Műhely 

Open access és osztályozás

BARTON Dávid – PÁL Vanda

Az open access mozgalom keretein belül közzétett, különböző digitális gyűjteményekben tárolt publikációk visszakereshetővé tételéhez a dokumentumok formai és tartalmi feltárása mellett az is fontos, hogy a felhasználóknak lehetőségük legyen egy közös felületen több digitális gyűjteményben keresni, amit a repozitóriumok által alkalmazott különböző technológiai megoldások mellett az eltérő nyelvek és terminológia is megnehezítenek. A jelen tanulmány azon megoldások rendszerbe foglalására törekszik, amelyek a nyílt hozzáférésű publikációk tartalmi feltárásának és visszakereshetővé tételének terén jelenleg használatosak. Ezek a megoldások az elmúlt két évtized különböző webes technológiáit használják, mint a Dublin Core-t és a kiterjeszthető jelölőnyelvet (XML).

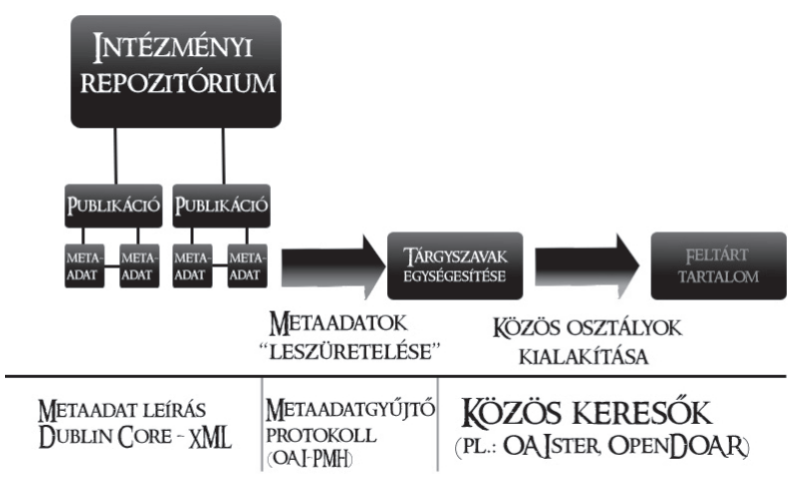
Az információszabadság létrejöttét támogató open, azaz nyílt kezdeményezések az elmúlt évtizedben jutottak el arra a szintre, hogy a tudományos kommunikációban komolyabb szerephez juthattak. Ezek legjelentősebbjei az open source (nyílt forráskód), az open standards (nyílt szabványok) és az open access, vagyis a nyílt hozzáférés. Utóbbinak köszönhetően a világon bárhol egyformán és egyenlő esélyekkel férnek hozzá a tudományos publikációkhoz az oktatásban és kutatásban résztvevő egyének, intézmények. Ezeket az open access (OA) dokumentumokat különálló digitális könyvtárak, intézményi repozitóriumok őrzik, amelyeknek formai és tartalmi feltárása épp oly fontos, mint hagyományos társaiké.

Az open access publikációk ingyenes, teljes szövegű online elérést, letöltést és (saját használatra való) nyomtatást tesznek lehetővé bárki számára. A nyílt hozzáférésű publikálás két lehetséges változatát, a *szerzői önarchiválás*t, illetve az *open access folyóiratokban való publikálás*t 2002-ben határozta meg a Budapest Nyílt Hozzáférés Kezdeményezés.[1] A szerzői önarchiválás elsősorban az intézményi repozitóriumba való feltöltésre vonatkozik, amely a Nyílt Archívum Kezdeményezés[2] szabványai mentén lehetővé teszi, hogy mindenki számára elérhetőek legyenek a publikációk, illetve hogy a fenntartó intézmény (egyetem, kutatóintézet, könyvtár) keretein belül létrehozott szellemi termékek egy közös tárhelyen legyenek elérhetők. A másik lehetséges út a nyílt hozzáférésű „alternatív” folyóiratokban való publikálás, amelyek az online megjelenési formát választották a nyomtatott formával szemben, így ugyanis a nyomdai és raktári költség is megszűnik.[3] A publikációk ingyenessé tétele a szerzők számára sem haszontalan – mivel szélesebb körben érhetőek el írásaik, több találati listában szerepelnek előkelő helyen, ezáltal jobb eséllyel fogják őket idézni is.[4]

Jelenleg nincsen egységes megoldás a nyílt hozzáférésű publikációk tartalmának feltárására. Problémát jelent a publikációs tárak közötti technológiai és terminológiai különbség, pedig lenne igény a relevánsabb open access találati listákra.

Az open access repozitóriumok esetében – mivel a kezdeményezés célja is az, hogy minél szélesebb körben elérhetővé váljanak a publikációk – alapfeltétel, hogy szemantikai szempontból megfelelően legyenek leírva a benne foglalt publikációk, lehetővé téve ezzel a közös keresők és az egyéb, felszíni webet indexelő keresők (például Bing, Google, Google Scholar stb.) számára, hogy releváns találatokat adjanak. A feltárásra két alapvető út létezik: a kereskedelmi elvek szerint működő keresőmotorok (például Bing, Google) tevékenysége erősen elkülönül a dokumentumokat a metaadatok útján, tartalmilag feltáró eljárásoktól. A Google-féle keresőmotorok „beleláthatnak” ugyan a repozitóriumok tartalmába és a publikációk teljes szövegébe, de a PageRank[5] keresési és hiperlinkelési népszerűségen alapuló algoritmusai kevésbé tematizálják a dokumentumokat tartalmuk szerint, indexelési módszerük mindössze egy szintaktikai aspektusa a dokumentum feltárásnak. A repozitóriumi rekordok, publikációk metaadatokkal való ellátása nem felesleges annak ellenére, hogy a keresőmotorok által már sikerül megmenteni őket attól, hogy a mély webre kerüljenek. A metaadatok biztosítják, hogy precízen, pontosan és mindenekelőtt szemantikailag legyenek feltárva a webes tartalmak, ami tudományos szempontból fontosabb, mint az, hogy a PageRank és Google hirdetések üzleti központú rangsorolása hová is helyezi azokat a találati listákban. (A tudományos írások feltárására fejlesztett Google Scholar saját idézettségi indexek létrehozásával már annál inkább törekszik az ettől való elrugaszkodásra.) Az internet feltártságát a tartalmak metaadatokkal való ellátása teszi a hagyományos, katalógusok által feltárt könyvtárakéhoz hasonlatossá. Kutatásunk során ez utóbbi megoldással és ennek eszközeivel foglalkoztunk, ugyanis a nyílt hozzáférésű tudományos publikációk esetében rendkívül fontos a szemantika, és a repozitóriumokban tárolt tartalmak esetében is a végső cél a szemantikus web létrehozása.

Amíg a szemantikus web, a Web 3.0 várat magára, közös keresők útján lehetséges a nyílt hozzáférésű dokumentumok visszakereshetővé tétele. A publikációk útja a szerzőtől az olvasóig az osztályozási funkciók szempontjából három szakaszra bontható *(1. ábra)*.



1. ábra
A publikációk útja az open access rendszerében és az egyes szinteken fellépő osztályozási funkciók

Elsőként a repozitóriumokba feltöltött dokumentumokat a feltöltő metaadatokkal látja el, amelynek során annak tartalmi ismérveit is leírja. Ezután a közös keresők részéről egy protokoll, az OAI-PMH lekéri, összegyűjti („leszüreteli") a metaadatokat, ami eszközt biztosít a tárgyszóegységesítéshez. Ennek segítségével végül a közös kereső osztályokat alakíthat ki, így a különböző terminológiával rendelkező, eltérő nyelvű repozitóriumok közös tudományfelosztással és tárgyszókészlettel jelenhetnek meg a keresők felületén, ami hozzájárul a sokkal relevánsabb találatokhoz a keresések során.

Metaadat-leírás

A metaadatok rögzítésének legelterjedtebb formátuma a Dublin Core elemkészlet.[6] A Dublin Core adatai közül a dokumentumok tartalmát hivatottak leírni a *tárgy*, a *leírás* és az *elhelyezés és időbeliség*.

A tárgy (subject) mezőben a dokumentum tartalmát kifejező fogalmakat, kifejezéseket, illetve különböző osztályozási rendszerek jelzeteit lehet leírni, nagyobb témakörök, illetve speciálisabb tárgy- és kulcsszavak megadására egyaránt használható.

A tárgyszavak forrása ideális esetben valamilyen rendszerezett tárgyszójegyzék vagy tezaurusz. A 2012-ben már 140 000 lexikai egységet tartalmazó Köztaurusz egyaránt tartalmaz szakkifejezéseket, földrajzi- és időneveket, valamint a dokumentumtípust meghatározó deszkriptorokat, de emellett is számos szaktezaurusz segítheti még a dokumentumok tartalmának pontos feltárását, mint például filmtechnikai, jogi és közigazgatási, távközlési, irattári, EU- és ETO-tezaurusz is. Ezek mellett léteznek elméleti, tudományos célú tezauruszok is, mint az Általános fogalmak ontológiája, az Időnévtér vagy a Szófajtezaurusz.[7]

A dokumentumok angol nyelven végzett feltárásához ajánlott adatbázisok megtalálhatók a Dublin Core honlapján.[8] Ilyen a Library of Congress Subject Headings (LCSH), a Kongresszusi Könyvtár átfogó, egyetemes tárgyszójegyzéke, amelyet az Egyesült Államok könyvtárai mellett nemzetközileg is számos könyvtár használ, gyakran fordítási célból. A Kongresszusi Könyvtár külön adatbázist szolgáltat a témaként használt tulajdonnevek egységesített alakjairól, külön kereshető a Kongresszusi Könyvtár saját osztályozási rendszere (Library of Congress Classification – LCC) és számos egyéb adatbázis.[9]

A Dublin Core-ban található subject mező kitölthető osztályozási rendszerek jelzeteivel is, ezt segíti például a Dewey Tizedes Osztályozásának (Dewey Decimal Classification – DDC),[10] illetve az Egyetemes Tizedes Osztályozásnak (Universal Decimal Classification – UDC)[11] online elérhető változata, amelyek az egyes fogalmakhoz rendelt jelzeteket teszik kereshetővé. A dublincore.org oldalon emellett egyes szakterületekhez tartozó tárgyszójegyzékeket is ajánlanak, mint az orvostudomány területén használatos Medical Subject Headings (MeSH), vagy a művészeti és építészeti szakterületet feldolgozó Art and Architechture Thesaurus (AAT).

Az ebben a mezőben tetszőleges számban megadható tárgyszavak és osztályozási jelzetek lehetővé teszik a tartalom messzemenőkig pontos, konkrét feltárását, illetve – megfelelő visszakereső felület létrehozása esetén – a tartalmilag összetartozó dokumentumokból kisebb csoportok képzését.

A Dublin Core tartalmi feltárást segítő következő mezője, a leírás (description) a dokumentum tartalmának összegző, szabad szöveges leírását teszi lehetővé. Bemásolható ide a dokumentum tartalomjegyzéke és tartalmi kivonata, akár több nyelven is.

Az elhelyezés és időbeliség (coverage) mezőben a dokumentum tartalmának tér- és időbeli vonatkozását lehet feltárni. Itt lehet megadni az olyan földrajzi és közigazgatási egységeket, időpontokat és korszakokat, amelyek a dokumentum témájához köthetők.

A földrajzi vonatkozást szabadszövegesen is lehetőségünk van leírni, de szabványosított formát is használhatunk, és pontos koordináta értékeket is megadhatunk. Szabványosított földrajzi nevek találhatók a Köztauruszt kiegészítő Geotaurusz és geohistauruszban, amely egy földrajzi és történelemföldrajzi témájú átfogó osztályozási rendszer. Feltünteti a fogalmak közti fontosabb relációkat, amelyek tükrözik például a különböző földrajzi egységek közti történeti összefüggéseket, a közigazgatási beosztást és – lehetőség szerint – annak változásait is. A tezauruszban található deszkriptorokhoz megjegyzések is kapcsolódnak, amelyek kézikönyvekből származó általánosan elfogadott tényeket, meghatározásokat tartalmaznak.[12]

Angol nyelven történő feltárás esetén használni lehet például a Thesaurus of Geographic Namest (TGN),[13] illetve az ISO 3166 szabványát,[14] amely az országok neveinek két és hárombetűs kódjait tartalmazza. Az idővonatkozás megadásához támpontot ad a World Wide Web Consortium Date and Time Formats (W3C DTF) segédlete.[15]

Emellett a Dublin Core más mezői is segítik a dokumentum tartalmi ismérveinek behatárolását – a dokumentum címe, típusa, forrása és a más dokumentumokkal fennálló kapcsolata is hordoz a dokumentum tartalmára vonatkozó információkat.

Az objektumtípus (object type) mezőben az adott dokumentum típusát, műfaját lehet leírni. Magyar nyelven használható ehhez az *Ungváry Rudolf* által összeállított *Dokumentumfajták tezaurusza*, amely a Köztaurusz részeként is elérhető, illetve önállóan a MOKKA Relex által kezelt tezauruszainak létrehozott felületéről is. A tezaurusz jelenleg majdnem 3700 lexikai egységet tartalmaz, amelyekhez meghatározás is kapcsolódik.[16] Angol nyelven a Dublin Core Metaadat Kezdeményezés ajánlását lehet használni.[17]

A forrás (source) mezőben lehetőség van a világhálóról valahonnan átvett anyagok, illetve a digitalizált dokumentumok eredeti forrásának adatait megadni. Ez is segítheti a felhasználót az információk keresése közben – ha például egy dokumentum eredeti, nyomtatott forrását már ismeri és feldolgozta, akkor az online elérhető dokumentum átolvasása nélkül is tudomást szerezhet arról, hogy az nem fog számára új információkat tartalmazni. Itt meg lehet adni egy dokumentum kiadási adatait, illetve szabványos azonosítókat (ISSN, ISBN, URL) is.

A kapcsolat (relation) mező alkalmas arra, hogy az adott dokumentummal valamiféle tartalmi vagy logikai kapcsolatban álló bármilyen másik dokumentumra hivatkozzunk. A dokumentumok között sokféle kapcsolat állhat fenn: előzmény vagy folytatás, egy sorozatba való tartozás, gyűjteményes mű különböző részei vagy egyszerűen tematikai vagy azonos szerző szerinti kapcsolat. Ennek segítségével könnyebb eljutni az adott dokumentummal tartalmilag összetartozó más dokumentumokhoz. A relation mezőben a source-hoz hasonlóan meg lehet adni egy dokumentum kiadási adatait, és szabványos azonosítókat (ISSN, ISBN, URL).[18]

Magyarországi repozitóriumok gyakorlata

A szerzői önarchiválás eljárásmódja intézményenként eltér. Bizonyos intézményeknél könyvtáros szakember kapcsolja a metaadatokat a dokumentumokhoz, bizonyos intézményeknél a szakember csak ellenőrzi a szerző által feltöltött metaadatokat, egyes repozitóriumoknál pedig még csak nem is ellenőrzik a szerzők által feltöltött metaadatokat (*1. táblázat*). Ennek eredménye, hogy a feltárás mélysége és minősége rendkívül ingadozó képet mutat a nyílt hozzáférésű dokumentumok körében, amely probléma kezelhető volna a publikációk feldolgozására, megőrzésére és hozzáférhetővé tételére vonatkozó eljárásoknak a különböző repozitóriumok körében való egységesítésével. Annak meghatározása például, hogy hány darab tárgyszót kell egy dokumentumhoz minimum megadni, hogy a tárgyszavak mely tezauruszokból származhatnak vagy hogy kell-e kötelezően valamelyik osztályozási rendszer jelzetét a publikációhoz kapcsolni, jelentősen csökkentené a különböző terminológiákból és a feltártság eltérő mélységéből fakadó, a dokumentumok visszakeresését nehezítő problémákat.

A szabad szöveges kereső mellett bizonyos repozitóriumok hierarchikus tudományfelosztáson alapuló böngészést tesznek lehetővé a szakterületek között. Ezek többnyire már meglévő felosztásokon alapulnak, mint a Kongresszusi Könyvtáré (az MTA REAL és REAL-d adatbázisa használja)[19] vagy a Magyar Elektronikus Könyvtáré (a MIDRA esetében).[20] A Szegedi Tudományegyetem Contenta repozitóriuma ezzel szemben egy egyéni, többszintű taxonómiát használ a doktori publikációk osztályozására.[21]

A taxonómiának megvan az az előnye, hogy az alá-, fölé- és mellérendeltségi viszonyokat is jól áttekinthetővé tévő „ágrajz" útján egyből megmutatja, hogy az egyes szakterületek milyen mélységig kapnak terepet az adott adatbázisban. Ez a fajta témakörök közti böngészési lehetőség azonban a felhasználó számára nehézségeket is rejthet magában, hiszen nem feltétlenül magától értetődő mindenki számára az adott felületen található tudományfelosztás. Egyes repozitóriumok a hierarchikus felosztás helyett a dokumentumaik tartalma közti böngészést a gyűjtemény tárgyszólistájának egyszerű betűrendes közlésével teszi lehetővé. A magyar nyelvű archívumok közül ilyen például a DEA.[22]

A Magyarországon legelterjedtebb repozitórium szoftverek, az EPrints és a DSpace nyílt forráskód alatt futnak, ami biztosítja a hosszú távú fejlesztés lehetőségét.

Repozitórium	Repozitórium szoftvere	Önarchiválási gyakorlat
Corvinus Egyetem Kutatások http://unipub.lib.uni-corvinus.hu/	EPrints	Önálló (szerzői) tárgyszavazás.
Corvinus Egyetem Doktori disszertáció http://phd.lib.uni-corvinus.hu	EPrints	Önálló (szerzői) tárgyszavazás.
Szent István Egyetem – HuVetA http://huveta.hu/	DSpace	–
Central European University http://ceu.archives.ceu.hu/	EPrints	Önálló (szerzői) tárgyszavazás.
Miskolci Egyetem – MIDRA http://midra.uni-miskolc.hu/	JaDoX	Önálló (szerzői) tárgyszavazás.
MTA – REAL http://real.mtak.hu/	EPrints	A szerzők önállóan töltik fel munkájukat szabadkulcsszavak és a Kongresszusi Könyvtár tudományfelosztása alapján. Mielőtt a dokumentum publikussá válna az MTAK munkatársai ellenőrzik.
MTA – REAL-d http://real-d.mtak.hu/	EPrints	Az MTAK végzi a Kongresszusi Könyvtár tudományfelosztása alapján.
Debreceni Egyetem Egyetemi és Nemzeti Könyvtára – DEA http://ganymedes.lib.unideb.hu:8080/dea/	DSpace	Önálló (szerzői) tárgyszavazás.
Szegedi Tudományegyetem repozitóriumai – Contenta http://contenta.bibl.u-szeged.hu/	EPrints	Központi vagy önálló feltöltés/tárgyszavazás.

1. táblázat

Magyarországi repozitóriumok által használt szoftverek és archiválási gyakorlatuk. (Forrás: OpenDOAR.org/ és az egyes repozotóriumok honlapjai.)

Metaadatgyűjtő protokoll

A közös felületen való visszakeresés biztosítását szolgálja a Nyílt Archívum Kezdeményezés adatgyűjtő protokollja (OAI-PMH – Open Archives Initiative Protocol for Metadata Harvesting), amely a nyílt adattárak metaadatait gyűjti be és biztosítja a repozitóriumok és a keresőmotorok közti kommunikációt. Az OAI-PMH két oldalon jelentkezik, az egyik a *data provider*, a (meta) adatszolgáltató, a másik a *service provider*, a szolgáltatási pont, vagyis metaadatgyűjtő (például OAIster, Videotorium).[23] A szolgáltatási pont maga is adatszolgáltatóvá válhat, ha az általa begyűjtött adatokat továbbítja egy másik szolgáltatási pontnak – így a szolgáltatási pontok között egy több szintű, hierarchikus rendszer is létrejöhet. Az egyes repozitóriumok belső struktúrája részhalmazokra osztható a 'set' kezdetű utasítások által, ami osztályozási szempontból fontos: akár témakörök szerint is. Az adatszolgáltatók tételeket tárolnak, amelyekhez a metaadatok kapcsolódnak.[24] A protokollra vonatkozó irányelvek a *DRIVER*-ben lettek közzétéve. Itt konkrét könyvtári osztályozással kapcsolatos iránymutatást is kapunk, pl. URI sémákkal, vagy a Dewey Decimal Classification jelzeteivel való leírást a Dublin Core Subject elemében.[25]

Az OAI-PMH-t alkalmazó repozitóriumok halmazképzéssel oldhatják meg az osztályozási taxonómiák kialakítását. A repozitóriumok tételeiket tetszőlegesen rendezhetik halmazokba, az XML nyelv segítségével poszt-koordináltan szabhatók meg a struktúrák. A halmazstruktúra lehet egyszintű vagy hierarchikus, tehát alá- és mellérendelő tárgyszóláncok alakíthatóak ki segítségükkel. Két osztályozási, jelzetalkotási szempontból leglényegesebb elemük a setSpec és a setName:

– *setSpec* – megadja egy halmaz elérési útvonalát a hierarchia gyökeréhez képest. Fontos, hogy egyedi azonosító legyen. A hierarchikus struktúra elemeit kettőspont választja el egymástól.
– *setName* – az előzőekben meghatározott elemek megjelenítendő elnevezései természetes nyelven.[26]
  Például az ELTE, mint intézmény a következőképp fejezhető ki. (1) *SetName: Intézmények;* (2) *SetName: Eötvös Loránd Tudományegyetem*, valamint (1) *SetSpec: intezmeny;* (2) *SetSpec: intezmeny:elte.*[27]

Egy közös kereső létrehozása ennek a technológiának az alkalmazásával, amely az összes magyarországi nyílt hozzáférésű repozitóriumból begyűjtené a metaadatokat, a jelenleginél sokkal egyszerűbb visszakereshetőséget biztosítana a felhasználóknak. A következőkben olyan közös keresőknél alkalmazott megoldásokról számolunk be, amelyeknél sikerrel hasznosították könyvtári osztályozási feladatokra a protokollt.

Közös keresők

Az OAI-PMH segítségével működik az Open DOAR (Directory of Open Acces Repositories – Nyílt hozzáférésű repozitóriumok mutatója)[28] is, amelynek példáján keresztül megvizsgáltuk, hogyan hasznosítható a nyílt kezdeményezés XML-alapú protokollja osztályozási feladatokra. A repozitóriumkereső felület lenyíló listájának tudományfelosztása már önmagában hordozza az OpenDOAR osztályozását.

A következő példában az OpenDOAR XML-alapú, OAI-PMH-t kiszolgáló, „osztályozási jelzeteket" leíró kódjának részletét ismertettük. A kódon belül a *setSpec* definiálja az osztályozási jelzeteket, a *setName* pedig a listákban ténylegesen megjelenő neveket természetes nyelven.

1. példa:

```
<psh>
<responseDate>2012-10-24T17:54:
23Z</responseDate>
<request verb="Count">http://w
ww.opendoar.org/demos/psh.php</
request>
  <Count>
  <header>
  <setType>subject</setType> 15
  <setSpec>Ci</setSpec>
```

```
  <setName>Technology General</
  setName>
  <datestamp/>
  <numItems>126</numItems>
  </header>
  <header>
  <setType>subject</setType>
  <setSpec>Cif</setSpec>
  <setName>Architecture</setName>
  <datestamp/>
  <numItems>32</numItems>
(...)
  <setSpec>Cil</setSpec>
  <setName>Civil Engineering</
  setName>
(...)
  <setSpec>Cin</setSpec>
  <setName>Computers and IT</
  setName>
(...)
```
[29]

Az alosztályok a 'Ci' (Technology General főosztály) jelzetét bővítették minden esetben egy-egy további betűvel, amely egyedileg azonosítja a halmazukat (Cif, Cil, Cin). Bár logikailag az 'Általános technológia' egy szinttel feljebb áll a másik három fogalomnál, mindazonáltal egymás mellett helyezkednek el *(2. ábra)* mellérendelő tárgyszóláncot alkotva.

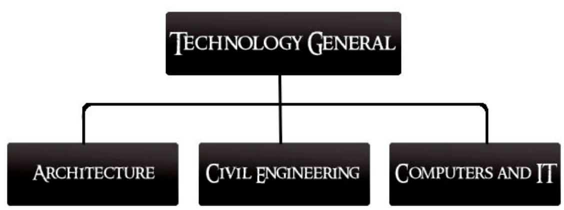
Technology General; Architecture; Civil Engineering; Computers and IT

2. ábra
Az OpenDOAR 'Általános technológia' főosztálya és három alosztályának egymáshoz viszonyított helyzete mellérendelő tárgyszólánc esetén

2. példa: Ugyanezek az adatok alárendelő tárgyszólánccal, kétszintű hierarchiában, ahol a fogalmi szintek kettőspontok segítségével vannak kifejezve (fiktív példa az OpenDOAR előző tárgyköreivel).

```
<request verb="Count">http://
www.opendoar.org/demos/psh.php</
request>
  <Count>
  <header>
  <setType>subject</setType>
  <setSpec>Ci</setSpec>
  <setName>Technology General</
  setName>
  <datestamp/>
  <numItems>126</numItems>
  </header>
  <header>
  <setType>subject</setType>
  <setSpec>Ci:arch</setSpec>
  <setName>Architecture</setName>
  <datestamp/>
  <numItems>32</numItems>
(...)
  <setSpec>Ci:civil_eng</setSpec>
  <setName>Civil Engineering</
setName>
(...)
  <setSpec>Ci:comp_IT</setSpec>
  <setName>Computers and IT</
setName>
(...)
```

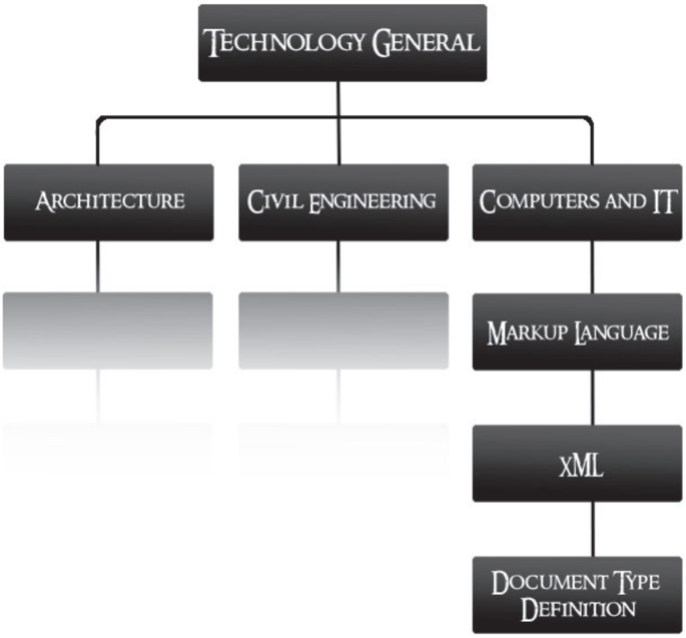
Az OpenDOAR saját jelzeteivel ellentétben (ld. előző példa) itt következetesen – a Cif, Cil és Cin helyett – az angol elnevezések rövid változatát használtuk a második szinteken, míg a gyökérben meghagytuk a 'Technology genereal'-t 'Ci' jelzetnek. Az osztályok egymáshoz viszonyított helyzete a következőképp néz ki ágrajzon *(3. ábra)*.



3. ábra
Az előbbi elemek alárendelő tárgyszólánc esetén

3. példa: Komolyabb taxonómiák is létrehozhatóak a kettőspontok halmozásával *(4. ábra)* így egész tárgyszókészletek leírhatóak a protokoll által. Például a Document Type Definition fogalmát így lehetne meghatározni, ha a hierarchia ötödik szintjén helyezkedne el (a jelzet fölé írt sorszámok nem képezik részét a kódnak, csupán az átláthatóságot segítik)

```
(...)      1.       2.   3.     4.    5. szint
<setSpec>Ci:comp_IT:Markup_Lang:XML:Document_Type_Def.</setSpec>
<setName>Data Type Definition</setName>
                                                   (...)
```



4. ábra
Ötszintű taxonómia a tárgyszókészlet szintjén

A WorldCat által üzemeltetett OAIster[30] a nyílt hozzáférésű publikációk legjelentősebb keresője, amely azt tűzte ki céljául, hogy ezek a dokumentumok ne a mély web részét képezzék. Az OAIster a keresés során a különböző repozitóriumok Dublin Core adatait a DLXS (Digital Library Extension Service) Bibliographic Class metaadat-formátumává fordítja, ezzel megalkotva a közös nevezőt: a tárgyszavak egységesítésével ritkábban adódnak hibák OAI-PMH harvestelés közben.[31]

Record 11 of 209

add to bookbag

Title	Three Frontiers in **Open Access** Scholarship
Author/Creator	Getz, Malcolm
Publisher	Internet-First University Press
Year	2005-01-13T20:21:35Z
Year	2005-01-13
Resource Type	Paper or Project
Resource Format	153534 bytes, application/pdf
Language	English
Note	There are three important frontiers in moving from subscription-based scholarly publications to delivery of scholarly works to readers without charge via the Internet. First are automated archives of preprints and post prints that do not require formal editorial review before posting. The arXiv service, now at Cornell, is emblematic of this frontier. (arXiv, 2004) Second are the quality-assured journals that are distributed on an open-access basis. The Public Library of Science initiative in launching journals in biology and medicine is emblematic of the second frontier. (Public Library of Science, 2004) Third are **open access** indices to the scholarly literature. Google Scholar, launched in beta version in November 2004, is emblematic of this frontier.(Google, 2004) Each frontier advances the prospect that the best scholarship will be readily available to all via the Internet. Our goal here is to identify where each frontier is today and how it may evolve.
Note	Vanderbilt University
1. Subject	**Open Access**; Scholary Publishing; Open Archives; Open Index
2. Subject	digital libraries; publishing; science; web programming
3. Subject	Information and Library Science; Business (General); Computer Science; Science (General)
4. Subject	Science; Engineering; Social Sciences; Business & Economics
URL	http://hdl.handle.net/1813/307
Data Contributor	DSpace at Cornell University
Score	141

5. ábra
A DLF Portál többszörös osztályozása

Egy másik innovatív megoldást használt az OAIstert fejlesztő közösség az open access publikációk osztályozására. A Digitális Könyvtári Szövetség (Digital Library Federation – DLF) portáljának[32] keresője volt a kísérleti projekt, ami az OAIster számára tervezett facettázó-klaszterképző eljárást használta, amely a dokumentumokat a High Level Browse (HLB) osztályozás terminológiájával címkézi fel a Dublin Core metaadatok alapján. A HLB a Kongresszusi Könyvtár osztályozási rendszerén (LOC Classification), a kereső klaszterképzési eljárása pedig automatizáláson alapul. A Topic Model elnevezésű eljárás segítségével az algoritmus szemantikusan kinyeri a releváns kulcsszavakat, és automatikusan klasztereket képez. A fölösleges, redundáns angol nyelvű szavak (and, the, with stb.) kihagyásáról is az algoritmus gondoskodik, amely közel 100 000 szóból álló szótárat készített, és 7,5 millió rekordot dolgozott fel az OAIsterrel kapcsolatban álló nyílt repozitóriumokból. A DLF portálon a találatok megjelenítése során úgynevezett kettős vagy többszörös osztályozást alkalmaznak. A találatoknál több 'Subject' mező is található *(5. ábra),* ahol az első az eredeti, önarchiválásnál létrehozott Dublin Core metaadatokat, a második és harmadik a Topic Model klaszter címkéket (a második tárgyszó, a harmadik tudományterületek szintjén), a negyedik pedig a High Level Browse terminológiát tartalmazza. Az OAIster a mai napig nem tért át erre a facettázó eljárásra a kísérleti projekt közben felmerült hibák és a visszakeresések pontatlansága miatt.[33] Tehát a közös kereső megmaradt Dublin Core/OAI-PMH-alapúnak, mindazonáltal a klaszterezést kihagyva is rendkívül hatékonyan keres vissza a repozitóriumok tartalmában.

Összegzés

Összességében azt láthatjuk, hogy a nyílt hozzáférésű dokumentumok esetében a technológiai lehetőségek már adottak a tartalom messzemenőkig pontos feltárásához és visszakereshetővé tételéhez, és a könyvtárosok szerepe továbbra is fontos marad a repozitóriumok térhódításával, hiszen a nyílt hozzáférésű publikációk feltárásához szükség lesz a megfelelő szakértelemre a tudományfelosztások, jelzetelések terén, valamint a minél relevánsabb metaadatok létrehozásában. Azonban ahhoz, hogy ezeknek a dokumentumoknak az osztályozása professzionálisan, egy könyvtár fizikailag létező állományával azonos szinten valósuljon meg, és a repozitóriumok közti átjárhatóság, interoperabilitás is megteremtődjön, szükséges volna a repozitóriumok eljárásainak valamiféle – irányelvekkel, ajánlásokkal való – egységesítésére. Ez megteremthetné az alapját egy a magyarországi repozitóriumokat magába foglaló közös kereső létrehozásának, amely által még könnyebbé válna a nyílt tudományos információkhoz való hozzáférés. Hasonló folyamatok indulnak el napjainkban az Unióban: a Horizon 2020[34] program célkitűzései között szerepel például a nemzeti open access rendeletek megalkotása, a LIBER ajánlásai[35] között pedig a könyvtárosok alkalmazásának és az átjárható infrastruktúra kialakításának szükségessége az adatok tárolása, elérése és megosztása céljából.

Irodalom

1. *Budapest Open Access Initiative.* Open Society Institute. Soros Foundation. (online) 14. February 2002. http://www.soros.org/openaccess/read.shtml [letöltve: 2012. október 12.]
2. http://www.openarchives.org/ [letöltve: 2012. október 12.]
3. BAILEY, Charles W.: What is Open Access? In: JACOBS, Neil (Ed.): *Open Access: Key strategic, technical and economic aspects.* Chandos Publishing, Oxford, 2006. 14. p.
4. Uő. uo. p. 14-15.
5. PAGE, Lawrence – et al.: *The PageRank Citation Ranking: Bringing Order to the Web.* January 29, 1998, p. 1-2. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf [letöltve: 2012. október 29.]
6. http://dublincore.org/ [letöltve: 2012. október 14.]

7. RIMÁR Miklós: *A Relex és tezauruszai.* Könyvtári Intézet, 2012. április 4. http://ki.oszk.hu/relex [letöltve:: 2012. november 1.]

8. *Dublin Core Qualifiers*. Dublin Core Metadata Initiative, July 11. 2000. http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/ [letöltve: 2012. november 1.]

9. *Library of Congress Subject Headings.* http://id.loc.gov/authorities/subjects.html [letöltve: 2012. november 1.]

10. *Subject Headings by Dewey Decimal Classification.* Central Mississippi Regional Library System. http://www.cmrls.lib.ms.us/subject_headings.htm [letöltve: 2012. november 1.]

11. *Universal Decimal Classification.* UDC Consortium. http://www.udcc.org/udcsummary/php/index.php [letöltve: 2012. november 1.]

12. UNGVÁRY Rudolf (szerk.): *Geotaurusz és Geohistaurusz.* Magyar Könyvtárosok Egyesülete – Országos Széchényi Könyvtár, 2001. november 1. http://mek.oszk.hu/00000/00070/html/ [letöltve: 2012. november 1.]

13. *Getty Thesaurus of Geographic Names Online.* The Getty Research Institute. http://www.getty.edu/research/tools/vocabularies/tgn/ [letöltve: 2012. november 1.]

14. *ISO 3166 Codes (Countries).* http://userpage.chemie.fu-berlin.de/diverse/doc/ISO_3166.html [letöltve: 2012. november 1.]

15. *Date and Time Formats.* W3 Consortium 1997. szeptember 15. http://www.w3.org/TR/NOTE-datetime [letöltve: 2012. november 1.]

16. *A Relex tezauruszai.* http://mokka.hu/relex/guest.html [letöltve: 2012. november 1.]

17. *DCMI Metadata Terms.* http://dublincore.org/documents/dcmi-type-vocabulary/ [letöltve: 2012. november 1.]

18. *MEK Dublin Core Generátor Súgó.* http://mek.oszk.hu/dc/sugo.html [letöltve: 2012. október 26.]

19. *MTA REAL-d.* http://real-d.mtak.hu/view/subjects/ [letöltve: 2012. november 2.]

20. *MIDRA.* http://midra.uni-miskolc.hu/jadox/portal/search.psml [letöltve: 2012. október 27.]

21. *SZTE Doktori Repozitórium.* Contenta. *http://doktori.bibl.u-szeged.hu/view/discipline/ [letöltv*e: 2012. november 2.]

22. A DEA gyűjteményének böngészése – Tárgyszó. http://ganymedes.lib.unideb.hu:8080/dea/browse?type=subject&order=ASC&rpp=20&starts_with=0 [letöltve: 2012. november 2.]

23. AHLBORN, Benjamin – NEJDL, Wolfgang – SIBERSKI, Wolf: OAI-P2P: A peer-to-peer network for open archives. = *Parallel Processing Workshops,* 2002. Proceedings. International Conference on. IEEE, 2002. p. 462-463.

24. KISS Gergő – et al.: *HEKTÁR: Hazai elektronikus könyvtári rendszerek összekapcsolása.* 2004. április 20. http://mek.oszk.hu/html/irattar/ajanlas/hektar2004.pdf [letöltve: 2012. november 2.]

25. *DRIVER irányelvek 2.0* : Irányelvek tartalomszolgáltatók számára - Szöveges források közzététele OAI-PMH használatával. Debreceni Egyetem. 2008. november. 104. p. http://www.open-access.hu/sites/www.open-access.hu/files/DRIVER_Guidelines_hun.pdf [letöltve: 2012. november 3.]

26. LAGOZE, Carl – VAN DE SOMPEL, Herbert: Az Open Archives Initiative Metaadatgyűjtési Protokollja. = *HEKTÁR.* MTA SZTAKI, Budapest, 2004. május 11. http://hektar.sztaki.hu/oai/protokoll.html [letöltve: 2012. október 12.]

27. Uő. uo. [Kiss Gergő (ford.) magyar példái alapján]

28. http://www.opendoar.org/ [letöltve: 2012. október 24.]

29. *OpenDOAR OAI-PMH.* http://www.opendoar.org/demos/psh.php?verb=Count&setType=subject&setQuery=Ci&setQueryType=spec&operator=starts [letöltve: 2012. október 25.]

30. http://oaister.worldcat.org/ [letöltve: 2012. október 27.]

31. HAGEDORN, Kat: OAIster: a „no dead ends" OAI service provider. = *Library Hi Tech.* 21 (2). p. 170-181.

32. http://quod.lib.umich.edu/i/imls/ [letöltve: 2012. október 27.]

33. HAGEDORN, Kat – CHAPMAN, Suzanne – NEWMAN, David: Enhancing Search and Browse Using Automated Clustering of Subject Metadata. = *D-Lib Magazine.* Volume 13 Number 7/8, July/August 2007. http://www.dlib.org/dlib/july07/hagedorn/07hagedorn.html [letöltve: 2012. október 27.]

34. Horizon 2020. http://ec.europa.eu/research/horizon2020/index_en.cfm [letöltve: 2013. június 2.]

35. 'LIBER ajánlások' http://www.open-access.hu/sites/www.open-access.hu/files/upload/KutAdatok2013marc13.pdf [letöltve: 2013. június 2.]

Beérkezett: 2013. szeptember 4.