyvtárak között egyenetlen. Hogyan bővüljön a katalógus? Gyorsan, a könyvek megjelenése után azonnal, vázlatosabb leírással, vagy lassabban, de alaposabban, analitikus feltárással? Kinek a rekordja legyen a „minta", amely már csak kiegészül a lelőhely-adatokkal? Legyen az elsőként beérkezett, vagy valamelyik kiemelt könyvtárból érkező rekord? Olyan kérdések ezek, amelyeket nehéz, sőt szinte lehetetlen eldönteni. A jelenlegi közös katalógusok próbálnak megoldást találni, de ennek mindig van vesztese. Véleményünk szerint az ilyen a kérdéseket *nem szabad* eldönteni. Az UTCA fejlesztésénél máshol próbáljuk megragadni az alapproblémát, s így e kérdések szükségtelenné válnak. A duplumellenőrzés helyett például rekordegyesítést szeretnénk alkalmazni. Nem egyetlen rekordot kívánunk alapként kiválasztani, és a többit duplumként megjelölni, hanem az egyes rekordok adattartalmát egyesíteni egyetlen „összesített" rekordban. Így az sem lesz eldöntendő kérdés, hogy felülírhatja-e valaki a közös katalógusban már meglévő bibliográfiai rekordot, s ha igen, akkor ki. A később érkező rekordok többlet-adattartalma – például egy tanulmánykötet analitikus feltárása – beépülhet a közös katalógusban már meglévő adatok közé. Az sem lesz probléma, ha a könyvtár újra beküld egy közben kibővített rekordot – az ugyanúgy beépül, tartalma nem vész el. Ezzel a módszerrel a „hamarabb felületesen", illetve „később alaposabban" kérdést is megoldottuk, ami jelenleg még kulcskérdés. A lényeg: arra törekszünk, hogy a közös katalógusban minden dokumentum leírása tartalmazza a forráskönyvtárakban bevitt leíró adatok összességét, inkább többszörözve, de sosem elvetve a részleteket. Bármely könyvnél előfordulhat, hogy az egyik helyen jobban tárgyszavazták, a másik helyen alaposabban feltárták – például a kötetben található egyes tanulmányok és szerzőik adatait feltüntetve –, az UTCA mindkettőt tudja használni.

A csaknem kizárólagosan alkalmazott MARC csereformátum (és belső formátumként való alkalmazása) önmagában sok probléma forrása. Először is több, egymással csak részben kompatibilis változata létezik (USMARC, HUNMARC, UNIMARC és MARC-21 stb.), másrészt túlságosan kötődik a papíralapú katalogizálás jellegzetességeihez. Példa erre: a MARC változtatás nélkül átvette a korábbi szabványok *cím és szerzőségi közlés* adatcsoportját (245-ös mező), ami szétbontva is leírható (100, 240, 600, 700, 730, 740, … mezők) – ezzel ellentmondásos és hibás rekordok létrehozása vált lehetővé. Ez a megoldás nyilvánvalóan az 1960-as évek informatikai színvonalának, a szűk erőforrásoknak és az említett könyvtáros/informatikus érdekegyeztetési problémáknak a következménye lehetett. Egy új (pl. az ugyancsak akkoriban megjelenő, relációs adatmodellre épülő) szabvány alkalmazásához talán teljes rekatalogizálásra lett volna szükség. Inkább egy köztes megoldást választottak – a 245-ös mező feltehetően azért jött létre, hogy helyet adjon a cédulákról digitalizált cím és szerzőségi közlés adatcsoportnak. Véleményem szerint e nélkül nem tudták volna a gyakorlatban bevezetni a MARC szabványt. Az UTCA csak be- és kimeneti formátumként alkalmazza a MARC-ot, saját adatbázisában azok adatelemeit elkülönítve, típusuknak megfelelően tárolja és kezeli. Így válnak például valóban kereshetővé a tárgyszavak kronologikus almezői.

A MARC alapproblémái mellett már elenyésző, és szerencsére könnyen kezelhető gond a különféle karakterkészletek (ISO, Ansel, Unicode) használata. E téren az UTCA egyértelműen az Unicode mint belső formátum mellett áll ki, ezzel lehetővé téve, hogy a címeket és szerzőket eredeti nyelvükön (akár cirill, héber, arab karakterekkel) is leírjuk, és kereshetővé tegyük. Aki például orosz nyelvű könyvet keres, vélhetően tud cirill betűkkel írni és olvasni. Automatikus transzliteráció természetesen alkalmazható, ha az olvasó úgy kívánja. Az eredeti kódolással leírt címeket – mivel azok többnyire hiányoznak a hazai katalógusokból – ISBN alapján külföldi katalógusokból emelhetjük át.

A könyvtárakban sokféle integrált katalogizáló szoftvert, és esetenként eltérő katalogizálási szabályt alkalmaznak. Bár van központi szabályzat, azt teljességében betartani képtelenség. A könyvtárak hatalmas munkával építették fel elektronikus katalógusaikat, és ezeket nem egyszerű feladat egy közös szabályzatnak megfeleltetni (pl. egy kiválasztott MARC formátum egységes alkalmazását elérni). A helyzet hasonló, mint amilyen a MARC megalkotásakor lehetett, most is valamiféle kompromisszumot kellene kötni, vagy ami ma már inkább lehetséges, informatikai eszközökkel kellene segíteni a probléma megoldását.

Az UTCA nem ismeri a „hibás MARC rekord” fogalmát, és nincs saját katalogizálási szabályzata. Ezzel szemben igyekszik rugalmas lenni, és minél többféle variánst befogadni. Számunkra minden rekord, amely informatikai értelemben megfelelő (a mezők és almezők tartalmi szempontjaitól függetlenül), elfogadható. Akkor is, ha például egy mezőben szerepel a főcím és az alcím, ha nincs, vagy éppen túl sok 100-as mező van benne, ha akár az 505-ös, akár a 730/740-es mezőket használják az analitikus feltáráshoz. Ha a szabálytalanságokat kizáró okként értelmeznénk, rengeteg hasznos adatot is eldobnánk. Katalógusépítő programunk megpróbál – a MARC formátumtól elvonatkoztatva – minél több adattartalmat kinyerni a forrásrekordokból, és ezt felhasználni az azonos dokumentumokat leíró rekordok csoportokba rendezéséhez, a közös rekord elkészítéséhez. Mindez azt a célt szolgálja, hogy a közös katalógusban minél több könyvtáros szakmai munkája megjelenhessen.

Az általunk alkalmazott katalógusegyesítési folyamat nem lineáris, hanem ciklikus, és nem párosával hasonlítja össze a rekordokat, hanem valamely főbb leíró adat (cím, közreműködő) alapján képzett csoportokon belül. A ciklikusság itt azt jelenti, hogy a közös katalógust nem egy lépésben, hanem egyes lépésekhez vissza-visszatérve építjük fel, miközben a végeredmény egyre javul. Egy példa: miután rájöttünk, hogy tíz, különböző forrásból származó rekord ugyanazt a dokumentumot írja le, mert egyezik a cím, a szerző és az ISBN is, megnézzük a többi adatot, és észrevesszük, hogy a kiadó nevét többféleképpen írták le. A névváltozatokat elraktározzuk egy szótárban, mint: „Kiskapu” = „Kiskapu K.” = „Kiskapu Kiadó”, és a következő hasonlításánál felhasználjuk. Ez jól jön majd a kiadó azon könyvénél, amelynek az ISBN-jét elírták valamelyik könyvtárban. Hasonló technika alkalmazható más besorolási adatok egységesítésénél is. Annak az azonosításával például, hogy *Merle, Robert (1908-2004)* és *Robert Merle (1908-)* ugyanazon személy. Egyetlen dokumentumot leíró, sok-sok könyvtárból érkező rekordok „sokszínűsége” nemcsak hátrány, hanem előny is lehet. A csoportos összehasonlítás technikája ezt a sokféleséget használja ki, amikor távolabbról, egységként néz a katalógus egy-egy szeletére, és abban nemcsak egy-egy rekord összehasonlításával pró-

bálkozik, hanem hasonló tulajdonságokat keres e csoport egészén belül.

További újdonság, hogy az azonos dokumentumot eltérő módon leíró rekordok felismerésén túl szeretnénk az azonos *művet* leíró rekordokat is együtt kezelni, mind „felfelé", például egy mű különböző nyelvi változatait összekapcsolva, mind „lefelé", egy-egy mű (pl. novella) többféle gyűjteményes kötetben való előfordulását is összekapcsolni, az analitikus feltárásokat felhasználva. Az olvasó így a műveket, és nem a befoglaló dokumentumokat keresheti a katalógusban – az IFLA FRBR (http://www.frbr.org) ajánlásának megfelelően.

A vázolt célok nagy része algoritmikus módon valósul meg: sokféle szabályt megtanítunk a gépnek, amely ezután a rekordok millióira alkalmazza azokat. A szép elképzelést, pusztán gépi munkával képtelenség lenne megvalósítani, ezért arra törekszünk, hogy az emberi és a gépi munkát egymást kiegészítő rendszerben tudjuk összekapcsolni. Jó példa erre, ha olyan szabályt fogalmazunk meg a gépnek, amely nem eldönt egy kérdést, például azonosnak ítél meg két rekordot, hanem megállapítja, hogy a kérdés általa nem eldönthető, vagyis kézi beavatkozást igényel. Így könnyebben megtaláljuk a rekordok azon kis hányadát (a hatalmas adatbázisban), amelyekkel tényleg kézzel kell dolgoznunk. Máskor „kézzel" építünk kisebb szótárállományokat, amelyeket a gép a rekordok hatalmas tömegén alkalmaz valamely cél érdekében.

Néhány szó a technikai háttérről: bemenő adatként MARC rekordokat (bármely variánst) vagy más strukturált adatot (például Excel táblázatot, ha nincs katalógus) tudunk befogadni. Ezeket karakterkonverziótól eltekintve, eredeti formájukban őrizzük meg, a további feldolgozást a belőlük leszűrt adatokon, egy relációs adatmodellben, *MySQL* adatbázisban tárolva végezzük el. A rendszer kifejezetten a közös katalógus építésére készült, más könyvtári funkciókat nem lát el. A fő alkalmazást *Delphi* nyelven írjuk, *Windows* környezetben, az OPAC-ot *Adobe Flexben*, *Flash* és *PHP* technológiával. Ezen kívül – csak a fontosabbakat említve – *Linux* operációs rendszert, *dotProjekt* munkamenet-szervezőt, *Hunspell* és *Hunmorph* nyelvészeti alkalmazásokat, *VMWare* virtualizációs szoftvert használunk (ennek segítségével futtatjuk a windowsos feldolgozó szoftvert a linuxos adatbázis- és webszerveren), valamint a *Google* webstatisztikáját alkalmazzuk (ebből tudjuk, hogy a látogatók többsége már rendelkezik az OPAC használatához szükséges 9-es változatú Flash lejátszóval).

Mindezek ingyenes, többnyire nyílt forráskódú szoftverek és szolgáltatások, *anyagi ráfordítást nem igényeltek,* kivéve a szervert, amely jelenleg egy használt *Compaq* gép. A szerverhez járt egy Windows 2000 licenc, amit a feldolgozó alkalmazást futtató virtuális géphez használunk.

Jelenleg a fejlesztés stádiumában vagyunk. Első lépésként az azonos dokumentumokat leíró rekordokat azonosító, és a besorolási adatokat (személynév, kiadó neve, földrajzi név, tárgyszó) egységesítő módszereken dolgozunk. Az eredmények bíztatóak, bár látszik, hogy még sok munka van hátra. Néhány példa: a MOKKÁ-ban valamivel több, mint 200 olyan rekord található, amelynek címe részben vagy egészben „általános pszichológia". Ez valójában kilenc művet takar, persze többféle kiadásban. Mi ezt egy hibaponttal, tíz műként tudjuk automatikusan azonosítani és megjeleníteni a felhasználónak. Egy másik példa az „Égi és földi szerelem" – az öt ilyen című művet mind különkülön felismerjük (tehát csak öt találatot adunk e kérdésre), holott sokkal több ilyen rekordot dolgoztunk fel. Természetesen hibákra is tudnánk példákat hozni, ezek közül még sokat ki tudunk küszöbölni, de teljesen hibátlan adatbázist szinte lehetetlen kialakítani. Fontos azonban, hogy a hibák hányada elenyésző lesz, és az azonosítást nem fogják akadályozni – az eredeti rekordok ugyanis mindig megjeleníthetők lesznek. A katalógus közösségi jellegének köszönhetően a hibákat bármely gyakorlott könyvtáros kolléga azonnal kijavíthatja. Mi pedig, értesülve erről, mindig tanulni fogunk a hibákból…

Az UTCA projekttel sokféle szándékunk van, de most a hangsúlyt a fennálló problémákra adott alternatív megoldásokra helyeztük. Az ez évi *Networkshop* konferencián már tartottunk egy bemutató előadást a projektről (http://vod.niif.hu/). Szeptemberben az *OSZK*-ban, a *Könyvtári Intézet* és az *MKE* műszaki szekciója szervezésében szakmai napon veszünk részt egy bemutatóval, új kezdeményezésekkel és vitalehetőséggel, részben az UTCA projekt, részben a Web 2.0 kapcsán. Erről és más hírekről is részletesebben olvashatnak a http://konyvtar.info/ oldalon.

Kardos András
(Informatikus, az ELTE könyvtárszakos hallgatója)