DOI: 10.31915/NWS.2019.17

Tömeges adatkonverzió és rugalmas export-import lehetőségek az EPrints, OJS és Omeka szoftverek körében

Nagy Gyula
SZTE Klebelsberg Könyvtár
gyula.nagy@ek.szte.hu
ORCID: 0000-0002-8391-2851

Nagy Dóra
SZTE Klebelsberg Könyvtár
dora.nagy@ek.szte.hu

Sándor Ákos
SZTE Informatikai és Szolgáltatási Igazgatóság
akos.sandor@ek.szte.hu

Possibilities of massive data conversion and flexible processes of export-import regarding EPrints, OJS and Omeka software
Since we have been using different types of systems to provide our digital contents, we had to migrate our data several times over the years. The lesson to be learned is ensuring interoperability and data exchange between these systems is a constant priority for libraries. Uploading several records at a time have become common practice in our library by now.

Each year since 2012 thousands of degree theses have been received from Modulo, which is an online platform assisting students and staff of the University. We have to convert the original XML file to a structure which is compatible with our EPrints based repository. The handling of SZTE Repository of Papers and Books, Miscellanea and UnivHistória repositories follows a similar procedure, but with different initial conditions. Last year we have started to work with Open Journal System, therefore establishing an efficient data conversion method between EPrints and OJS and vice versa was necessary.

One of our long-term plans was to have a search engine which can discover all of our repositories and we were able to achieve this with the help of the Vufind, which is an open-source search engine especially for libraries. A critical point of the project was to develop a data exchange format for MARC using OAI-PMH in EPrints.

Our previous Marc-based databases (Bodza) were exported to EPrints years ago, and with that experience we were able to start building our new Omeka-based photo archive. In this presentation we demonstrate the above mentioned processes through a few practical examples.

Keywords: repository, data conversion, data import and export, bulk data import, EPrints, Open Journal System, Omeka, VuFind, MARC, OAI-PMH

Bevezetés

Az SZTE Klebelsberg Könyvtár digitális tartalmainak szolgáltatása terén az évek során különféle szoftveres megoldásokat használtunk, így többször előfordult, hogy az adatok migrálására volt szükség. Fontos tanulság, hogy a különböző rendszerek közötti átjárhatóság és adatcsere biztosítása folyamatosan kiemelt feladatként van jelen a könyvtárak munkájában. Ezen migrálások mellett mára bevett gyakorlattá vált az új rekordok tömeges adatbetöltése repozitóriumainkba, amely munkamenetnek mindig az adott archívum sajátosságaihoz kell illeszkednie. Többek között 2012 óta ilyen módon zajlik az évente több ezer kurrens szakdolgozat átvétele a Szegedi Tudományegyetem tanulmányi rendszeréből. A hallgatók diplomamunkáikat a Modulo-ba töltik fel, amelyek a Neptunból származó, összefésült metaadatokkal együtt kerülnek exportálásra. Az így kapott XML fájlt EPrints XML formátumra (EP3 XML) szükséges konvertálni. Az adatkonverzió után történik a szakdolgozatok tömeges betöltése. Hasonló elvek mentén, de más kiinduló feltételekkel történik az SZTE Egyetemi Kiadványok, az SZTE Miscellanea, az SZTE UnivHistória és a Tiszatáj archívumának kurrens és retrospektív gyarapítása is.

A tavalyi évben egy EFOP 3.6.3 projektnek köszönhetően elindult az SZTE OJS folyóirat-platform. A Szegedi Tudományegyetemhez köthető folyóiratok archívumának gyors kiépítése és a hatékony munkavégzés megteremtése miatt szükség volt az EPrints-OJS, majd a repozitóriumok gördülékeny gyarapításának biztosítása miatt az OJS-EPrints automatizált konverziós irányok megteremtésére is. Az adatcsere egy másik megközelítését alkalmazva, repozitóriumaink speciális tagoltsága miatt régi tervünk volt ezek közös kereshetőségének biztosítása, amelyet egy EFOP 3.4.3. projekt keretében a VuFind rendszer segítségével valósítottunk meg. A projekt kritikus eleme volt az EPrints OAI-PMH kimenetén előállított MARC formátumok adatcsere lehetőségének kidolgozása.

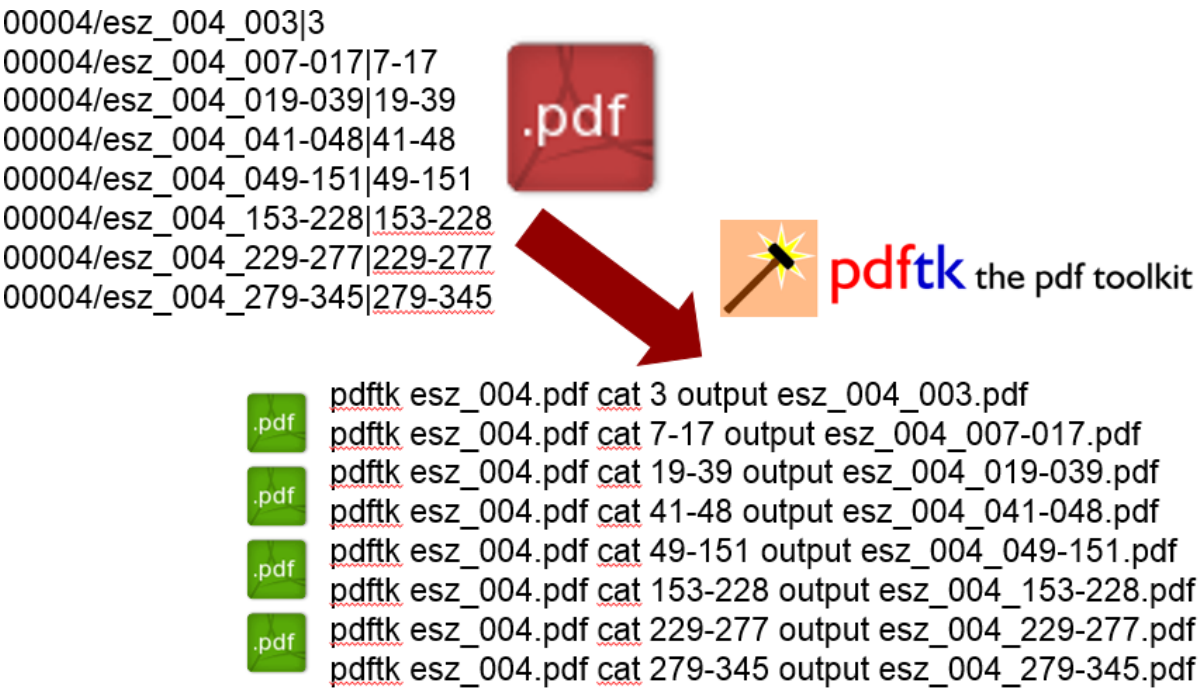
Az eddig említett szoftverek körében szerzett gyakorlat és korábbi MARC alapú adatbázisaink (pl. Bodza) tartalmának EPrints-be való átköltöztetése évekkel ezelőtt megtörtént, amely tapasztalatokat már fel tudtuk használni új, Omeka alapú képgyűjteményünk (SZTE Képtár és Médiatéka) létrehozásakor. Tanulmányunkban a fent említett projektek munkafolyamatai közül a legtanulságosabbakat kíséreljük meg bemutatni, elsősorban a tömeges adatkonverzió és a rugalmas export-import lehetőségek területéről.

1. Repozitóriumi metaadatbetöltési gyakorlat

Érdekes volna alaposabban körüljárni a hazai (és nemzetközi) repozitóriumok gondozói által használt tömeges adatbetöltési gyakorlatot, hogy mely intézményi repozitóriumoknál számít bevett rutinnak a tömeges export-import lehetőségek kihasználása, illetve a legtöbb szabványos repozitóriumrendszer esetében rendelkezésre álló OAI-PMH protokoll által biztosított potenciál kiaknázásának jó példái is tanulságosak lennének. Sajnos erre jelen tanulmány keretében nincs módunk, helyette elsősorban saját, ilyen irányú megoldásainkról tudunk csak beszámolni.

A metaadatok repozitóriumba való injektálásának egyik kézenfekvő módját jelentheti egy adott könyvtár OPAC-jából származó katalógusadatok repozitóriumi betöltése. Erre kiváló példaként szolgál az MTA Könyvtár és Információs Központban az Aleph és az EPrints rendszerek közötti adatcsere kapcsolat megteremtése vagy az Országos Széchényi Könyvtár adatbázis konszolidációs projektje kapcsán alkalmazott megoldások[1]. Egy másik kiválóan működő példa lehet a metaadatok automatizált továbbítására az MTMT felőli, SWORD[2] protokollon át történő intézményi repozitóriumokba való adattovábbítás, ahol valójában nemcsak metaadatok utaznak a hálózaton keresztül, hanem a publikációk teljes szövegét tartalmazó fájlok is részei a továbbított csomagnak.

Saját tömeges adatbetöltési gyakorlatunk alapját a minél automatizáltabb megközelítés adja. Egyes repozitóriumok esetében már a kezdetektől célul tűztük ki a teljes analitikus feldolgozást, amihez az egyes folyóiratokat, egyetemi actákat és tanulmánykötetet magától értetődő módon, fájlszinten is részekre kellett bontani. Ehhez a PDFtk[3] nevezetű parancssori eszközt használjuk. Ennek szintaktikája és működési módszere az 1. ábrán látható.



1. ábra – A PDFtk segédprogram alkalmazása: a cikkek logikai és fizikai oldalhatárai és a parancssorban futtatható, kötegelt fájl szintaxisa

1 Balázs László. Adatbázis konszolidáció az OSZK-ban. Networkshop 2017, Debrecen, 2016.03.29.-2016.04.01. Hozzáférés: 2019.06.17. https://kifu.videotorium.hu/hu/recordings/12965/adatbazis-konszolidacio-az-oszk-ban

2 Allinson, Julie, Sebastien François, and Stuart Lewis. "SWORD: Simple Web-service offering repository deposit." (2008). Hozzáférés: 2019.06.17. http://scholarworks.csun.edu/handle/10211.3/118201

3 PDFtk – The PDF Toolkit. https://www.pdflabs.com/tools/pdftk-the-pdf-toolkit

Az ilyen módon előállított PDF fájlok alapját képezik a következő lépésnek, hiszen az akár több ezer soros fájllistából egy ugyanennyi soros XLS vagy CSV fájlt állítunk előtt, ahol az egyes mezők fogják tartalmazni az egyes metaadat-elemeket. A módszer segítségével jó néhány adatelem tömegesen, illetve fél-automatikusan kitölthető (pl. számozási adatok, típusra vonatkozó adatok, azonosítók, stb.). Mivel az analitikus feldolgozás mellett fontosnak tartjuk a borítótól-borítóig terjedő teljes kötetek repozitóriumi megőrzését is, ezért ezek is bekerülnek egy, a cikkek metaadatait tartalmazó táblázathoz hasonló listába (a 2. ábrán piros karikával jelöltük a két megközelítés közötti különbségeket). A teljes lapszámokat és köteteket „full", míg az egyes tanulmányokat, cikkeket „part" néven hivatkozzuk. Ezek az elnevezések tükröződnek a táblázatok elnevezésében, illetve a későbbi munkafolyamatokban keletkező különböző kimenetekben is.

A következő lépésben az ilyen módon előállított, akár több tízezer soros táblázatokból elő kell állítanunk az EPrints által az automatikus adatbetöltések esetében preferált EP3 XML fájlokat. Ez egy repozitóriumonként meghatározott XML-skeleton alapján történik, a „full" és a „part" táblázatok eltérő adatelemeit természetesen ez az XML-skeleton is követi. Az elkészített XML-sémából az XMLBlueprint[4], vagy újabban a saját fejlesztésű CSV2XML Python-alapú segédeszköz segítségével készülnek el az EP3 XML fájlok, amelyeket az EPrints importfelülete már fogadni tud. A betöltések során a teljes szövegű PDF-ek is automatikusan bekerülnek a rekordokba, melyet úgy oldunk meg, hogy a <documents> rész megfelelő <url> tag-jében egy általunk üzemeltetett Apache webszerveren elhelyezett, csak a betöltés idejéig élő URL címek találhatóak. Természetesen ehhez a megoldáshoz a repozitórium oldalán engedélyezni kell a web-import lehetőséget. Az itt röviden felvázolt módszert használjuk évek óta mind a kurrens, mind a retrospektív betöltések és időnként a migrálások esetében is, melynek segítségével immár több százezer rekordot tettünk közzé különböző archívumainkban.

	A	B	C	D	E	F	G	H	I	J	K
1	1 - type	2 - title	3 - dat	4 - publication_full	5 - volume	6 - number	7 - issn	8 - isbn	9 - format	10 - security	11 - filename
2	book	Aetas - 33. évf. (2018) 1.sz.	2018	Aetas	33	1	0237-7934		full	public	aetas_2018_001.pdf
3	book	Aetas - 33. évf. (2018) 2.sz.	2018	Aetas	33	2	0237-7934		full	public	aetas_2018_002.pdf
4	book	Aetas - 33. évf. (2018) 3.sz.	2018	Aetas	33	3	0237-7934		full	public	aetas_2018_003.pdf
5	book	Aetas - 33. évf. (2018) 4.sz.	2018	Aetas	33	4	0237-7934		full	public	aetas_2018_004.pdf
6	book	A Szegedi Alföldkutató Bizottság	1930	A Szegedi Alföldkutató Bizottság ko	1				full	public	alfoldkutato_001_001
7	book	A Szegedi Alföldkutató Bizottság	1928	A Szegedi Alföldkutató Bizottság ko	1				full	public	alfoldkutato_002_001
8	book	A Szegedi Alföldkutató Bizottság	1928	A Szegedi Alföldkutató Bizottság ko	2				full	public	alfoldkutato_002_002
9	book	A Szegedi Alföldkutató Bizottság	1928	A Szegedi Alföldkutató Bizottság ko	3				full	public	alfoldkutato_002_003
10	book	A Szegedi Alföldkutató Bizottság	1928	A Szegedi Alföldkutató Bizottság ko	4				full	public	alfoldkutato_002_004
11	book	A Szegedi Alföldkutató Bizottság	1929	A Szegedi Alföldkutató Bizottság ko	5				full	public	alfoldkutato_002_005
12	book	A Szegedi Alföldkutató Bizottság	1930	A Szegedi Alföldkutató Bizottság ko	7				full	public	alfoldkutato_002_007
13	book	A Szegedi Alföldkutató Bizottság	1930	A Szegedi Alföldkutató Bizottság ko	8				full	public	alfoldkutato_002_008
14	book	A Szegedi Alföldkutató Bizottság	1931	A Szegedi Alföldkutató Bizottság ko	9				full	public	alfoldkutato_002_009
15	book	A Szegedi Alföldkutató Bizottság	1928	A Szegedi Alföldkutató Bizottság ko	2				full	public	alfoldkutato_003_002

full

	A	B	C	D	E	F	G	H	I	J	K	L
1	1 - type	2 - title	3 - dat	4 - publication	5 - volume	6 - numbe	7 - pagerange	8 - issn	9 - isbn	10 - format	11 - security	12 - filename
2	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	5-18	2064-5929		part	public	belvedere_2018_00
3	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	19-39	2064-5929		part	public	belvedere_2018_00
4	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	40-60	2064-5929		part	public	belvedere_2018_00
5	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	61-82	2064-5929		part	public	belvedere_2018_00
6	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	83-95	2064-5929		part	public	belvedere_2018_00
7	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	96-107	2064-5929		part	public	belvedere_2018_00
8	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	108-123	2064-5929		part	public	belvedere_2018_00
9	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	124-140	2064-5929		part	public	belvedere_2018_00
10	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	141-159	2064-5929		part	public	belvedere_2018_00
11	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	160-180	2064-5929		part	public	belvedere_2018_00
12	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	181-190	2064-5929		part	public	belvedere_2018_00
13	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	191-202	2064-5929		part	public	belvedere_2018_00
14	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	203-210	2064-5929		part	public	belvedere_2018_00
15	article	feldolgozásra vár	2018	Belvedere Meridionale	30	4	211-215	2064-5929		part	public	belvedere_2018_00

part

2. ábra – A teljes és rész PDF-ek metaadatait tartalmazó táblázatok

4 XML Editor – XMLBlueprint. https://www.xmlblueprint.com

full

```
<?xml version='1.0' encoding='utf-8'?>
<eprints xmlns='http://eprints.org/ep2/data/2.0'>
{repeat}
<eprint>
    <eprint_status>archive</eprint_status>
    <metadata_visibility>show</metadata_visibility>
    <type>$(column_1)</type>
    <full_text_status>public</full_text_status>
    <title>$(column_2)</title>
    <date>$(column_3)</date>
    <publication_full>$(column_4)</publication_full>
    <volume>$(column_5)</volume>
    <number>$(column_6)</number>
    <issn>$(column_7)</issn>
    <isbn>$(column_8)</isbn>
    <documents>
        <document>
            <format>$(column_9)</format>
            <security>$(column_10)</security>
            <files>
                <file>
                    <filename>$(column_11)</filename>
                    <url>$(column_12)</url>
                </file>
            </files>
        </document>
    </documents>
</eprint>
{repeat}
</eprints>
```

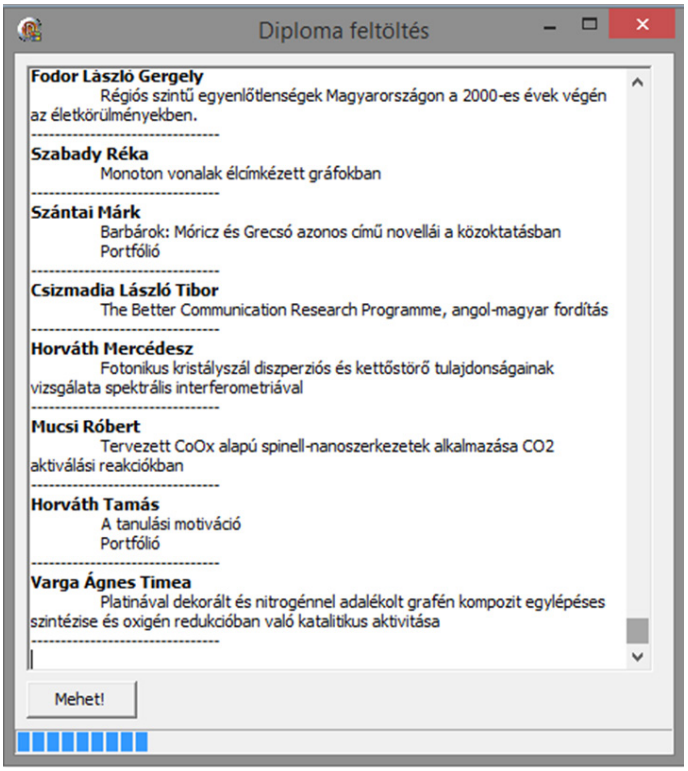
part

```
<?xml version='1.0' encoding='utf-8'?>
<eprints xmlns='http://eprints.org/ep2/data/2.0'>
{repeat}
<eprint>
    <eprint_status>archive</eprint_status>
    <metadata_visibility>show</metadata_visibility>
    <type>$(column_1)</type>
    <full_text_status>public</full_text_status>
    <title>$(column_2)</title>
    <date>$(column_3)</date>
    <publication>$(column_4)</publication>
    <volume>$(column_5)</volume>
    <number>$(column_6)</number>
    <pagerange>$(column_7)</pagerange>
    <issn>$(column_8)</issn>
    <isbn>$(column_9)</isbn>
    <documents>
        <document>
            <format>$(column_10)</format>
            <security>$(column_11)</security>
            <files>
                <file>
                    <filename>$(column_12)</filename>
                    <url>$(column_13)</url>
                </file>
            </files>
        </document>
    </documents>
</eprint>
{repeat}
</eprints>
```

3. ábra - A "full" és a "part" betöltéseknél alkalmazott XML-skeletonok, kiemelve az eltérő XML tag-eket

Az SZTE Diplomamunka repozitórium[5] kurrens gyarapítására egy, a bemutatotthoz elviekben nagyon hasonló, ám más szoftveres megoldáson nyugvó módszert használunk, mivel ennek kimunkálása időben megelőzte a fenti módszer általános bevezetését. A Szegedi Tudományegyetem hallgatói szakdolgozatukat a Modulo adminisztrációs rendszerben adják le, ami szoros kapcsolatban áll a Neptun tanulmányi rendszerrel. Könyvtárunk az ezekből a rendszerekből exportált metaadatokat és teljes szövegű fájlokat kapja meg 2012 óta, amely több tízezres kurrens gyarapodást tett lehetővé a szakdolgozatokat tároló adatbázisunk esetében. Terveink szerint a 2019-es év végére a kurrens és retrospektív projektek keretében gyarapított rekordok száma eléri az 55 ezret.

5 SZTE Diplomamunka Repozitórium. http://diploma.bibl.u-szeged.hu

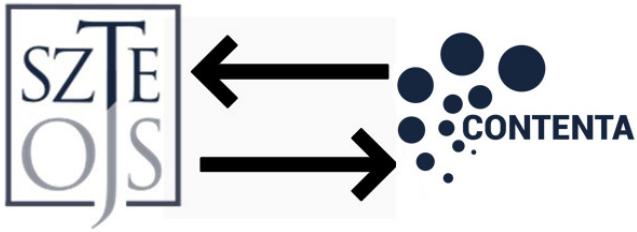


4. ábra - A szakdolgozatok metaadatainak konvertálása

2. A repozitóriumok és az OJS közötti kétirányú adatkapcsolat megteremtése

Mivel a módszer bizonyított, ezért a felmerült igények alapján elkezdtük kiterjeszteni más általunk használt szoftverekre is. A 2018-as év nyarán beindított Open Journal System folyóirat-szerkesztőségi keretrendszer használatakor már kézenfekvő lehetőségként merült fel a fentiekben bemutatott módszeren keresztüli archívumépítés a platformra beköltöző 15 folyóirat mintegy 20 ezer körüli tanulmánya esetében. Ez a következő munkafolyamatot jelentette: EPrints EP3 XML export → EP3 XML táblázattá alakítása[6] → Manuális korrekciók → OJS natív XML fájl előállítása → OJS XML import.

A munkafolyamat kvázi megfordításával pedig az OJS platform alatt megjelenő kurrens lapszámok EPrints repozitórium alatti archiválásának folyamatos biztosítását tudjuk automatizálni: OJS DOAJ Export Plugin XML → XML fájl táblázattá alakítása[4] → Manuális korrekciók → EP3 XML fájl előállítása → EPrints XML import.



5. ábra – Az SZTE OJS folyóirat-platform és az SZTE Contenta rendszerek közötti kétirányú adatkapcsolat

6 Convert CSV/Excel To... http://www.convertcsv.com

3. Tömeges adatbetöltési lehetőségek az Omeka rendszer esetében

A 2019-es év tavaszán elindított SZTE Képtár és Médiatéka[7] szolgáltatás alapjául az Omeka Classic[8] nyílt forráskódú rendszert választottuk. Több tízezernyi fotónk metaadatai rendelkezésre álltak MARC formátumban, ezért mindenképpen meg kellett teremtenünk ezek Omekába való injektálásának lehetőségét. Mivel az Omeka rendszer CSVImport+ pluginja metaadatokat, fájlokat és geokódokat is tud importálni, továbbá egy rekordhoz több fájl is betölthető egyszerre, sőt akár az adatok csoportos módosítására is lehetőség van, valamint az adminisztrátori felületen visszavonható a betöltés, így nagy rugalmasságot biztosítottak ezek a funkciók. Ehhez „csupán" arra volt szükség, hogy a Bodza rendszerből kiexportált MARC rekordokat a MarcEdit ingyenes szoftvercsomag[9] segítségével – illetve a megfelelő MARC mezők és almezők Dublin Core megfeleltetésével – import-kompatibilis, táblázatos formára alakítsuk. A kurrens, kötegelt betöltéseknél már eleve alkalmazható ez a táblázatos forma, hiszen az említett CSVImport+ plugin kiválóan fogadni tudja az így betöltött rekordokat.

Dublin Core:Identifer	Dublin Core:Creator	Dublin Core:Title	Dublin Core:Medium	Dublin Co	Dublin Core	Dublin C	Dublin Core:Extent
shvoy-001-001-001	Shvoy Kálmán	Autótúra 1930. Ausztria, Németország (dél)	címlap	1930	p. 1.		32,1x24,5 cm (3702
shvoy-001-003-000	Shvoy Kálmán	1. album, 3. oldal	oldalkép	1930	p. 3.		35x26 cm (3816x25
shvoy-001-003-001	Shvoy Kálmán	Aspang nyaralóhely	fénykép	1930	p. 3.		13x8,1 cm (2656x16
shvoy-001-003-002	Shvoy Kálmán	Aspang nyaralóhely	fénykép	1930	p. 3.		13x8,1 cm (2600x16
shvoy-001-003-003	Shvoy Kálmán	Aspang nyaralóhely	fénykép	1930	p. 3.		13x8,2 cm (3420x21
shvoy-001-003-004	Shvoy Kálmán	Aspang nyaralóhely	fénykép	1930	p. 3.		13x8 cm (3368x209
shvoy-001-004-000	Shvoy Kálmán	1. album, 4. oldal	oldalkép	1930	p. 4.		35x26 cm (3264x23
shvoy-001-004-001	Shvoy Kálmán	Alsó Aspang nyaralóhely	fénykép	1930	p. 4.		13x8,1 cm (3444x21
shvoy-001-004-002	Shvoy Kálmán	Kirchberg am Wechsel, Schneeberg	fénykép	1930	p. 4.		13,1x7,8 cm (3198x
shvoy-001-004-003	Shvoy Kálmán	Kirchberg am Wechsel A. Ausztria	fénykép	1930	p. 4.		13,4x8,6 cm (3231x
shvoy-001-004-004	Shvoy Kálmán	Kirchberg am Wechsel A. Austria	fénykép	1930	p. 4.		13,6x8,7 cm (3260x
shvoy-001-005-000	Shvoy Kálmán	1. album, 5. oldal	oldalkép	1930	p. 5.		35x26 cm (3012x22
shvoy-001-005-001	Shvoy Kálmán	Kirchberg am Wechsel	fénykép	1930	p. 5.		13,5x8,6 cm (2890x
shvoy-001-005-002	Shvoy Kálmán	Kirchberg am Wechsel : Kilátás a Kernstockwartéról a	fénykép	1930	p. 5.		13,5x8,6 cm (1853x
shvoy-001-005-003	Shvoy Kálmán	Kirchberg am Wechsel : St. Wolfgang templom	fénykép	1930	p. 5.		8,7x13,5 cm (1470x
shvoy-001-005-004	Shvoy Kálmán	Kirchberg am Wechsel	fénykép	1930	p. 5.		13,8x8,7 cm (3168x

6. ábra - Az Omeka rendszerbe szánt Dublin Core adatelemek táblázata

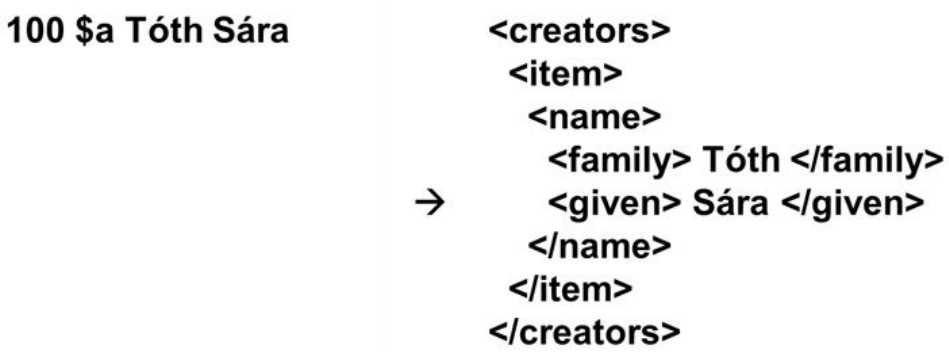
4. A MARC formátum és az EPrints lehetséges kapcsolódási pontjai

A támogatását vesztett Bodza keretrendszer kiváltása során szembekerültünk azzal a problémával, hogy nagy tömegű (összességében százezres nagyságrendű rekordszámról van szó) MARC rekordot kellett EP3 XML formátumra konvertálni. Szintén a Bodza alól való kiköltözés igényét erősítette a Java Applet támogatásának kivezetése a böngésző programokból, mivel így megszűnt a Bodza MARC szerkesztői felülete. Néhány érintett adattárunk: SZTE Egyetemi Kiadványok, SZTE Miscellanea, SZTE UnivHistória, DélmagyArchív, Földrajzi Közlemények, Magyar Nyelvű Filozófiai Irodalom. A Bodzában tárolt nagy számú MARC21 alapú rekord konverziója EP3 XML formátumra egy saját fejlesztésű Java alkalmazás segítségével történhetett meg.

7 SZTE Képtár és Médiatéka. https://mediateka.ek.szte.hu

8 Sirhán Bálint. Repozitóriumépítés: válasszuk az Omeka open source rendszert! Tudományos és Műszaki Tájékoztatás 64. 12. sz. (2017) 619-622.

9 MarcEdit. https://marcedit.reeset.net



7. ábra – Egy kiragadott MARC-EP3 XML konverziós példa részlete

5. A repozitóriumok közös kereshetőségének megteremtése VuFind alapokon

A korábban alkalmazott Bodza-keretrendszer nemcsak metaadatokat és teljes szövegű állományokat szolgáltatott, hanem egyúttal több különálló adatforrás közös keresőjeként is működött, amely funkció a Bodza kivezetésével ellátatlanul maradt, ugyanakkor érzékelhető módon erre a szolgáltatásra folyamatos igény mutatkozik a felhasználók részéről. Ezért a korábban alkalmazott megoldást megkíséreltük kiváltani az EPrints repozitóriumaink VuFind alapú közös kereshetőségének biztosításával. A megfelelő szoftveres háttér kialakításához az ötletet egy 2016-os Networkshop előadás adta1. Informatikusaink különböző megfontolások miatt a MARC szabvány mellett tették le voksukat az adatcsere formátumát illetően, ezért már a kísérletezés elején megvizsgáltuk a GitHub-on megtalálható EPrints MARC export-import eszközt[10]. Sajnos ennek használata körül adódtak nehézségek, mivel egy több mint tíz éves kódról van szó, melyet az akkori EPrints verziókhoz fejlesztettek[11]. Ilyen nehézség volt például, hogy a megfelelő MARC mezőbe a rekord exportálásának időpontja íródott az eredeti létrehozási dátum helyett, illetve a tesztek során az OAI exportot követően az Apache webszerver többször lefagyott. Az eszköz szerencsére parancssorból is működőképesnek bizonyult, ami végül eredményre vezetett. Ütemezett feladatok segítségével így is biztosítható a VuFind alapú közös kereső friss metaadatokkal való folyamatos ellátása.

10 EPrintsMARC. https://github.com/eprintsug/EPrintsMARC

11 Neugebauer, Tomasz, and Bin Han. Batch Ingesting into EPrints Digital Repository Software. Information Technology and Libraries 31. 1. sz. (2012) 113-125.

```
<collection xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns="http://www.loc.gov/MARC21/slim" xsi:schemaL
ocation="http://www.loc.gov/MARC21/slim
http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd">
<record>
 <leader>01329nam a2200265 i 4500</leader>
  <controlfield tag="001">1567</controlfield>
  <controlfield tag="005">20180522121405.0</controlfield>
  <controlfield tag="008">
    120725s2008 hu om 0|| hun d
  </controlfield>
  <datafield tag="040" ind1=" " ind2=" ">
    <subfield code="a">SZTE Doktori Repozitórium</subfield>
    <subfield code="b">hun</subfield>
  </datafield>
  <datafield tag="100" ind1="1" ind2=" ">
    <subfield code="a">Tóth Sára</subfield>
  </datafield>
```

8. ábra – Az SZTE Doktori Repozitórium egyik rekordjából konvertált MARC XML egy részlete

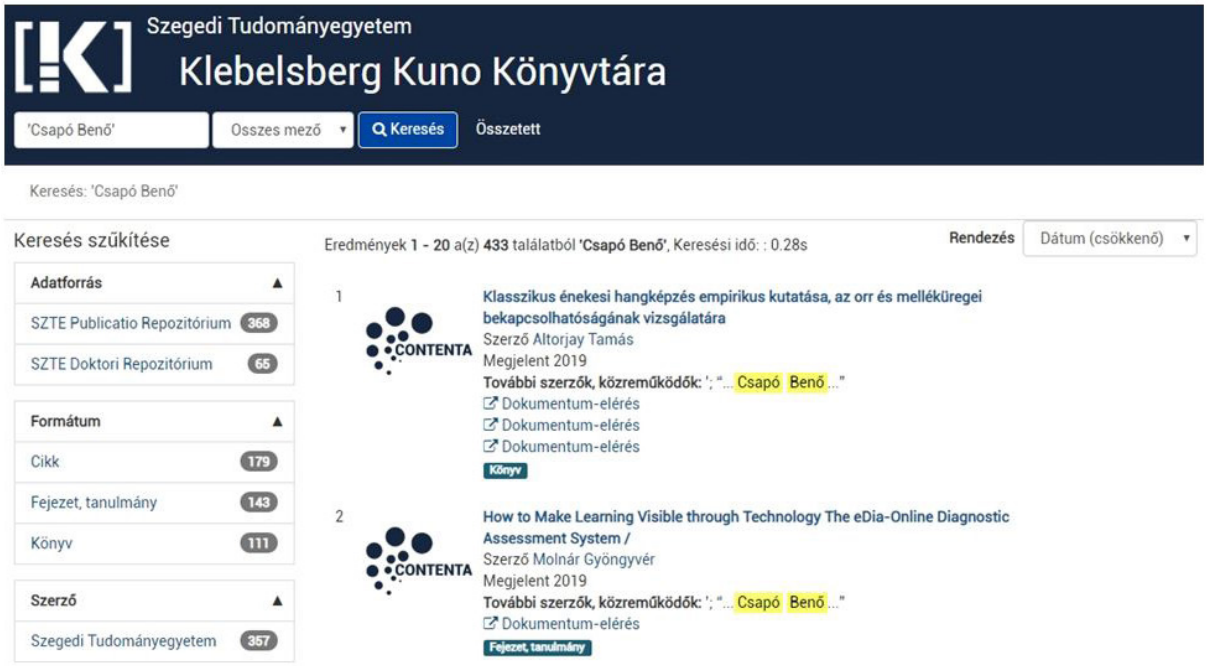
Mivel célunk többféle adatforrásból egy közösen kereshető adathalmaz létrehozása volt, amely célkitűzés szinte magában hordozza a duplumok problémáját, ezért foglalkozni kellett az esetleges duplumrekordok kérdéskörének kezelésével is. Ezzel kapcsolatban szintén a GitHub-on találtunk megoldást a RecordManager nevezetű fejlesztés formájában[12]. Az algoritmus leírása és az ellenőrzési szempontok részletesen megtalálhatóak a hivatkozott oldalon.

A VuFind alapú kereső létrehozásával a metaadatok közös kereshetőségének biztosítása mellett természetesen a teljes szövegű indexelést is szerettük volna megoldani. Ezt különböző kiegészítő szoftverkomponensek segítségével lehet megvalósítani, az egyik ilyen például az Apache-projekt részét képező Tika elnevezésű megoldás[13]. A teljes szövegű kereséshez a Tika megfelelő telepítése és paraméterezése mellett a VuFind konfigurációs fájljaiban is el kellett végezni néhány beállítást.

12 RecordManager Deduplication. https://github.com/NatLibFi/RecordManager/wiki/Deduplication

13 Apache Tika - a content analysis toolkit. http://tika.apache.org

A Contentas névre hallgató közös kereső jelenleg tesztüzemben működik, két repozitóriumunk (SZTE Doktori Repozitórium és SZTE Publicatio Repozitórium) teljes állománya található meg benne. A rendszer a http://contentas.bibl.u-szeged.hu URL címen keresztül nyilvánosan kipróbálható. Terveink szerint hamarosan a Contenta rendszer (amely jelenleg 12 független adattárból áll) tartalmának teljes egésze be fog kerülni.



9. ábra – Találati lista a VuFind alapú Contentas közös keresőben