Burmeister Erzsébet – Drótos László

EPA – MATARKA együttműködés: a közös cikkek ellátása egyedi azonosítókkal

Az időszaki kiadványokkal foglalkozó két országos szolgáltatás, az EPA (Elektronikus Periodika Archívum és Adatbázis: epa.oszk.hu) és a MATARKA (Magyar Folyóiratok Tartalomjegyzékeinek Kereshető Adatbázisa: www.matarka.hu) közötti együttműködés, mely 2004 óta létezik, újabb mérföldkőhöz érkezett: a közös folyóiratok esetében csaknem az összes cikk egyedi azonosítót kapott. Az EPA-ban nyilvántartott több mint 3200-féle kiadvány egynegyede saját szerveren teljes szöveggel archiválásra is kerül, s utóbbiak 53%-át teszik ki a MATARKA-val közös folyóiratok. A MATARKA adatbázisában pedig mintegy 25%-nyi az EPA-val közös kiadványok aránya. Nemrég befejeződött a két szolgáltatás közös cikkcímeinek (számuk átlépte a 450 ezret) egyedi azonosítóval való ellátása. 2016 közepe óta már eleve ilyen azonosítóval ellátva kerülnek át az EPA-ból az új folyóiratok adatai a MATARKA-ba, a retrospektív feldolgozás pedig 2018 elején ért véget. A cikk a célokat, az elvégzett munkát és ennek jelentőségét mutatja be.

Tárgyszavak: *Egyetemes Tizedes Osztályozás; ETO-jelzet; elemzés*

A MATARKA és az EPA rövid bemutatása

A MATARKA magyar vagy magyar nyelvterületen kiadott, nem feltétlen csak magyar nyelvű cikkeket tartalmazó szakfolyóiratok tartalomjegyzékeit dolgozza fel 2002 óta. Az idők folyamán sok évkönyv és egyéb periodikum is bekerült a szakfolyóiratok mellé: főleg múzeumok évkönyvei és felsőoktatási intézmények tudományos közleményei. A tartalomjegyzékek a feldolgozás után azonnal kereshetők a címben levő szavak és a szerzők neve alapján, a találatok különböző szempontok szerint szűkíthetők, a találati listák sokféle formátumban menthetők, ezáltal a listák további, többirányú felhasználása is lehetővé válik. Maguk a tartalomjegyzékek is böngészhetők és a szerzői illetve kulcsszó-indexek szintén kereshetők és böngészhetők. A cikk írásának időpontjában az adatbázisban 1843 folyóirat, 2 608 321 cím, 371 165 szerző és 758 747 ugrópont van a teljes szövegre.

Az EPA a Magyar Elektronikus Könyvtár (MEK) „folyóirat részlege". A MEK-en belül már a kilencvenes évek közepén megkezdődött a weben elérhető elektronikus időszaki kiadványok nyilvántartása. A részgyűjtemény önállóvá válása, vagyis saját adatbázisának és honlapjának fejlesztése 2003-ban indult. Az EPA-ban nagyrészt más szervereken is fellelhető kiadványok találhatók, de sok tételnek ez az archívum az egyetlen lelőhelye. Az archivált e-folyóiratok számára az EPA-szolgáltatás stabil hozzáférést, megjelenés- és formátumbeli egységet biztosít. A cikk írásának időpontjában az adatbázisban 869 archivált, 2283 élő vagy már megszűnt távoli elérésű és 115 csak offline hozzáférhető kiadvány, valamint több mint félmillió, a teljes szövegű cikk(ek)re mutató URL-cím található.

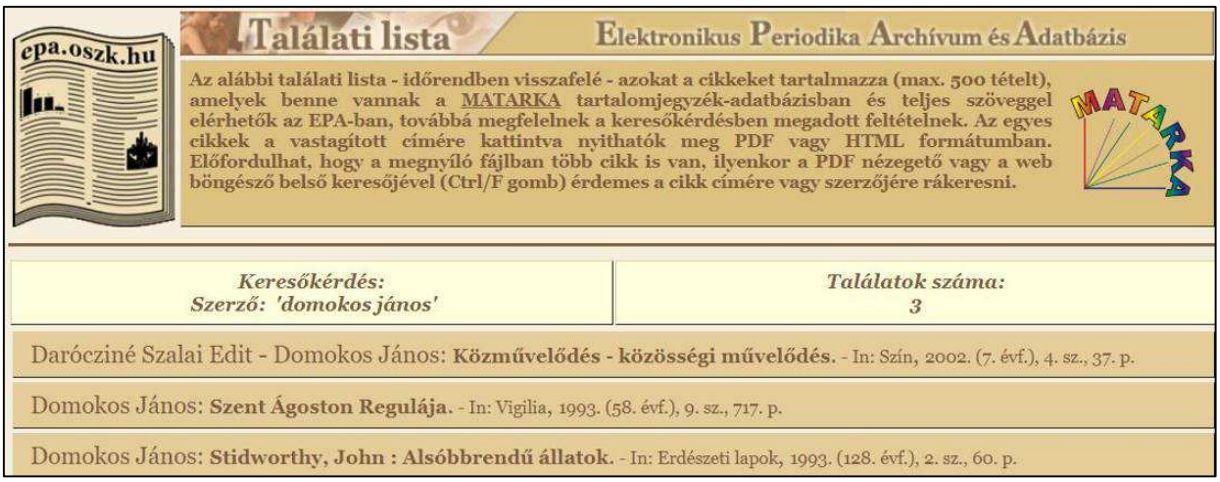
Az együttműködés formái

A két szolgáltatás 2004-től, tehát igen korán elkezdett együttműködni. [3] Ez az alábbi tevékenységeket jelenti:

- Adatcsere: a MATARKA átvesz az EPA-tól tartalomjegyzékeket, erre az elmúlt 15 év alatt több program is készült. Az EPA is átvesz a MATARKA-tól kész tartalomjegyzékeket, amire az idők folyamán szintén több program jött létre. Az EPA-nak természetesen utólag még a teljes szöveg PDF vagy HTML fájljait is hozzá kell linkelnie a cikkeihez. Jelenleg mind a két irányban

az EPA XML formátuma szolgál az adatcsere alapjául.

- Az EPA-ból szerző- és cikkszintű keresést lehet indítani a közös folyóiratokban a MATARKA kereső robotjával és a találati lista itt is az EPA külalakjával jelenik meg *(1. ábra).*
- A MATARKA-ból teljes szöveges keresést lehet indítani a közös folyóiratok EPA-s archívumában és a találati lista EPA-dizájnnal jelenik meg *(2 ábra).*



1. ábra **Az EPA-ból szerző- és cikkszintű keresés indítása a MATARKA-val közös folyóiratokban – találati lista**

2. ábra **A MATARKA-ból teljes szövegű keresés indítása az EPA-ban – találati lista**

Egyedi azonosítók

A publikációk egyre növekvő tömege már az elektronikus korszak előtt szükségessé tette a bibliográfiai adatok leírásának szabványosítását és egyedi azonosítók bevezetését. Jó példa erre az ISBN, ISSN stb. számok megjelenése a 20. század második felében. Az adatbázisok és az internet világában talán még fontosabb a szerzők, a földrajzi nevek, az elektronikus dokumentumok stb. egyedi azonosítása a gyors fellelhetőség érdekében. [1]

Szerzők esetében többféle, a személyek egyedi azonosítását megoldó szolgáltatás létezik, például ORCID (Open Researcher and Contributor ID), VIAF (Virtual International Authority File), ISNI (International Standard NameIdentifier). Magyarországon ilyen azonosító rendszert biztosít a felsőoktatási és akadémiai hálózat szerzőinek publikációit nyilvántartó MTMT (Magyar Tudományos Művek Tára).

A digitális dokumentumok – különösen a tudományos publikációk – azonosítására pedig a Handle rendszeren alapuló DOI (Digital ObjectIdentifier) terjedt el, melynek a „gazdája" az IDF (International DOI Foundation: www.doi.org). A DOI-n kívül használható még az URN (Uniform Resource-Name) vagy például az ARK (ArchivalResource) is ilyen célra.

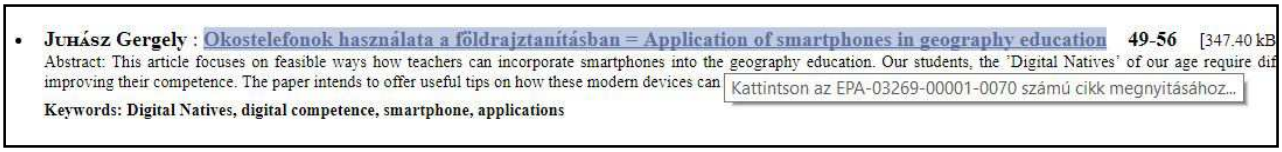
Ahhoz, hogy az EPA és a MATARKA cikkrekordjait egyértelműen összekapcsoljuk, szükségessé vált egy ilyen egyedi és stabil azonosító bevezetése. Bár a DOI használata már a magyar folyóiratokban is egyre elterjedtebb, ilyen azonosítóval az EPA-ban található cikkeknek csak töredéke rendelkezik, hiszen például a régi, papírról digitalizált periodikák, vagy a már ugyan eleve digitálisan születő, de nem tudományos jellegű kiadványok esetében természetesen nincs DOI-ja az egyes cikkeknek, és mivel a DOI-számok igénylése költségekkel jár, az nem is jöhetett szóba, hogy tömegesen igényeljünk ilyen azonosítót hozzájuk. Az URN használatát pedig végül azért vetettük el, mert bár az ingyenes és nagy tömegben generálható, viszont külön URL címen kell lennie minden dokumentumnak és be kell tenni az URN-azonosítót a dokumentumot tartalmazó weboldalba ahhoz, hogy az URN-szerver validálni tudja és „beélesítse". Ez a mechanizmus a MEK esetében a kezdetektől be van építve a rendszerbe, vagyis a könyvek automatikusan kapnak egy URN-t, amikor kikerülnek a nyilvános felületre, viszont az EPA-nál a cikkek esetében jelentős átalakításokat igényelt volna az utólagos bevezetése, tekintve, hogy a cikkeknek nincs önálló weblapjuk, hanem egy teljes folyóiratszám minden cikkadata egyetlen oldalon van felsorolva, továbbá arra is van példa, hogy az egyes cikkeknek nincs saját URL-címük sem, mert egy PDF-fájlban van a teljes füzetszám. Ezért a gyors megoldás érdekében egy saját EPAcikk_ID generálása mellett döntöttünk, amivel megoldható a két rendszer adatbázisainak összekapcsolása és megteremti annak a lehetőségét is, hogy az egyéb egyedi azonosítók (pl. DOI, ORCID, VIAF) is átvehetők legyenek azoknál a rekordoknál, amelyeknél léteznek ilyenek.

Egyedi cikkazonosító

Az EPAcikk_ID felépítése: EPA-XXXXX-YYYYY-ZZZZ, ahol
az első 5 jegyű szám a folyóirat alkönyvtárának sorszáma az EPA-ban (az EPA_ID).
A második 5 jegyű szám a füzet száma a folyóiraton belül.
A harmadik 4 jegyű szám a cikk sorszáma a tartalomjegyzéken belül (ez a sorszám tízesével nő, hogy szükség esetén be lehessen szúrni kifelejtődött vagy utólag megkapott cikkeket).

Egy konkrét példa: EPA-03269-00001-0070, ami a GeoMetodika folyóirat 2017. évi 1. számának 7. cikkét azonosítja, melynek címe: Okostelefonok használata a földrajztanításban.

Mind az EPA-ban, mind pedig a MATARKA-ban a cikk címe fölé mozgatva az egeret az azonosító is megjelenik *(3., 4. ábra).*



3. ábra **A Juhász Gergely által írt cikk azonosítójának megjelenítése az EPA-ban**

4. ábra **Ugyanezen cikk azonosítójának megjelenítése a MATARKA-ban**

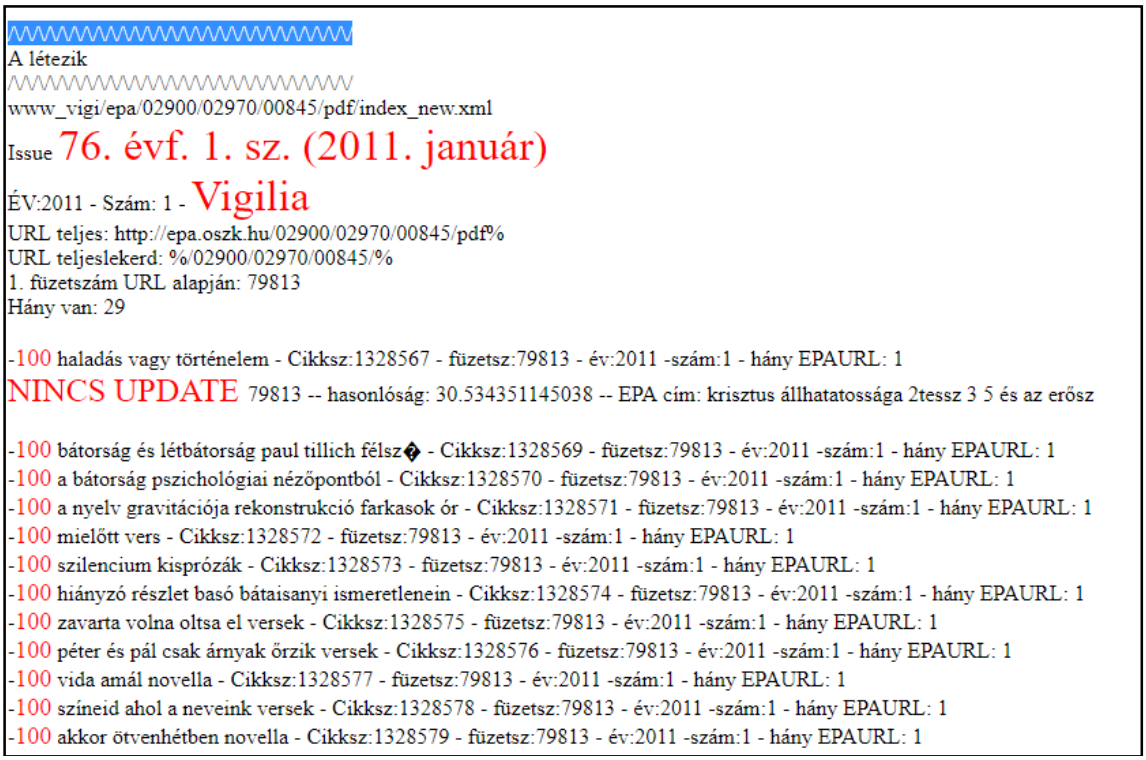
2015 őszén az EPAcikk_ID megtervezésekor a célunk az EPA és a MATARKA adatbázisok közös régebbi rekordjainak egyedi azonosítókkal való ellátása volt. Az új rekordok 2016 közepétől pedig már eleve az EPAcikk_ID-vel kerülnek át. Akkoriban azt tervezgettük, hogy a két adatbázist egyesítjük és a cikkazonosító bevezetése ezért elkerülhetetlen volt. Az adatbázisok összeolvasztásának mindenképpen lettek volna előnyei, például munkamegtakarítás mind a két oldalon, a folytonos adatcsere megszűnése, egyetlen adatbázis és szoftver üzemeltetése, és a felhasználóknak is csak egy szolgáltatást kellett volna használniuk.

Megvalósítás

A munkát egy, az *Internet Szolgáltató Tanácsa* (ISZT) által támogatott projekt keretében kezdtük el 2015 végén és folytattuk 2016-tól. Naivan az volt az elképzelésünk, hogy automatizált módszerekkel, egy erre a célra fejlesztett program segítségével mindkét adatbázis rekordjai rövid idő alatt ellátható lesznek az egyedi azonosítóval. Az első, erre a célra kifejlesztett megoldás csődöt mondott, mert csak a rekordok feléhez sikerült EPA-azonosítót hozzárendelni. A Burmeister Erzsébet által kifejlesztett php programmal lehetett végül a teljes munkát elvégezni. Az EPA-adatokat ehhez a MEK és az EPA informatikai hátterét nyújtó *Vitéz Bt.*, személy szerint *Vitéz Gáborné* szolgáltatta. A manuális munkákban nagyon sokat segített a *Miskolci Egyetem* könyvtárának közfoglalkoztatott kollégája: *Nagy Zsolt*. Az EPA-csapatot *Uri-Kovács József* képviselte az egyeztetések során.

A munka majd két évig tartott. Hamar kiderült, hogy a teljes automatizálás lehetetlen. Először lefutott a program a MATARKA tesztszerverén, amely az összehasonlítást végezte el adott folyóirat számainak MATARKA és EPA tartalomjegyzékei között. AZ EPA-ból ehhez minden folyóirathoz *index_new.xml* nevű XML fájlokat kapott a MATARKA, alkönyvtárakba rendezve (*5. ábra*). A program az XML fájlok tartalmát hasonlította össze a MATARKA-adatbázis megfelelő tábláinak tartalmával (*6. ábra*). (A MATARKA relációs adatbázis, táblái SQL-parancsokkal lekérdezhetők.) Ahol 95% feletti egyezést talált, ott rögtön aktualizálta az URL-t és a megfelelő mezőbe beírta az EPAcikk_ID-t. Ahol nem volt egyezés, azoknál utólag ún. *update* parancsok készültek, melyek a megadott cikkszámokhoz beszúrták az EPAcikk_ID-ket. Ezeket az update parancsokat egyszerre le lehetett futtatni a MATARKA éles szerverén. Folyóiratonként a munka ellenőrzését egy külön php script futtatásával lehetett elvégezni és a még előbukkanó hibákat további update parancsokkal javítani (*7. ábra*).

5. ábra **A Vigilia 2011/1. számához tartozó XML fájl**



6. ábra **Képernyőrészlet: a Vigília folyóirat tartalomjegyzékeinek összehasonlítása az EPA XML fájljaival**

```
update cikk setEPAcikk_ID="EPA-02362-00032-0090" wherecikksz=272437;
update cikk setEPAcikk_ID="EPA-02362-00032-0100" wherecikksz=272438;
update cikk setEPAcikk_ID="EPA-02362-00032-0110" wherecikksz=272439;
update cikk setEPAcikk_ID="EPA-02362-00032-0120" wherecikksz=272440;
update cikk setEPAcikk_ID="EPA-02362-00032-0130" wherecikksz=272441;
update cikk setEPAcikk_ID="EPA-02362-00032-0140" wherecikksz=272442;
update cikk setEPAcikk_ID="EPA-02362-00032-0150" wherecikksz=272443;
```

7. ábra **Az *update* parancsok sorozata a 2362 számú EPA folyóirat 32 számú füzeténél**

Néhány megállapítás

- A közös folyóiratok egy része nem az EPA-ból került át közvetlenül a MATARKA-ba, hanem más könyvtárak dolgozták fel őket és ezért a címleírások eltértek, rövidebben vagy hosszabban, párhuzamos címmel vagy anélkül készültek stb. Itt ki kell hangsúlyozni, hogy TARTALOMJEGYZÉKEK BIBLIOGRÁFIAI LEÍRÁSÁRA NINCS SZABVÁNY!! Ezért a cikkcímek programmal történő összehasonlítása sokszor nem hoz kielégítő eredményt.
- URL-cím alapján történő összehasonlítás is csak olyan URL-eknél lehetséges, melyek egyediek, vagyis nem több cikket fognak össze.
- A közvetlenül (és teljes egészében) az EPA-ból az évek során átkerült folyóiratok esetében az EPAcikk_ID-k átemelése automatizálható volt, a program 100 százalékosan átvette ezeket.
- Minden folyóiratra külön le kellett futtatni a programot és a sikertelen átvételeknél manuálisan kellett az EPAcikk_ID-ket pótolni.

A munka abból a szempontból is nagyon sikeres volt, hogy melléktermékként mind az EPA-ban, mind a MATARKA-ban levő (elég sok) hibát tudtunk kijavítani. Ehhez egy, a Google Drive-on megosztott közös fájlt használtak a MATARKA-s és EPA-s munkatársak, amelyet a továbbiakban a napi átvételek során észrevett hibák esetében is használni fognak.

A munka végén készült egy összesítés a feldolgozott folyóiratokról (2017 január végén 460 db ilyen volt), illetve egy kimutatás arról, hogy mely folyó-

iratoknál nem sikerült minden cikkhez EPAcikk_ID-t hozzárendelni. Két oka volt ennek:

- A MATARKA bővebb, több cikkhez tartozik ugyanaz az URL, míg az EPA összefoglalva írta le a cikkeket. Ilyenkor csak az első cikkhez lett EPAcikk_ID hozzárendelve, hisz az egyedi azonosító csak egy cikkhez tartozhat.
- Az EPA-ban az adott füzet nincs „felszeletelve", vagyis ugyanaz az egy URL tartozik minden cikkhez, amely a teljes füzetre linkel. Ilyenkor nem létezik *index_new.xml* a tartalomjegyzékhez (pl. Partium folyóirat).

Összességében 2018. január 25-én a fő jellemző számok:
A MATARKA-ban EPA URL-lel rendelkezik 456 711 rekord. Ezek közül EPAcikk_ID-t tartalmaz 444 475 rekord. A cikkek 97,3 %-a rendelkezik EPAcikk_ID-vel. (Vannak olyan EPAcikk_ID-k is, melyeknél az EPA-adatbázisban és ezért a MATARKA-ban sincs URL, természetesen ezek is át lettek véve.)

Összegzés

Az egyedi cikkazonosító bevezetése mindkét szolgáltatás esetében egy-egy nagyobb folyamat egyik lépéseként is tekinthető. A MATARKA-ban néhány éve intenzíven folyik az egyedi azonosítók (pl. VIAF, ORCID, MTMT) hozzákapcsolása a magyar szerzők neveihez, hogy megkülönböztethetővé váljanak az azonos nevű személyek, illetve összerendelhetők legyenek ugyanazon személy különböző névváltozatai. [2] A névazonosítók és a már kezdettől fogva nyilvántartott DOI cikkazonosítók mellett logikus lépés volt, hogy az EPA-ban teljes szöveggel megtalálható cikkek mindegyike kapjon egy egyedi azonosítót. Az EPA esetében pedig egy-két éve elindult egy adatkonszolidációs munka, ami az Országos Széchényi Könyvtár 2018-ban bevezetett új könyvtári és digitális könyvtári rendszerébe való átköltözést készíti elő. Ennek keretében zajlik a metaadatok javítása és egységesítése, és ide tartozik az egyedi azonosítók ügye is. Az EPAcikk_ID bevezetésével egyrészt könnyen beazonosíthatók és linkelhetők lesznek a MATARKA-ból a cikkek akkor is, ha ezek a digitális dokumentumok átkerülnek az új rendszerbe és megváltozik az URL-címük. Másrészt lehetőség lesz arra, hogy a cikkek EPA-s metaadatai közé átemeljük a MATARKA-ból a személynevekkel összekapcsolt névtér-azonosítókat, valamint a DOI-számokat is, azoknál a cikkeknél, amelyeknél már vannak ilyenek.

Irodalom

1. BURMEISTER Erzsébet: Egyedi szerző- és dokumentumazonosítók használata a magyar könyvtári adatbázisokban – A 2015. nov. 10-én Kaposváron megrendezett konferencián elhangzott előadás átdolgozott, lektorált változata.
   In: Könyvtárak a tudomány és a felsőoktatás szolgálatában. – Kaposvár : KE Egy. Kvt., 2016, p. 27- 35.
   http://lib.ke.hu/emimg/konferencia/Konyvtarak_Tanulmanyok_2016_BELIV_P001-144_NYOMDA.pdf (Letöltve: 2018. 02. 12.)
2. BURMEISTER Erzsébet: Szerzők nevének egységesítése, szerzők szétválasztása, egyedi azonosítók.
   In: Tudományos és műszaki tájékoztatás, 2016. 6-7. sz. p. 244-250.
   http://tmt.omikk.bme.hu/tmt/article/view/77 (Letöltve: 2018. 02. 12.)
3. URI-KOVÁCS József: A MATARKA és az EPA közötti együttműködés.
   Magyar Könyvtárosok Egyesülete 49. Vándorgyűlése, Miskolc, 2017. július 6.
   http://mek.oszk.hu/html/irattar/eloadas/2017/VGY_UKJ_2017_V1.pptx (Letöltve: 2018. 02. 12.)

Beérkezett: 2018. II. 20-án.

Burmeister Erzsébet
a Miskolci Egyetem
Könyvtár, Levéltár, Múzeum
főkönyvtárosa.
E-mail: *erzsi@uni-miskolc.hu*

Drótos László
könyvtáros
OSZK – E-könyvtári Szolgáltatások Osztály.
E-mail: *drotos.laszlo@oszk.hu*