Lengyel Mónika – Andaházi Szeghy Viktor – Magyar Gábor

Mindenki másképp csinálja! A retrospektív konverzió két útja

Összehasonlító esettanulmány

Vajon mit kívánna napjaink finanszírozási gondokkal küzdő könyvtárvezetője egy bánatos hajnalon a horgára akadt aranyhalacskától? Szerepelne-e a három kívánság között a könyvtárban porosodó katalóguscédulák retrospektív konverziója? A cikk szerzői feltételezik, hogy erre nagy valószínűséggel igen a válasz. Sajnos aranyhallal nem tudunk szolgálni cikkünkben, de megpróbáljuk két könyvtár retrokonverziós projektjének folyamatát és tanulságait összevetni. Kérdés, lehet-e még hinni a hétköznapi csodákban?

A tengerentúlról kiinduló, mára Nyugat-Európa könyvtárain is átsöprő retrokonverziós hullám immár Magyarországon is elérte a közép- és kisebb méretű könyvtárakat. Bizonyított tény, hogy annak a könyvtárnak a látogatottsága, és ezáltal a jelentősége is erősen megnő, amelynek teljes állománya hozzáférhető elektronikus katalógusában [1]. Az online elérhetőség nemcsak az eddig elfekvő állományrészek kihasználtságát és a könyvtár jelentőségét növeli, de hozzájárulhat a belső munkafolyamatok javulásához is, amely további plusz szolgáltatások megjelenését generálhatja. Ennek megfelelően a közép- és kisebb méretű könyvtárak számára a retrokonverzió mára semmiképpen sem tekinthető „úri huncutságnak", sokkal inkább létkérdésnek.

Cikkünkben két eltérő jellegű gyűjtemény cédulakatalógusának retrospektív konverzióját mutatjuk be. Az első esettanulmány a *Hadtörténeti Könyvtár* (*HK*), mintegy 20 ezer katalóguscédulájának retrokonverzióját írja le, a második pedig a *Magyar Táncművészeti Főiskola Könyvtárá*nak (*MTF*) kevesebb, mintegy 16 ezer cédulájának feldolgozását.

Általános megfontolások

A hazai és a nemzetközi szakirodalom alapján a retrokonverzió elvégzésére az alábbi módszerek állnak rendelkezésre [2]:

- cédulakép digitalizálása, azt követően pedig adatbevitel a szkennelt képről;
- OCR karakterfelismertetéssel létrehozott szöveges információ segédprogramokkal történő átalakítása az adatfelismerés és tipizálás megkönnyítésére;
- rekordletöltés az OSZK, MOKKA, valamint külföldi adatbázisokból és a rekordok honosítása;
- az előzők kombinációja.

A retrokonverzió módszerének kiválasztásakor az alábbi szempontok figyelembevételével célszerű döntést hozni:

- *Rendelkezésre álló / megpályázható keret;* ez a pont önmagáért beszél, nyilván a leginkább költséghatékony megoldást kell előnyben részesíteni.
- *Feldolgozandó cédulák száma / kora;* az előzőek függvényében a mennyiségi és minőségi kérdéseket kell megvizsgálni.
- *A feldolgozandó cédulakatalógus egyedisége;* ez már sokkal erősebben gyűjteményspecifikus kérdés, mely nagyban meghatározza, mit részesítünk előnyben az általánosságban vázolt módszerek közül.
- *A meglévő kurrens adatbázis tulajdonságai;* ez szintén nem elhanyagolható kérdés, figyelembe véve, hogy a retrokonverzió terve alapvetően olyan könyvtárakban merül fel, amelyek már több éve dolgoznak valamilyen IKR-ben.

Bármilyen módszert választunk is, a cédula képe mindenképpen gépre kerül, ami már önmagában is könnyebb feldolgozási lehetőséget kínál. Így kezdődött ez a fentiekben jelzett két könyvtár retrokonverziója során is, azonban ettől kezdve más utakon folyatódott.

A HK 20 ezer cédulájának feldolgozása

A vállalkozás méreteinek érzékeltetéséhez érdemes tudni, hogy a könyvtár 2005-től kezdve a *HunTéka* integrált könyvtári rendszert használja. Ezt megelőzően, 1997-től a TINLIB rendszert alkalmazta, amelyben 15 ezer rekordot dolgozott fel. A retrokonverzió elkezdésekor ennek több mint kétszerese, kb. 33 ezer volt megtalálható a huntékában, a konverzióval bekerülő állományrész, a mintegy 20 ezer cédula pedig ennek csaknem 2/3-át képezi, miközben további 100 ezer cédula vár még feldolgozásra.

A folyamat előkészítése

Meg kell említeni, hogy az előkészítés során elméleti alternatívaként felvetődött az autopszián alapuló rekatalogizálás lehetősége is. A cédulaalapú retrokonverzióval összevetve azonban az idő – költség – minőség szempontrendszerben [3] vizsgálva ez utóbbi módszer nem bizonyult versenyképesnek.

Idő

Az autopszia a dokumentumokhoz való folyamatos hozzáférésen alapul, tehát csak munkaidőben végezhető. Ezzel szemben a retrokonverzió manuális korrektúrája során alkalmazott szerver–kliens rendszer 0-24 órás rendelkezésre állást tesz lehetővé. Egy Java alapú platformfüggetlen megoldás minden szoftverkörnyezetben biztosítja a hozzáférést, függetlenül a távmunkában dolgozók operációs rendszerétől. A munkavégzési mobilitás fenti lehetőségei csak képdigitalizálással rögzített adatok távoli hozzáférésével válhatnak lehetővé.

Költség

Az élőerős munka jelenti a feldolgozási folyamat fajlagosan legköltségesebb részét. Növeli a költséghatékonyságot, ha bizonyos munkafolyamatok kiválthatók automatizálható gépi megoldásokkal. A rekatalogizálás teljesen manuálisan történik, így ott nem beszélhetünk az emberi munkaidő olcsóbb gépi kiváltásáról.

Minőség

A rekatalogizálás során kizárólag manuálisan bevitt adatokból épül fel az adatbázis. Könyvtáros szakemberek alkalmazásával (bár ez a költséghatékonyság rovására mehet) az adattévesztés minimalizálható, ugyanakkor az elütési hibák elkerülhetetlenek. Az elgépelés szűrésére több módszer létezik. Nyilván hatékony, de a legdrágább megoldás, amikor ketten gépelik ugyanazt az adatsort, és a különbségeket ellenőrzi a keretprogram.

Kompromisszum

A retrokonverzió során adatbázisba kerülő rekordok több automatizált gépi fázis után manuális korrektúrán esnek át. Ezzel elérhető a másik módszert megközelítő pontosság, alacsonyabb áron.

Várakozások, előnyök

A HK retrokonverziós projektjének előkészítésekor a következő gyűjteményspecifikus jellemzőket kellett mérlegelni:

- Az állomány jelentős része, mintegy 35-37%-a, idegen nyelvű dokumentum: nagy a dokumentumok nyelvi szórása. Sorrendben a leggyakoribb nyelvek: német, orosz, angol, francia, román, cseh/szlovák. (Előfordulnak még: szerb/horvát, spanyol, portugál, török, koreai, kínai.)
- A gyűjteményben komoly darabszámot képviselnek az 1945 előtti dokumentumok – egészen 1534-ig visszamenőleg.
- Gyűjtőköri sajátosságokból adódóan nagyobb mennyiségű, máshol nem szerzeményezett dokumentum is megtalálható a gyűjteményben. Ilyen például a katonai szabályzatok mintegy 45 ezer tételből álló kollekciója.

A gyűjteményi sajátosságok alapján tehát itt nem volt járható a teljes mértékben rekordimportra alapozott katalógusépítési megoldás, hiszen az anyag egy jelentős részét (pl. katonai szabályzatokat) még sehol nem írták le elektronikus katalógusban. Egy jelentős rész ugyan letölthető volna, viszont várhatóan csak több nemzeti könyvtár átfésülésével, ami az egy rekordra eső munkaidőt (és költséget) tetemesen megnövelte volna (feltételezve a szükséges Z39.50 kapcsolat meglétét).

A fenti megfontolásokat figyelembe véve alapvetően a katalóguscédulák képfeldolgozására és karakterfelismertetésére lett felépítve a rekatalogizálás munkafolyamata. Ennek alapfeltétele a lehetőség szerinti legteljesebb, szabványos cédulakatalógus volt. Forrásként a könyvtár szolgálati katalógusa lett kiválasztva, melynek céduláin az olvasótermi katalógustól eltérően a leltározási és a példányszámadatok is jelen vannak. Jellemző, és valószínűleg nem egyedi jelenség, hogy az itt 1958 óta épített cédulakatalógus egyes cédulái között – már csak a leírási szabvány időközbeni változása

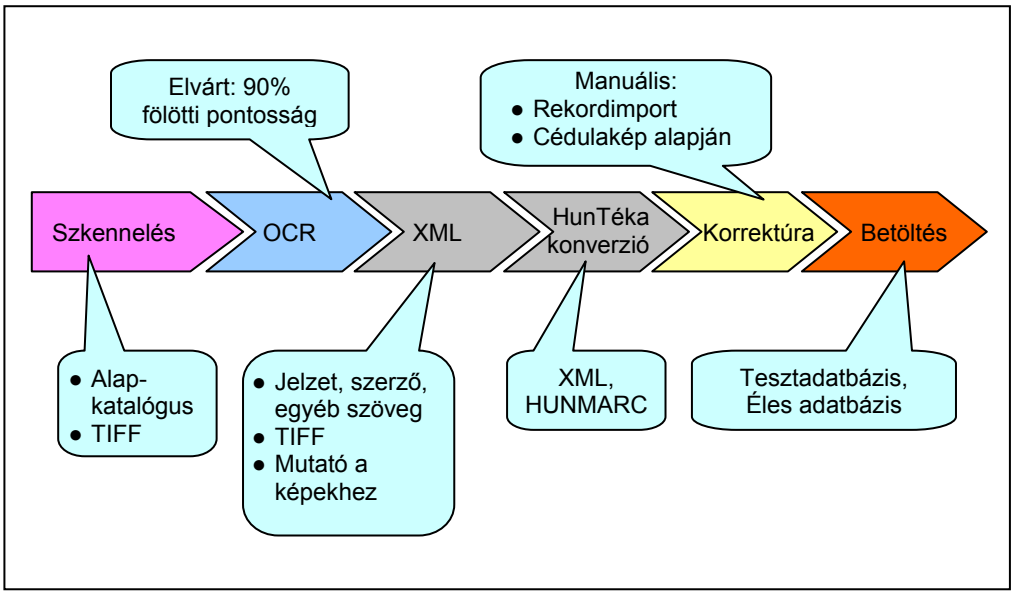
miatt is – jelentős eltérések figyelhetők meg. Leginkább a szeparátorok változása szembetűnő.

Az állományi és nyilvántartási adottságokon, lehetőségeken túl, a folyamat előkészítésekor számba kellett venni a munkamenet során kihasználható előnyöket. A képfeldolgozás szempontjából előnyként volt értékelhető, hogy az adathordozók szabvány cédulák, továbbá, hogy a részben szabványos leírások várhatóan segítik majd az egyes adatcsoportok elkülöníthetőségét, ezzel gyorsítva a feldolgozást. Az előzetes várakozások szerint a cédulaképet jól fel lehet majd használni az adatok korrekciójánál. Adottságként lettek kezelve az egynemű gyűjteményi részek (szabályzatok, volt zárt anyag stb.), melyekre nézve elkülönítve terveztük a feldolgozást megvalósítani.

A várható kockázatok vizsgálatakor merültek fel a következők: a cédulák többféle leírási szabvány alapján készültek, így a rajtuk szereplő adatok szétválogatásának eredményességét is eltérően lehetett prognosztizálni; az OCR és a konverziók hatásfoka, hiszen az adatok helyessége a végleges rekordokban leginkább ezeken a részfolyamatokon múlik; továbbá az sem volt lényegtelen, hogy a korrektúra során mennyit kell javítani az egyes rekordokon.

A HK retrokonverziós folyamat lépései (1. ábra)

- A képfeldolgozáson alapuló eljárás során a legtöbb adatot tartalmazó alapkatalógus céduláiról készült, jó minőségű képek szövegfelismertetése történt meg.
- A karakterfelismerés hatékonysága meghatározza az összes további munkafolyamat sikerességét, így elengedhetetlen az elérhető legnagyobb pontosság.
- Az elsődleges XML állomány a teljes felismertetett szöveget, a cédula képét, és a kettőt összekapcsoló azonosítót tartalmazta.
- Az ebből készült elsődleges MARC rekord a dokumentum raktári azonosítóját, a megjegyzés mezőbe ömlesztett teljes szöveget és a cédulaképre utaló indexet tartalmazta.
- Ezt követte – már a HunTéka tesztadatbázisban – az általános megjegyzés mezőbe ömlesztett adatok automatizált szétválogatása a HUNMARC-nak megfelelő mezőkre.
- Az így létrejött rekordokat manuálisan korrektúrázták a cédulaképre, mint elsődleges adatforrásra támaszkodva.
- Végezetül megtörtént a korrektúrázás után a 20 000 db. retrokonvertált rekord éles adatbázisra töltése.

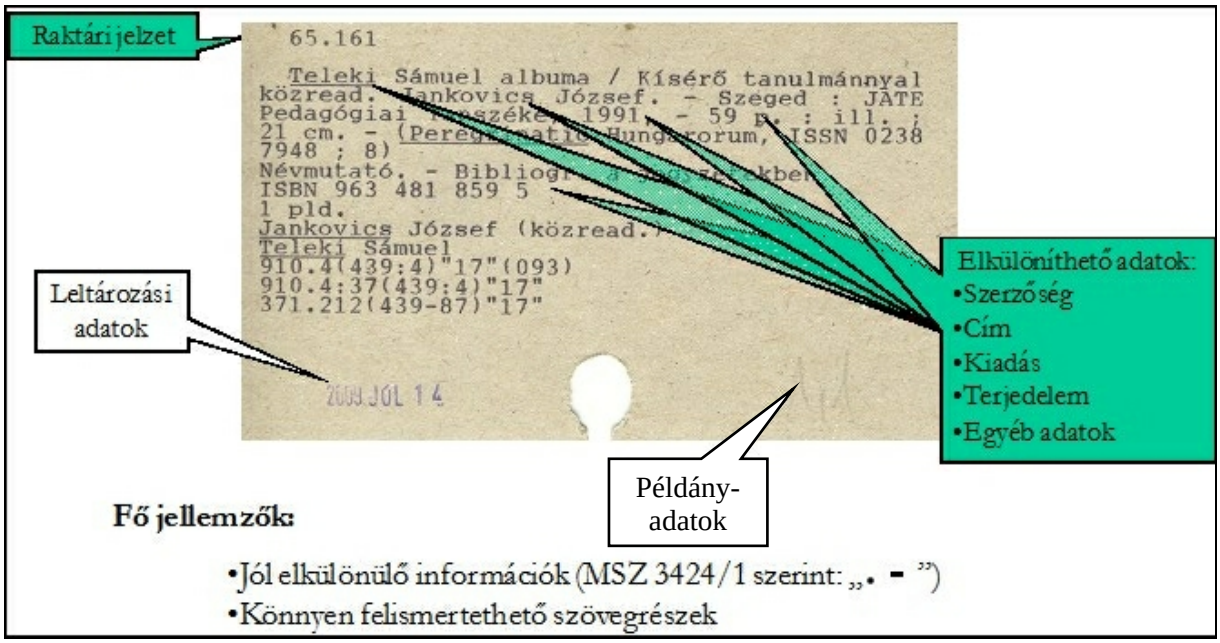


1. ábra **A könyvtárban megvalósított folyamat elve**

Képfeldolgozás és karakterfelismertetés

A cédulaképek digitális változatai a folyamat első fázisában karakterfelismertetésen estek át. A képfeldolgozás és az OCR hatásfokát, eredményességét legjobban a folyamat eredményeként kapott MARC hívójeleket tartalmazó XML formátum szemlélteti. Egy ISBD szabványt követő cédula képe a 2. ábrán, az ebből képzett redukált MARCXML a 3. ábrán látható. Ez az elsődleges (redukált adattartalmú) MARCXML mindösszesen a 856 (Elektronikus hely és hozzáférés), a 852 (Elhelyezés) és az 500 (Általános megjegyzés) mezőkben tartalmazott adatot.

A kiemelt példákon (zölddel) a bedolgozandó adatok vannak jelölve. A fehér mutatók a még szóba jöhető, de ez alkalommal figyelembe nem vett kiegészítő információkat jelölik. Az xml mintában a piros pedig már a később korrigálásra szoruló, nem megfelelően értelmezett adatokat jelöli.



2. ábra **ISBD szabványt követő cédula képe**

```
<record>¶
··<datafield·code="856">·¶
····<subfield·code="f">6e99c3e8c0cf4d38ac4306b1e65816d1.tif</subfield>·¶
··</datafield>·¶
··<datafield·code="852">·¶
····<subfield·code="m">65161</subfield>·¶
··</datafield>·¶
··<datafield·code="500">·¶
····<subfield·code="a">Teleki·Sámuel·albuma·/·Kísérő·tanulmánnyal·közread.·Jankovics·
József.·—·Szeged·:·JATE·Pedagógiai·tanszéké,·1991.·-·59·p.·:·ill.·7948·;·8·)·Perearinatio·
Hungarorum,·ISSN·0238·Névmutató.·-·Bibliogr.·a·jegyzetekben·ISBN·963·481·859·5·1·pld.·
Jankovics·József·(közread.)·Teleki·Sámuel</subfield>·¶
··</datafield>¶
</record>¶
```

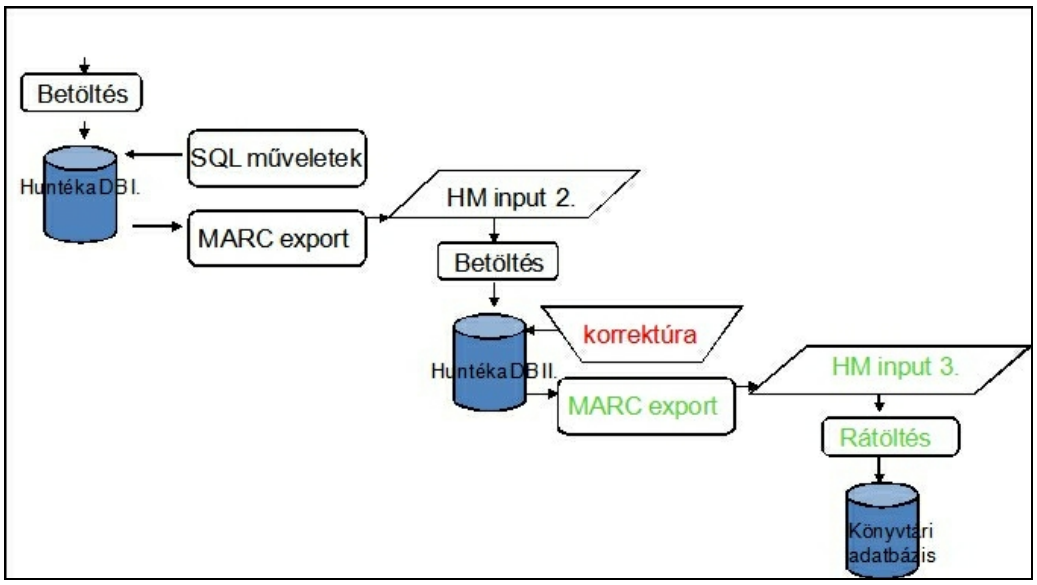
3. ábra **Redukált MARCXML**

Az ISBD alapú példán és annak XML formátumban megkapott rekordján ugyanakkor sajnos jól látszanak a karakterfelismertetés hibái: a sortörések helytelen felismerése és kezelése folytán a terjedelemi adatcsoportba „tolt” ISSN és sorozatszám, az ISBN végéhez ragasztott példányszám, a megjelenési évet követő adatcsoportot jelölő „pont gondolatjel” vesszőbe torzult pontja, a lemaradt ETO jelzetek és melléktételek. Az alapvető problémát a központozási jelek félreismerése, illetve fel nem ismerése okozta. Ennek megfelelően nem volt sokkal rosszabb a helyzet az ISBD-t nélkülöző, azt megelőző korszakban készült leírásoknál sem (l. a 4. és az 5. ábrát). A gyakorlatban ezt jelenti az a nagyon pontosnak tűnő statisztikai adat, amit az OCR-ről itt-ott olvasni lehet, miszerint annak felismerési pontossága 96%-os. Ez 100 karakteres sorokkal számolva a valós alkalmazásban tehát soronként 4 karaktereltérést jelent! És mindebbe nem számíthatjuk bele a cédulatartalom szeparátorkaraktereit és a kötött helyzetű szóközöket, valamint a kézzel írt szövegrészek felismerését, mert csupán karaktereket ismer fel, miközben azt jól tudjuk, hogy a cédulákon minden területnek, írásjelnek, szóköznek és a szövegcsoport elhelyezkedésének megvan a maga speciális jelentése és szerepe.

4. ábra **ISBD előtti cédula képe**

```
<record>↵
<datafield code="856">↵
<subfield code="f">afa058c43df64225a4238ecc1f7af57d.tif</subfield>↵
</datafield>↵
<datafield code="852">↵
<subfield code="m">sz 3493</subfield>↵
</datafield>↵
<datafield code="500">↵
<subfield code="a">Fegyverzeti teljességi utasítás a 10.5 cm 40/43 M. roham tarack és a 7.5 cm 43 M. roham-ágyú részére. Bp. ny.n. 1944. 13 p. 28 cm. F-23</subfield>↵
</datafield>↵
</record>¶
```

5. ábra **ISBD előtti céduláról készült MARCXML**



6. ábra **Az 1. ábra „huntéka konverzió" elemének kifejtése**

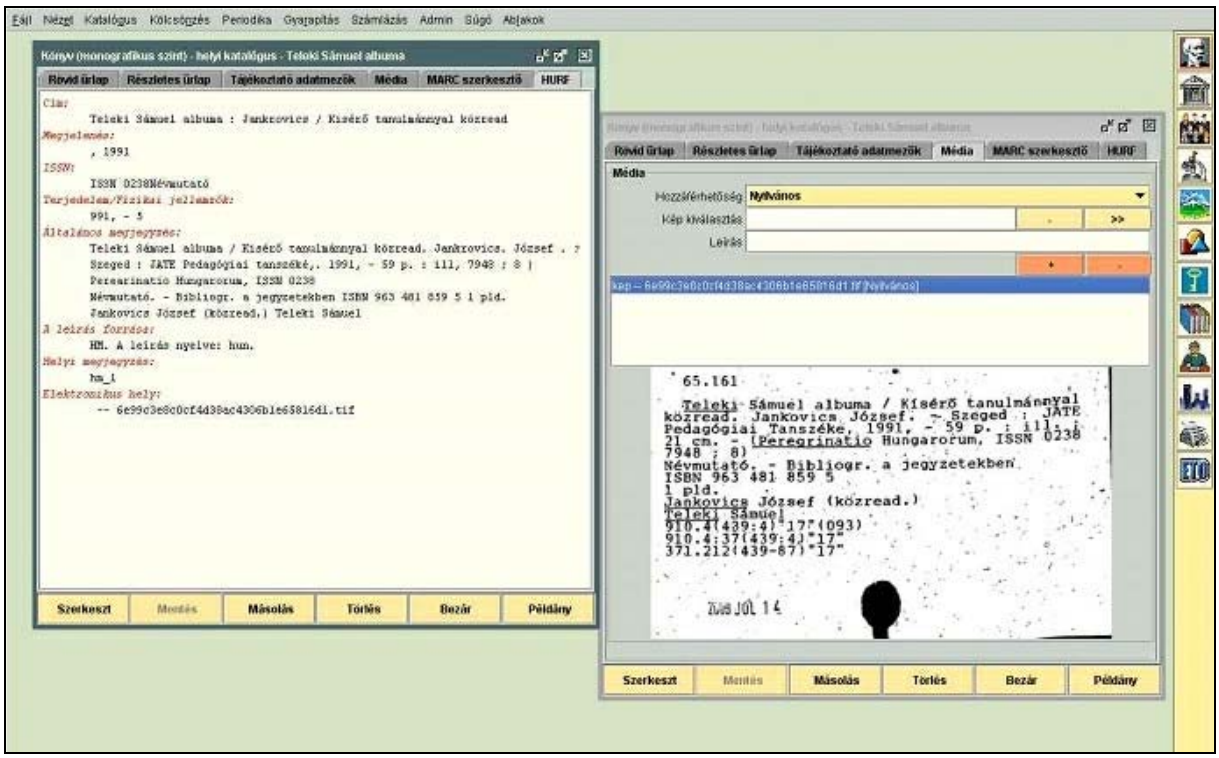
Az 1. ábra „huntéka konverzió" elemének kifejtését láthatjuk a 6. ábrán. Ez a lépés 3 ütemben zajlott le.

Az első adatbázisba egy tagban az egész cédula tartalma, tehát a fenti OCR utáni adatfolyam került. Ezen próbáltuk elvégezni az adatcsoportok elkülönítését, ami a vártnál gyengébb OCR eredménynek is betudhatóan, sajnos némileg alulmúlta az elvárásokat.

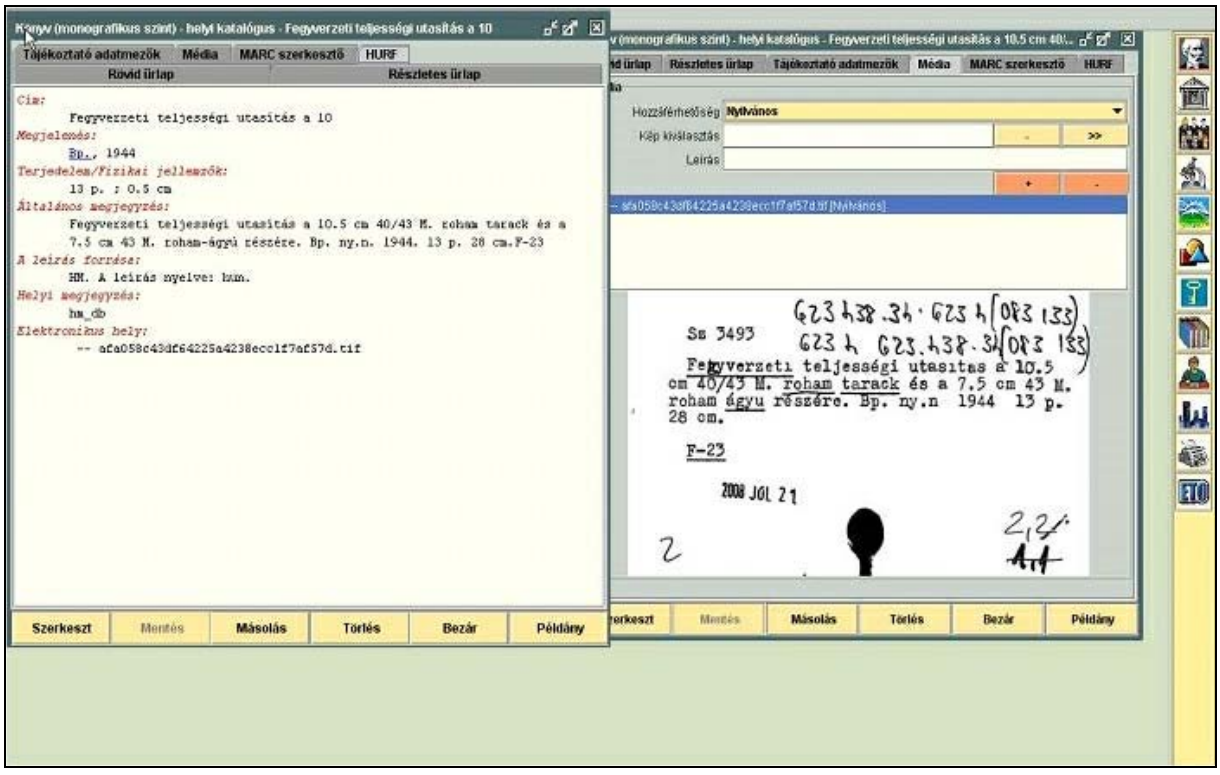
Az adatfolyam feldarabolása az alábbi algoritmusokkal zajlott. Első lépében az 500$a adattartalom került feldarabolásra a központozási jelek mentén. Az így kapott szövegdarabok egy része „hátulról" került visszafejtésre, jellemzően ebbe a csoportba tartozott a méret a 'cm'-t megelőző számjegyeivel, a terjedelem a 'p.' 'o.' 'l.'-t megelőző számjegyeivel, vagy az illusztráltság jellegzetes kifejezései ('ill.', *részben színes*, *színes*). Az egyes visszafejtések lépéseinek sorrendje itt nagyon fontos volt. Ezt követte a cím kivágása az első pontig vagy '/'-ig, az ezt követő szövegrész szerzőségi közléssé alakítása, valamint a cím utáni oszlop alcímként való kezelése. A szerzők alaki ellenőrzésen mentek keresztül, a kiemelt vezetéknév itt sokat segített. A maradék részek kezelése már tartalomelemzéssel párosult a bizonytalan sorrend és a hiányok miatt. Ilyen volt például a megjelenés helye, leggyakrabban Budapest, Berlin, Wien, vagy a leggyakoribb kiadók (melyet a sorrendiség is megerősített). A megjelenés éve ugyancsak jól kivehetőnek tűnt, bár a címben szereplő évszámok bezavarhattak. A kiadásjelzés adata is jellegzetes karaktersorozat ('xx Aufl.', 'x Kiad. Stb.) alapján volt felismerhető. Utólagos ellenőrzések alapján szűrésre kerültek a szerzőkbe került címek, illetve a címek közé került szerzők (névelő megléte alapján, ill. nagybetűs írásmód alapján). A zárójelezett adatsorról feltételeztük, hogy megfeleltethető a sorozatcímnek.

Az így létrehozott adatbázis természetesen nem volt korrekt, mert a darabolások alapvetően „két dimenzióban" zajlottak, az adatbázison belül nem képeződtek le valós relációk, azaz nem jöttek létre valós besorolási állományok. Ezt az adatbázist tehát nem adhattuk volna oda a korrektoroknak, mert nem tudtak volna vele dolgozni. Ezért szükség volt egy köztes exportra és importra, melynek során már MARCXML-t előállítva és azt egy második huntékába betöltve immár egy korrekt, szerkeszthető huntéka adatbázist kaptunk eredményül.

A felkért korrektorok tehát a második huntéka adatbázissal dolgoztak, a fenti darabolások eredményeképpen létrejövő rekordok itt kerültek manuális javításra (l. a 7. és a 8. ábrát). Minden rekord tartalmazta a cédula eredeti digitalizált képét is, a javítás során elsősorban ehhez és nem karakterfelismertetett szöveghez képest történt a behasonlítás.



7. ábra **Szabványos leírás HunTéka munkafelülete a korrektúrát megelőzően**



8. ábra **Nem szabványos leírás HunTéka munkafelülete a korrektúrát megelőzően**

A korrektúra alapvető tapasztalatai, tanulságai

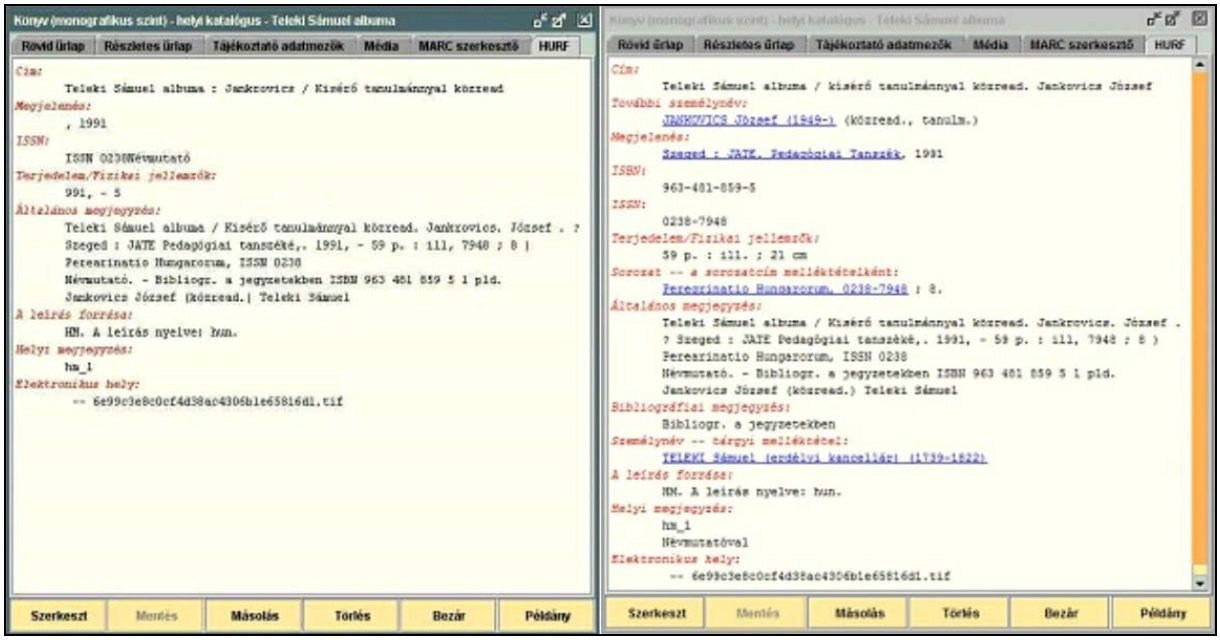
A másodlagos MARC rekordok korrektúrája során több, a hasonló eljárást választó gyűjtemények számára is hasznosítható tapasztalatot sikerült felhalmozni. Ezek közül kiemelnénk a leglényegesebbeket. Az egész eljárásról elmondható, hogy az előzetes várakozásoktól eltérően az ISBD szabvány szerinti katalóguscédula és a korábbi keletkezésű, nem szabvány cédulákból készült rekordok között érdembeli minőségi eltérés nem volt megfigyelhető. Az egyes adattípusokról összességében elmondható, hogy gyakorlatilag 90-100% közötti volt az egyezés az automatizált munkafázisokat (OCR, konverzió) követően, ami önmagában nem rossz szám. A „cím” adatok karakterei (kivéve: aláhúzott szöveg, â, è, ş), a cm-ben megadott terjedelem, a mellékletek megléte, leltári szám 100%-ban jöttek át a korrektúra fázisába. Ugyanakkor a megjelenési helyek, kiadók, oldalszámok behatárolása már csak 90% körüli értékre sikeredett (főleg a nagyobb, elterjedtebb helység- és kiadói nevek jöttek át jobb hatásfokkal). Egyes adatokat azonban, mint például: alcímek, sorozati adatok, manuálisan kellett kiválasztani.

A 20 000 rekordból mindössze fél % (tehát alig 100 db!) volt, amin nem kellett manuálisan módosítani. A manuális ellenőrzés és korrekció megkerülhetetlenségét támasztja alá az is, hogy a szinte 100%-ban korrekt adatcsoportoknál is előfordult, hogy esetileg máshová kerültek a folyamat során az adatok, például: leltári szám. A rekordok korrektúrát követő véglegesítése előtt ezért volt kötelező az összes adatot megjelenítő HURF nézet áttekintése minden alkalommal (*9. ábra*).

A korrektúrázott állomány a harmadik fázisban elvégzett migrációval került a könyvtár éles adatbázisába. Ez a mozgatás is alapvetően a bibliográfiai és a példányrekordokra vonatkozott, azonban az ellenőrzéskor kiderült, hogy a korrektúra strukturális mélységben érintett bizonyos besorolási állományokat is, legnagyobb mértékben természetesen az egységesített besorolási neveket. Az így létrejött névváltozatokat, némi áldozatok árán, nem szabványos megoldások alkalmazásával átmentettük az éles adatbázisba.

A Hadtörténeti Könyvtár retrokonverziója során felhalmozott tapasztalatok közül néhány megállapítás kiemelhető azok számára, akik hasonlót terveznek:

- Kimondható, hogy a *karakterfelismertetés és a konverzió összesített hatásfoka leszámolt azzal az illúzióval, hogy az OCR egyedüli és általános megoldást jelenthet*.
- Az adatok teljes újragépelésének elkerülésével a további szövegromlás lehetősége megszűnik.



9. ábra **Egy rekord korrektúra előtt és után**

Néhány, a retrokonverzió során kiaknázható további lehetőségek közül:

- A feldolgozás során egy jó alapkatalógus kiegészítő adatai (leltár, ETO szám, példányadatok stb.) is bekerülhetnek a végleges rekordokba.
- Megfelelő gyűjtemény esetén a rekord import kiterjesztése (pl. 1976-ra szűrés a magyar kiadványoknál) az adatok bevitelét nagyban segítheti.
- Az adatpontosság növelése természetesen költségcsökkentést jelenthet, hiszen az ezt követő manuális munka volumene csökkenhet.

A HK projekt költségei

Az 1. táblázatban a projekt egyes fázisainak idő- és költségvonzatai láthatók, nagyságrendi számadatokkal. Mint megfigyelhető, a manuális adatbevitel tette ki mind időben, mind költségben a legnagyobb tételt.

1. táblázat
HK retrospektív projekt költségei

Munkafázis	Időtartam*	Bekerülési költség/db**
Képfeldolgozás	2 nap	25 Ft
OCR	4 hét	
Konverzió	2 hét	60 Ft
Korrektúra	5 hónap***	80 Ft
Összesen	6,5 hónap	165 Ft

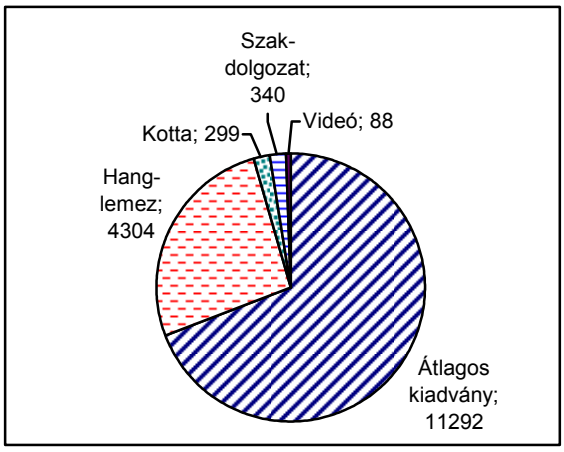
* Nettó idők: betöltéseken, tesztüzemen, próbaszériákon túl
** Nagyságrendi, kerekített összegek
*** Heti átlagok: 124–5778 (!) db között

Az MTF 16 ezer cédulája

Az MTF könyvtára üdítő kivételt jelent a retrokonverziós projektek sorában, ugyanis a munka elkezdését megelőzően nem rendelkezett semmiféle gépi nyilvántartással, azaz az állomány nagyobbik hányada csak cédulán, egy kisebbik hányada pedig még cédulán sem volt feldolgozva. A feldolgozatlan kottaállomány leírása egy erre a célra szerveződött ideiglenes munkacsoport segítségével, katalogizálási szabályzatot követve, az autopszia elvén zajlott, egy üres huntéka rendszerben. Ezzel párhuzamosan kezdődött meg a 16 ezer cédula konverziója, amelyből a duplumcédulák miatt végül 14 ezer rekord született.

Általános megfontolások

A könyvtár állományát (*10. ábra*) figyelembe véve feltételezhető volt, hogy annak döntő többsége elérhető lesz már más katalógusokban. Fő átvételi forrásként a könyvtár kérésére az immár saját retrospektív katalogizálási projektjén túljutott, 400 ezer cédulával bővített OSZK katalógusa [4] lett megjelölve. A gyakorlatban azonban az OSZK főforrásként elsősorban a szépirodalmi vonatkozású és általános műveknél vált be, jóllehet közelítőleg ezek 30%-a sem volt megtalálható benne. A könyvtár főprofiljának tekinthető táncművészeti szakirodalmat, valamint a zenéhez, tánchoz kapcsolódó speciális kiadványokat, a hanglemezeket, kottákat, tánctörténeti irodalmat, videofelvételeket pedig egyáltalán nem lehetett megtalálni az OSZK adatbázisában. A fenti szakmaspecifikus kiadványok rekordjainak begyűjtéséhez további forrásokat kellett tehát keresni, úgymint a *FSZEK* katalógusát, valamint a *Karlsruhe-i Virtuális Katalógus* (*Karlsruhe Institute of Technologie, Karlsruhe Virtual Catalog* = *KIT-KVK*) mintegy 500 millió rekordra tehető adatállományát.



10. ábra **A MTF állományának tartalmi megoszlása számokban**

Kezdettől két alapvető problémával kellett szembenézni:

- Duplumcédulák megléte akár ugyanazon a katalóguson belül is, valamint leltári számok ismétlődése más szempontú rendezésben.
- A másik alapvető helyi sajátosság a hangzóanyagok analitikus leírásainak megléte volt. Ezek átvétele más katalógusokból korlátozott, itt tehát megint más módszert kell követni.

Megvalósítását tekintve ez tehát vegyes helyzet volt.

A digitalizált cédulák megmunkálása

A szkennelés, mint minden hasonló esetben, a munka egyik legegyszerűbb része. Szinte tovább tart a fel- és levonulás a szkennerrel, mint maga a beolvasás. Az olvasási műveletet egy Fujitsu gyártmányú automata lapadagolós lapolvasó berendezéssel végeztük (*11. ábra*). A teljes mennyiség beolvasása, a cédulák fiókból történő kiszedésével és visszarendezésével együtt összesen 14 órányi munkaidőt vett igénybe.

11. ábra **Automata lapadagolós lapolvasó berendezés képe**

Ezt követte az OCR művelete, melynek során megtörtént a szkennelt képállomány szöveges adatsorrá alakítása. Jóllehet az OCR eredményességét már fentebb is firtattuk, a feldolgozás folyamatába való beiktatása ennek ellenére megkerülhetetlen, mivel segítségével a munkatársak a tartalmi feldolgozásra koncentrálhatnak. A felismertetés során a karakterfelismerő program kibővítésével egy előzetes adatszeparációt is elvégeztettünk, ami egy elválasztó karaktersort illesztett be a leltári szám és a cédula tényleges szövege közé, valamint besorszámozta a szkennelt képek sorozatszámai alapján a szövegblokkokat.

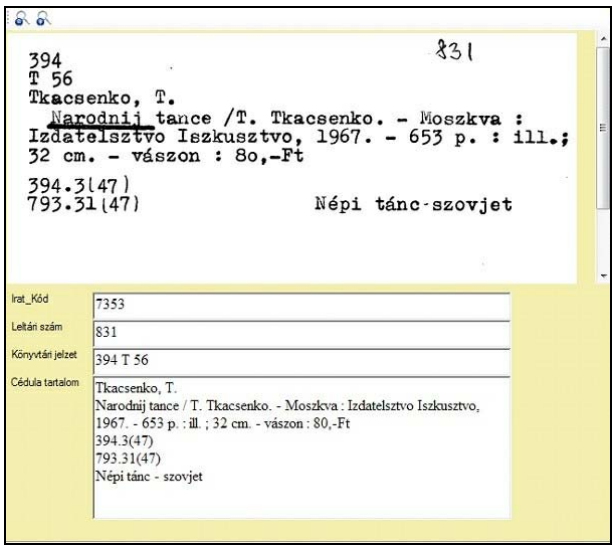
Az OCR-rel előállított adathalmazt és képállományokat ezt követően feltöltöttük az *MKBLUX* által fejlesztett *PraktiDok* feldolgozó rendszer adatbázisába. A retrospektív konverzió minden további lépése ebben a könyvtári munkafolyamatokra is felokosított dokumentumkezelő rendszerben zajlott. Ez a program eredetileg nagytömegű tetszőleges iratanyag iparszerű feldolgozására lett kifejlesztve. Az eredeti kívánalmaknak megfelelően az érdemi munkavégzést maximálisan támogató funkciókon felül, a feldolgozási folyamat nyomon követésére, valamint az egyéni teljesítmények értékelésére és elszámolására szolgáló funkciókat is tartalmazza, amit a feldolgozás során jól ki tudunk használni. Az adatbázisokat és a képállományokat tartalmazó kiszolgáló egy nagysebességű internetkapcsolatot biztosító szerverhotelben van elhelyezve, így a távmunkában, otthonról dolgozó feldolgozók egyszerű, minimális irodai célú felhasználásra szánt számítógépekkel, normál „háztartási” internetkapcsolattal csatlakoztak a szerverhez. A munkatársak egyedi felhasználónévvel és azonosítóval, a képességeiknek és feladatuknak megfelelő jogosultságokkal rendelkeznek. A „vastag-kliensalkalmazás” minden egyes bejelentkezéskor program vagy konfigurációs adat változásakor frissül, így biztosítva azt, hogy mindig mindenki a legfrissebb programverzióval rendelkezzen. Az egyes cédulákkal végzett műveletek során az átmozgatott adatmennyiség − a cédula képével együtt − nem haladja meg a 40 kilobájtnyi méretet, tehát még nagyszámú felhasználó esetében is igen gyors működést biztosít.

Az elvégzendő első lépés tehát a karakterfelismertetett anyag ellenőrzése, azaz a szövegjavítás művelete volt. Az adatrögzítő munkatársaknak ebben a fázisban kellett összevetni az egyes cédulaképeket a hozzájuk tartozó adatcsomagokkal, elsősorban szövegpontosság és a szövegblokkok elhelyezkedésének szempontjából. A könyvtár által is jóváhagyott cédulákon a feldolgozási szabályzat alapján sortöréskarakterek beszúrásával szeparálni kellett a szerzőségi közlés tartalmát, a cédulatartalom szövegblokkját, az ETO számokat, és a tárgyszavakat. A szövegblokkban további szóközkaraktereket kellett beszúrni az egyes adattípusok közé (*12. ábra*).

Az összes cédulát érintő ellenőrzési és javítási fázis után következett a duplumszűrés művelete, melynek során immár különböző kritériumok alapján lehetett leválogatni az ismétlődő cédulákat. Az egyértelműen többször szereplő cédulákat töröltük az adatbázisból. A beszkennelt 16 320 db cédulából az ismétlődések kiszűrése és eltávolítása után 11 494 db maradt. A feldolgozás további folyamataiban már csak ezek vettek részt.

Cédulából HUNMARC rekord

A fenti előkészítő műveletek után következett csak a rekordok begyűjtés vagy létrehozás általi tényle-



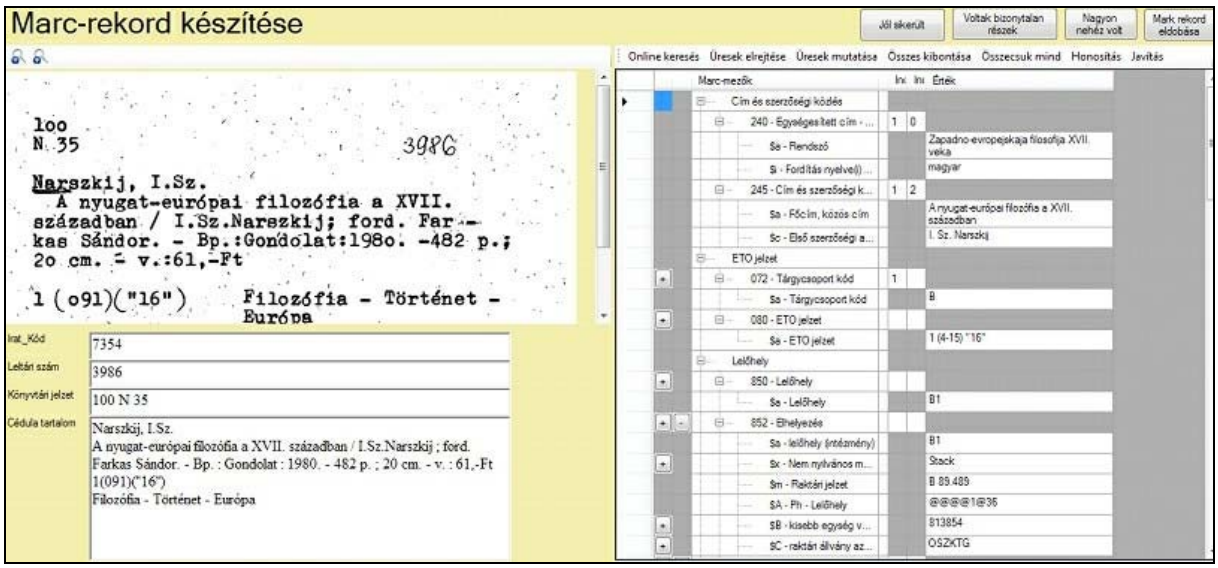
12. ábra **Rekord tagolása az ellenőrzés első fázisában (validálás)**

ges előállítása. Ennek alapja a munkák megkezdése előtt, a könyvtárral történt megállapodásokat, kívánságokat magába foglaló részletes feldolgozási szabályzat volt. Ez a szabályzat, a feldolgozási munka során a projekt előrehaladásával párhuzamosan maga is fejlődött, és kiegészült a feldolgozók által megjelölt általánosítható, mégis különlegesebb „esetek” figyelembevételével.

A hatékonyabb munka érdekében a PraktiDok rendszerbe közvetlenül is beépítésre került az elsődlegesnek tekintett OSZK katalógusának kereső és megjelenítő felülete. Egy találatot a szabályzat szerint csak akkor tekintettünk pontosnak, ha az legalább 5 paraméterben (szerző és címadatok, a megjelenési adatok és az ISBN vagy ISSN szám) megegyezett a kiinduló cédulán olvasható tartalommal. Az ISBN vagy ISSN szám pontos egyezése, a bármely oldali elírás lehetősége miatt önmagában nem volt elegendő a találattá minősítéshez (*13. ábra*). Pontos találat esetén az adatokat egy kattintással lekértük az OSZK gyűjteményéből és az adatfeldolgozó képernyőn a HUNMARC szerkezetnek megfelelően megjelenítve megkezdődtek a honosítás műveletének lépései (*14. ábra*).

Ennek kezdetén egy háttérben zajló automatikus műveletsor letárolta a forrásrekord azonosítóját, törölte az idegen könyvtár adatait és kapcsolati értékeit, helyüket pedig a MTF egyedi adataival töltötte fel. Ezt követően a megfelelő HUNMARC hívójelekben kerültek rögzítésre a cédulán lévő, a helyi kiadványra jellemző információk, úgymint a példányspecifikus adatok, ETO számok, tárgyszavak stb.

13. ábra **Találattá minősítés**



14. ábra **HUNMARC rekord létrehozása**

Ennek a munkafázisnak a végén a feldolgozók minősítették a saját maguk által létrehozott rekordokat, ami lényeges információként szolgált a következő fázisban dolgozó munkatársak, illetőleg az ezt követő automatikus folyamatok számára. Ha példának okáért valaki nem birkózott meg a feldolgozandó cédulával, akkor azt „eldobhatta”, a rendszer pedig ezeket később újból kiosztotta magasabb minősítésű munkatársak számára.

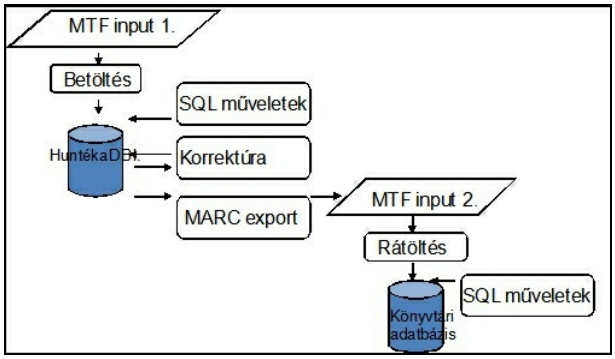
Az OSZK-ban nem található rekordok előállítására a fentebb már ismertetett források (FSZEK, KIT-KVK) szolgáltak, legrosszabb esetben pedig a feldolgozási szabályzat útmutatása alapján, a cédulán található tartalom bontásával kellett a megfelelő hívójelek alá sorolni az adatelemeket.

Az utolsó előtti fázis az elkészített rekordok ellenőrzése volt. A hivatalos forrásból származó, ISBN számmal rendelkező, készítője által jónak minősített rekordokat a PraktiDok rendszer automatikusan leválogatta és hibátlannak minősítette. A teljes munka befejezését követő tapasztalatok alapján ezek az automatikusan ellenőrzött cédulák valóban 100%-ban hibátlanok voltak. A készítők által gyengébbnek minősített, vagy ISBN szám nélküli, ám jó minősítésű letöltött rekordok ugyanakkor tételes ellenőrzésen estek át. Ebben a munkafázisban a könyvtárosi végzettségű munkatársak még egyszer összevetették a cédula képét és szövegtartalmát a kész HUNMARC rekord adataival, szükség esetén pedig javították az esetleges hibákat. Az ellenőrzési folyamat nehézségét és minőségét itt is cédulánként kellett minősíteni!

Az utolsó fázisban az elkészült és hibátlannak minősített rekordokból MARCXML adatcsomagokat generáltunk, melyek átadásra kerültek a könyvtári rendszer üzemeltetőjének.

Migráció a huntékába

A PraktiDok rendszerből generált MARCXML rekordjai a 15. ábrán látható folyamat során, két lépésben kerültek be a könyvtár huntéka adtabázisába.



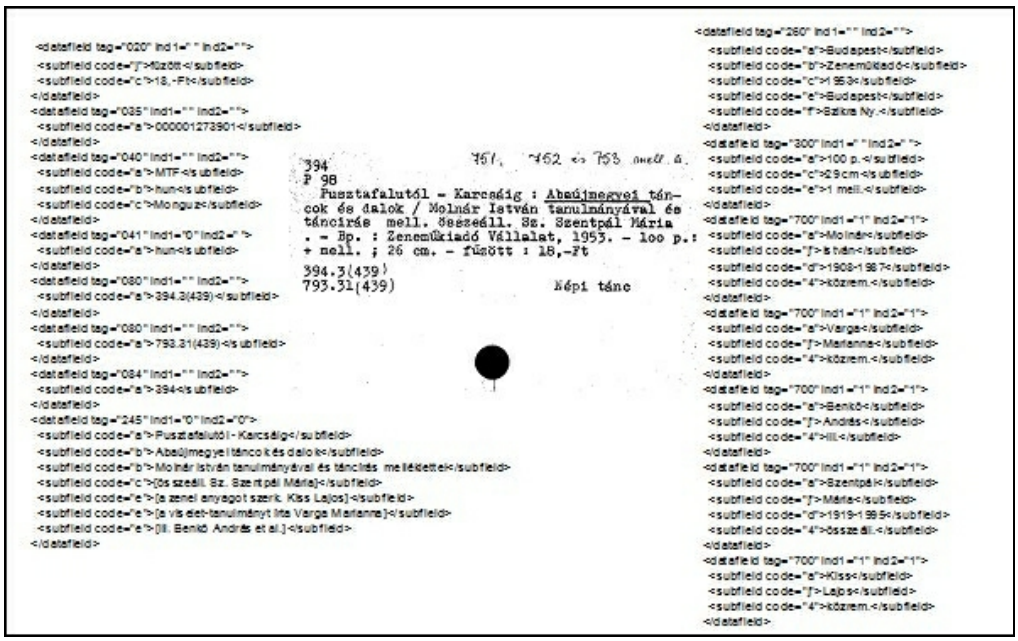
15. ábra **Huntéka migráció lépései a MTF-en**

Előzmények híjával a már honosított rekordokat nem volt mihez igazítani, tehát a könyvtári éles adatbázisba való integrálásuk nem okozott gondot, viszont jelentős többletinformációval bírtak az eredeti cédulához képest (*16. ábra*). Nemcsak a köte-

lező adatokra kell itt gondolni (pl. nyelv és országkód), hanem egységesített nevekre, valamint tartalmi feltárásra. A retrokonverzióból származó MARCXML az autopszián alapuló kottaadatbázisra itt is csak második lépésben lett rátöltve. Problémát csak a többkötetes kiadványokról készült leírások jelentettek, mert a kapcsolt rekordok szinte soha nem jöttek át, így végső soron ennek megkerülésére egy hibrid megoldást kellett alkalmazni a migráció során. A szabványosság és az adatbázis konzisztenciájának érdekében ezeket később kell pótolni. Egy későbbi javítás esetére a cédulák szkennelt képei természetesen itt is rendelkezésre állnak.

Az MTF projekt költségei

A költségeket a 2. táblázat foglalja össze. Mint látható, az egy kötetre jutó nettó költség 180,00 Ft. Ez az összeg sok élőmunkát tartalmaz, de összevetve egy leíró könyvtáros egy kötetre vetített 270,00 Ft-os átlagköltségével, és a feldolgozás eredményeként keletkező igen részletes és pontos rekordokkal, önmagáért beszél.



16. ábra **MTF MARCXML inputja és az eredeti cédula képe**

2. táblázat
MTF retrospektív konverziójának költségei

	Db	Megnevezés	Ft/db	Kötet
1	X	Cédulaszkennelés	4,00 Ft	16329
2	X	OCR gépi szövegbeolvasás	8,00 Ft	16329
3	X	Képi feltárolás PDS	3,00 Ft	16329
4	X	Indexelés PDS	15,00 Ft	16329
5	X	Katalóguscédula-validálás	25,00 Ft	11494
6	X	MARC rekordra bontás	40,00 Ft	11494
7	X	MARC rekordellenőrzés	75,00 Ft	11494
8	X	Rekordfeltöltés (HunTéka)	– Ft	
9	X	PraktiDok (program használati díj – kötet szerint)	10,00 Ft	11494
		Kötet ár összesen	180,00 Ft	
10	X	PraktiDok (konfiguráció, fejlesztés, tárhely, elérés), egyszeri díj	300 000,00 Ft	1
		Összesen:		

Összegzés

A két projekt tanulságai egyképpen összegezhetők: bár a technikai feltételek egyre fejlettebbek, az automatizmusok egyre kiforrottabbak, az emberi munka és értelem nem hagyható ki a retrokonverziós folyamatokból (sem). A karakterfelismertetés messze nincs még azon a szinten, hogy vakon megbízzunk benne. Az adatcsoportok felismerésére szolgáló algoritmusok és heurisztikák a tapasztalatok fényében tovább finomíthatók, de az OCR-t követő ellenőrzés hiányában nem hozhatják meg a kívánt eredményeket. Az MTF projekt ugyanakkor rávilágított arra a meglepő tényre, hogy a nagykönyvtárak állománya még napjainkban sem tekinthető mindenek felettinek, hiszen nagy számban vannak még olyan, a közelmúltban megjelent kiadványok, amelyek kívül esnek ezeken. A több forrásból „összevadászott" rekordhalmazoknál, így a rekordletöltésen és honosításon alapuló retrospektív eljárásoknál fokozottan kell őrködni az ilyen módon létrejövő besorolási állományok egysége felett. Bármelyik utat is választjuk állományunk visszamenőleges feldolgozásához, a biztonságot szem előtt tartva, nem kerülhető meg az a gond, amit a többlépcsős leírások rekordjainak betöltése okoz.

Irodalom

[1] BAKÓ Dorottya: Retrokonverziós körkép: német és svájci példák. = TMT, 51. köt. 9. sz. 2004. *http://tmt.omikk.bme.hu/show_news.html?id=3746&issue_id=454*

[2] BERKE Barnabásné: Első falat a nagy kalácsból. A nemzeti könyvtár cédulakatalógusainak retrokonverziós munkájáról. = Könyv, könyvtár, könyvtáros, 2004. augusztus. *http://epa.oszk.hu/01300/01367/00056/pdf/04muhelykerdesek.pdf*

[3] DANCS Szabolcs: Retrospektív konverzió nagyüzemi módon: az ADAM-projekt. = TMT, 57. köt. 2. sz. 2010. *http://tmt.omikk.bme.hu/show_news.html?id=5279&issue_id=512*

[4] BERKE Barnabásné: A könyvek cédulakatalógusának retrospektív konverziója az Országos Széchényi Könyvtárban. Networkshop, 2005. *https://nws.niif.hu/ncd2005/docs/ehu/026.pdf*

[5] BÁNKESZI Katalin – KOLTAY Klára: Mi újság a MOKKA háza táján? A közös katalógus továbbfejlesztése az Országos Dokumentumellátó Rendszer és a könyvtárak szolgálatában. = TMT, 58. köt. 2. sz. 2011. *http://tmt.omikk.bme.hu/show_news.html?id=5453&issue_id=523*

Beérkezett: 2012. V. 16-án.

Lengyel Mónika
a Monguz Kft. informatikusa.
E-mail: *lmoni@monguz.hu*

Andaházi Szeghy Viktor
a Hadtörténeti Könyvtár könyvtárvezetője.
E-mail: *andaszegi@yahoo.com*

Magyar Gábor
az MKB-LUX Kft. tulajdonosa, tanácsadója.
E-mail: *magyar.gabor@mkblux.hu*