DOI: 10.31915/NWS.2019.11

MTA SZTAKI DSD – 25 éve a digitális könyvtárak szolgálatában

Kovács László, Micsik András
MTA SZTAKI Elosztott Rendszerek Osztály
laszlo.kovacs@sztaki.mta.hu, andras.micsik@sztaki.mta.hu

Bevezetés

25 évvel ezelőtt, 1994. január 1-én alapítottuk meg az MTA SZTAKI DSD, Elosztott Rendszerek Osztályt (dsd.sztaki.hu). Az osztály kezdetben, az akkori Magyarországon még újdonságnak számító World Wide Webes technológiák hazai bevezetésében, elterjesztésében volt úttörő. Az ország első webszolgáltatásai (pl. SZTAKI Szótár), kormányzati honlapjai (www.kormany.hu, www.kancellaria.gov.hu), webes műalkotásai (pl. Nightwatch, SZTAKI Gallery) létrehozása mellett gyorsan kialakult az a tematikus profil, mely azóta is töretlenül jellemzi az osztály kutatás-fejlesztési tevékenységét. Ebben a szakmai profilban az (elosztott) digitális könyvtári és archívum rendszerek kutatás-fejlesztése kiemelkedő és meghatározó szerephez jutott és jut a mai napig is, a digitális könyvtárak az MTA SZTAKI DSD osztály alapvető, szakmai identitásképző témaköre.

A 25 év során elért szakmai eredményeinket áttekintve, néhány jelentősebb, a témakörbe eső projektet mutatunk be, konkrét szakmai feladatok köré csoportosítva azokat.

A keresés aspektusai

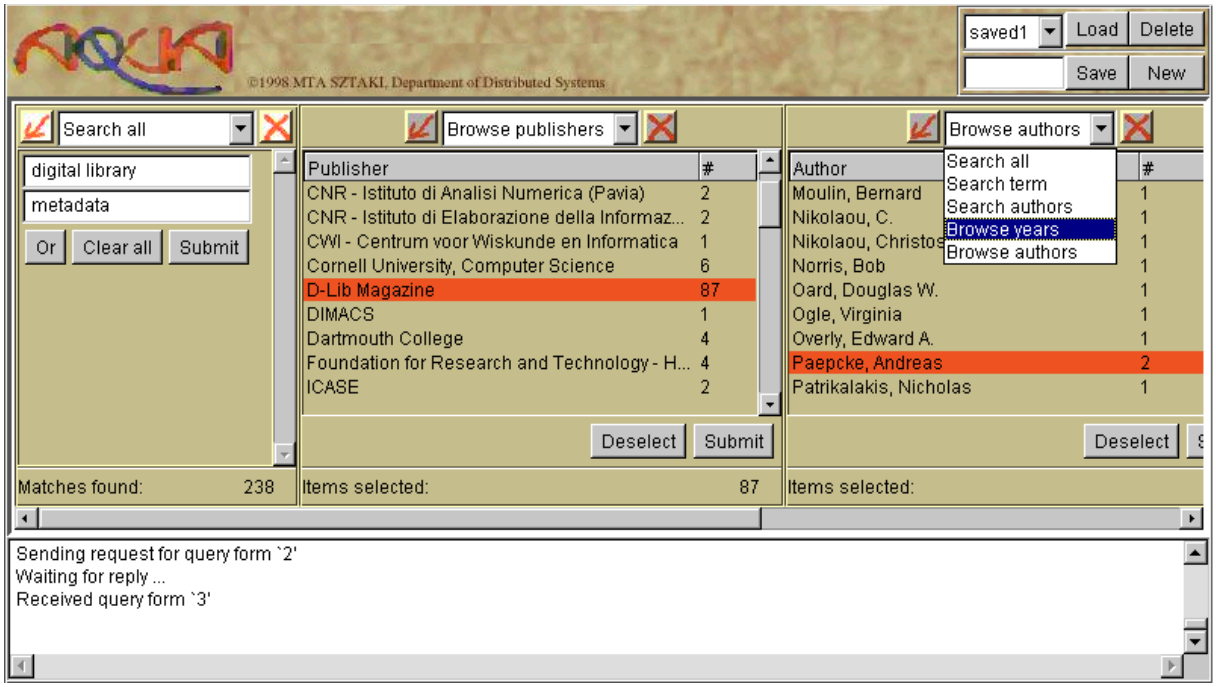
A digitális könyvtárak a World Wide Web elterjedésének egészen korai szakaszában megjelentek (sőt, voltak FTP és Gopher alapú digitális könyvtárak is, de ki emlékszik már rájuk?). Akkoriban azonban még nem voltak univerzális webes keresők (lásd Google), és így a több helyen egyszerre történő, elosztott keresés igénye hamar felmerült. A számítástechnika és számítógéptudomány diszciplínán belül az első digitális könyvtár az USA-ban épült meg és az NCSTRL[1] (Networked Computer Science Technical Report Library) nevet kapta. Az NCSTRL elosztott rendszerként egymáshoz kapcsolt könyvtári csomópontok hálózata volt, melyben egy elosztott keresési algoritmus alapján lehetett megtalálni a keresett digitális objektumokat (kutatási jelentéseket, tudományos cikkeket, publikációkat). A DSD osztály megalakulásával szinte egyidejűleg még 1995-ben felállítottuk az MTA SZTAKI-ban az NCSTRL első európai csomópontját és bekapcsoltuk az NCSTRL hálózatba. A rendszer az elosztott és központi keresési módszereket kombinálta, és sok azóta alapvetőnek tartott szolgáltatást vezetett be, mint például egységes API, perzisztens linkek stb. és a hamarosan megjelenő OAI-PMH protokoll is sokban hasonlít erre a korai API-ra.

A digitális könyvtári kutatási területen Európa jelentősen elmaradt az USA mögött, amit az Unió a DELOS LTR (Long Term Research) ESPRIT projekt indításával igyekezett behozni. A DELOS LTR projekt alapkutatási projektként meghatározó szerepet játszott abban, hogy Európában létrejött a digitális könyvtári

1 http://www.cs.cornell.edu/NCSTRL/

szakmai témakörrel foglalkozó kutatók és fejlesztők nemzetközi közössége, és megteremtődött az európai digitális könyvtári kutatások humán erőforrás alapja, egyben akkor jelentős pénzügyi forrásokat, támogatást is kapott e szakmai terület. Az MTA SZTAKI DSD a kezdet kezdetén történő részvétele a DELOS LTR projektben tette lehetővé azt, hogy elkezdődjön Magyarországon is a digitális archívumok kutatása és fejlesztése, immáron harmonikusan beágyazva az európai digitális könyvtári K+F erőfeszítésekbe. Később az MTA SZTAKI DSD folyamatosan jelen volt az európai digitális könyvtári kutatói közösségben a DELOS NoE, illetve a DELOS NoE2 európai (FP4 és FP5) projektek tagjaként és számos más hazai és nemzetközi K+F projektben kamatoztathatta az európai szinten kifejlesztett, illetve az ott elsajátított technológiákat, műszaki megoldásokat.

Az MTA SZTAKI az ERCIM (European Research Consortium for Informatics and Mathematics) konzorciumba való belépése után jelentős szerepet kapott az ERCIM saját elosztott digitális könyvtári rendszerének létrehozási folyamatában. Elkészült az ETRDL[2] (ERCIM Technical Reference Digital Library), melyben az MTA SZTAKI hazai csomópontot üzemeltetett és az ETRDL (Európa) - NCSTRL (USA) kapcsolat létrehozásáért volt felelős. Később az ETRDL technikai alapjain készítettük az AQUA keresőnket, amely korai Java applet technológiával az ismétléses finomítás (iterative refinement via facets) keresési módszert támogatta[3].



1. ábra: Az AQUA keresőfelület 1998-ból

2 A. Andreoni et al.: The ERCIM Technical Reference Digital Library. D-Lib Magazine December 1999

3 L. Kovács, A. Micsik, B. Pataki: AQUA: query visualization for the NCSTRL digital library. Proceedings of the fourth ACM conference on digital libraries. Berkeley, 1999.

A nagy keresőmotorok hatására a központosított keresőszolgáltatások váltak népszerűvé, mivel gyorsabbak és megbízhatóbbak voltak elosztott társaiknál. Ezek elterjedését roppant mód felgyorsította az OAI-PMH protokoll, melyet először a magyar OSZK-SZTAKI Hektár[4] projekten belül próbáltunk az OSZK MEK-kel együtt itthon népszerűsíteni. A Hektár projektben kialakított kereső megoldás az általunk kifejlesztett NDA@SZTAKI digitális könyvtári rendszer felületen is megjelent, mely az időközben hasonló elvek mentén megvalósult Nemzeti Digitális Adattár alternatív keresőfelülete volt. Ez a szolgáltatásunk jelenleg az OAI kereső[5] nevű reinkarnációjában él tovább, és a nyílt teljes szövegű repozitóriumok és folyóiratok közös, országos keresőfelületét nyújtja.

Továbbra is foglalkoztatott minket azonban az elosztott digitális könyvtárak és archívumhálózatok problémaköre, és a kis- és közösségi rádiózás szükségleteinek megfelelő peer-to-peer elosztott rádióarchívum-hálózatot fejlesztettünk ki a StreamOnTheFly EU projektben[6]. Itt a csomópontok meghatározhatták nyelvi, regionális, vagy bármely más alapon, hogy mely más csomópontokkal cserélnek metaadatokat.

A StreamOnTheFly európai hálózat több mint 10 éven keresztül üzemelt megbízhatóan és biztosította az európai közösségi rádiócsatornák archiválási igényeit, illetve a csatornák közötti műsorcserét, mintegy virtuális, ingyenes piacteret hozva létre a multimédia (leginkább audió) tartalmak kicserélésére és újrafelhasználására. A kis- és közösségi rádiócsatornák ugyanis mindig is erőforráshiánnyal küzdöttek, ezért a StreamOnTheFly hálózat jelentősen hozzájárult e csatornák mindennapi takarékos működéséhez, fennmaradásához.

A StreamOnTheFly hálózat felbomlása után több mint 10 évvel digitális "maradványként" a radio.sztaki.hu[7] oldalon ma is belehallgathatunk magyar városi rádiók korábban archivált műsoraiba.

A nem szöveges média formátumok (audió, kép, videó stb.) terjedésével az ezekben való keresés lett az aktuális probléma. Az MTA SZTAKI CrossMedia projektünkben a képi információk és metaadatok (szemantikus) keresésének kombinálási lehetőségeivel kísérleteztünk[8].

Végső soron a plágiumkeresés is egy ilyen újfajta keresési technika, amely sokféle dokumentumformátumból kivonja a szöveget, és észleli az egyező szövegrészeket a különböző dokumentumokban. Az MTA SZTAKI KOPI[9] plágiumkeresőjét 2004-ben

4 http://hektar.sztaki.hu/index.html
5 http://oaikereso.sztaki.hu/
6 Kézdi Tamás, Kovács, László, Micsik András, Pataki Máté: Elosztott digitális hangtárak a közösségi rádiózásért. Networkshop 2003, Pécs
7 http://radio.sztaki.hu/
8 Gergó Márton, Havasi László, Mátételki Péter, Micsik András, Kovács László, Szirányi Tamás: Képi és szemantikus keresőalgoritmusok kutatását támogató közösségi platform Networkshop 2012, Veszprém
9 https://kopi.sztaki.hu/

hoztuk létre, és azóta üzemeltetjük. Ez volt az első plágiumkereső Magyarországon, amely jól kezelte a magyar nyelvű szövegeket, és később 2011-ben, a világon elsőként, valósított meg jó minőségű fordítási plágiumkeresést[10] is, amely például detektálni tudta, ha valaki a dolgozatába az angol nyelvű Wikipedia-ból magyarra fordított szövegrészeket illesztett be.

Metaadatok, dokumentumok, kapcsolt adatok

Az 1998-ban megjelent Dublin Core (DC) metaadatleíró rendszer forradalmasította a metaadatok világát. A Dublin Core megnyitotta az utat az eddig egymástól elkülönülve fejlődő leíró rendszerek (mint például a MARC), átjárhatóvá, egymásra leképezhetővé tételére, meghatározva a leglényegesebb, esszenciális metaadatok legszűkebb körét. Kezdetben úgy képzelték, hogy az egyes szakágak (pl. kereskedelem, könyvtárak, közigazgatás stb.) mind kialakítják a DC, majd később qDC alapú, de specializált változataikat, az ún. alkalmazási profilokat (application profile). Ezek szabványosítására, áttekintésére és kényelmes, grafikus szerkesztésére szolgált az azonos nevű EU projekt keretében létrehozott CORES rendszer, melynek fejlesztésében segédkeztünk, és melyet sokáig üzemeltettünk[11].

Az alkalmazási profilok szakmai körökben történő lassú terjedésével egyidejűleg megjelent egy alapvetőbb - persze bonyolultabb - koncepció, a Szemantikus Web, amellyel egyszerű metaadatrekordokat, tezauruszokat és logikai modelleket (ontológiákat) egyaránt le lehetett írni. A cél tehát az lett, hogy minden területen létrejöjjön a tudásreprezentációhoz szükséges ontológiák megfelelő halmaza. A Szemantikus Web később egyszerűbb és hatékonyabb formában kapcsolt adatok (linked data (LD), linked open data (LOD)) néven indult rohamos fejlődésnek.

A kapcsolt adatokat a már korábban említett CrossMedia projektben használtuk szemantikus keresésre, vagyis konkrét szóelőfordulások helyett egy tezaurusz részgráfja alapján kerestünk (pl. szinonimák, hiponimák bevonásával) megfelelő képeket.

2011-ben elindult a lod.sztaki.hu szolgáltatás az Intézetben, amely az OAI alapon összegyűjtött hazai kulturális adatokból előállított kapcsolt adathalmaz szolgáltatás, és több mint 11 millió egyszerű tényből (RDF triple) áll össze.

Végül 2016-ban, a COURAGE EU projekt indulásával sikerült egy a kezdetektől és alapjaiban is kapcsolt adatos (RDF alapú) adattár-megoldást létrehoznunk.

Teljes rendszerek fejlesztése és üzemeltetése

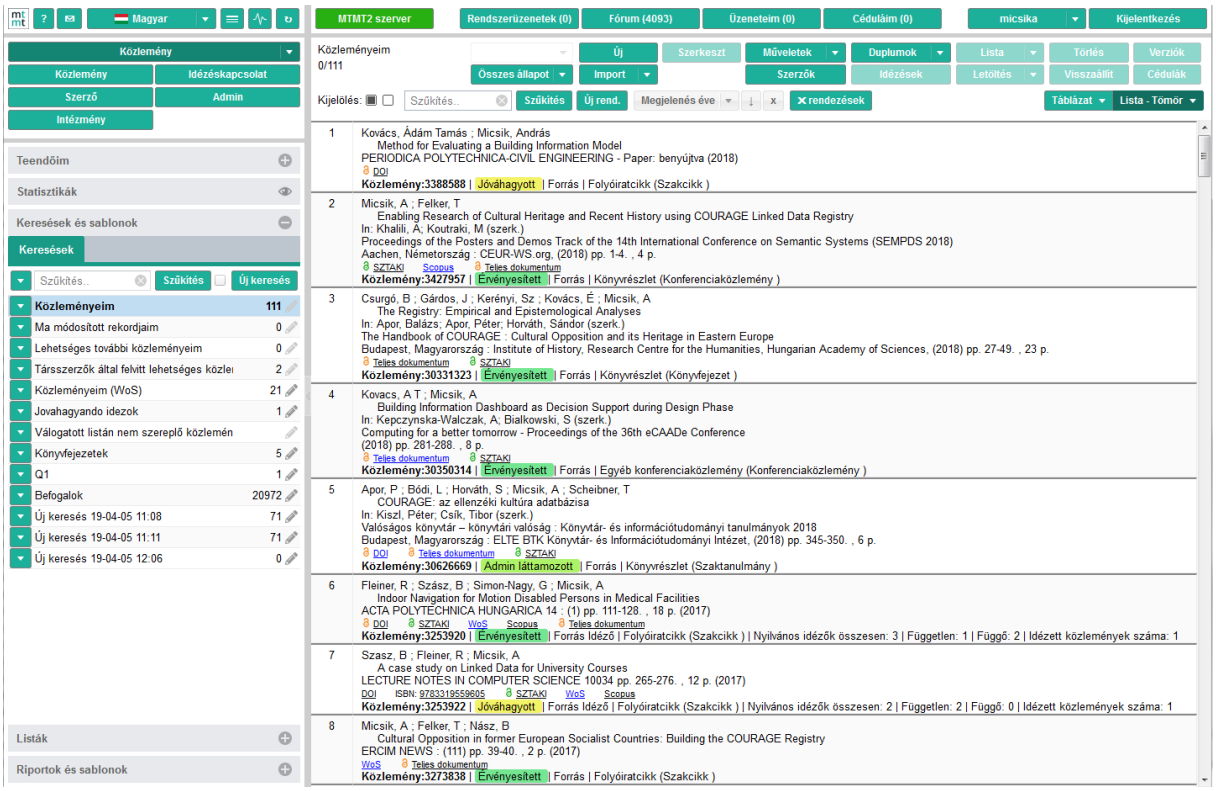
A fejlesztések mellett mindig is szerettük, ha a szoftvereink működnek és széles körben használják is azokat. Ezért fokozott gondot fordítottunk az üzemeltetésre és a folyamatos fenntartásra, támogatásra. Néhány komplexebb rendszert kiemelnénk a létrehozott szolgáltatás-portfóliónkból.

10 Pataki Máté: Plágiumkeresés különböző nyelvek között. Networkshop 2011, Kaposvár

11 Fülöp Csaba, Kovács László, Micsik, András: Metaadatsémák nyilvántartása szemantikus web alapon. Networkshop 2004, Győr.

A már említett StreamOnTheFly európai rádióarchívum rendszer több csomópontból álló, elosztott rendszer volt, ahol az egyes csomópontok dönthettek arról, hogy mely más csomópontokkal lépnek partnerségre. Egy csomóponton belül több "rádióadó" archiválhatta műsorait, a szerkesztői jogok szabályozásával akár sorozatonként más-más ember végezhette kényelmesen a hanganyagok, metaadatok feltöltését és/vagy szerkesztését. A felhasználók pedig saját, akár több órás egyéniesített (perszonalizált) rádióműsort állíthattak össze az archívumból és stream formájában élvezhették azt.

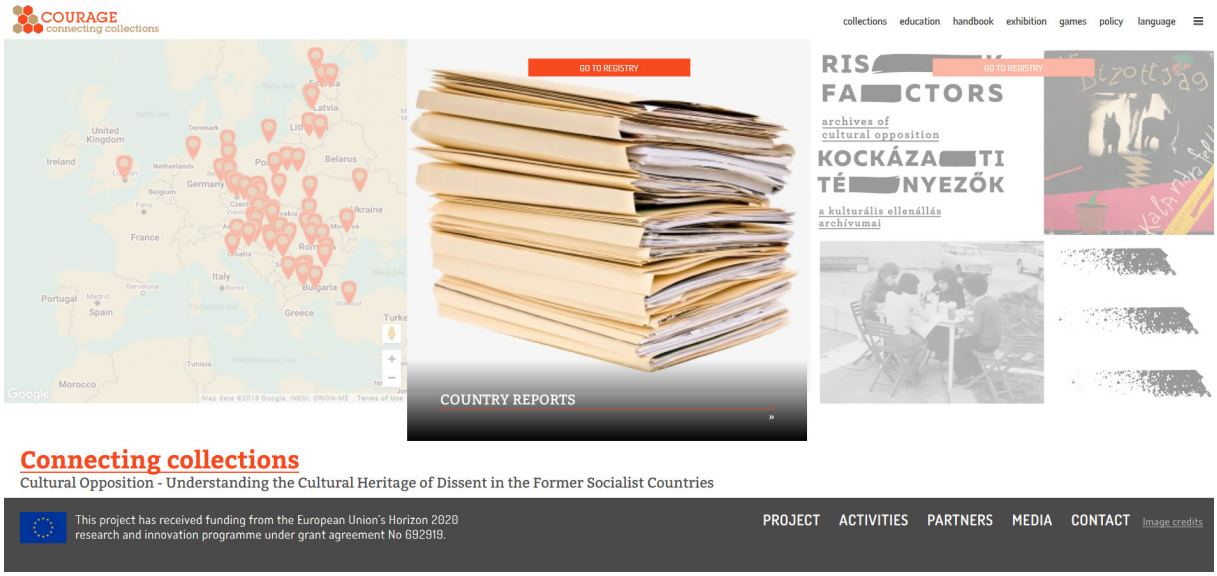
A KOPI plágiumkereső is összetett osztott rendszer, mivel a keresés összetett belső folyamatának lépéseit különböző célszerverek végzik. A SZTAKI Szótár is ide kívánkozik, bár az csak szavak archívuma, de napi 80-100 ezer látogatót szolgál ki 8-10 szerver precíz együttműködésével. A szótár tartalmak tárolását pedig egy gráfadatbázis (neo4j) szolgálja ki. A SZTAKI Szótár kapcsán érdemes megjegyeznünk, hogy itt gondosan ügyeltünk a szótári kereső URL-ek hosszútávú megőrzésére. Ezért, ha valaki 1995-ben a szótár indulásakor berakott egy linket (keresést) egy szóra a honlapjára, ha erre ma 2019-ben ráklikkelnek, akkor a szótár ma is megadja a megfelelő fordítást.



2. ábra: Az MTMT2 szerkesztői felülete

Az MTA SZTAKI Elosztott Rendszerek osztályán fejlesztettük ki a Magyar Tudományos Művek Tára 2018. novembere óta éles üzemelésű 2-es (MTMT2) szoftververzióját[12]. Ez volt pályafutásunk során eddig a legbonyolultabb és legnagyobb saját fejlesztésű szoftver-rendszer. Az MTMT2 a külvilággal egy REST API-n keresztül kommunikál, ezt használja a teljesen új nyilvános felület, amelyen bejelentkezés nélkül akár mobilon is lehet böngészni a szerzők, csoportok munkásságát, vagy az egyes témakörökben megjelent cikkeket. A szerkesztői felület ún. egyablakos Javascript alkalmazás, amely egy teljes professzionális munkakörnyezetet (workspace) ad a közlemények felviteli, kiegészítési stb. teljes körű adatkurátori feladatainak elvégzésére. Az MTMT2 rendszert több mint 60.000 felhasználó, az ország teljes kutatói szférája használja rendszeresen.

Végezetül a COURAGE[13] EU projekt keretében az általunk létrehozott digitális archívum rendszerét említjük meg, amely a technikai fejlettségével emelkedik ki. Az adatokat RDF triple store tárolja a külön e célra készült COURAGE ontológia, mint adatséma alapján. A felhasználói felület nagy részét is kapcsolt adat konfiguráció, illetve SPARQL lekérdezések vezérlik. Az adatfelvitel során egyből létrejönnek a kétirányú adatkapcsolatok, relációk, melyek mentén a létrehozott szemantikus tudásgráf sokoldalúan böngészhető. A rendszergazda pedig menet közben a rendszer leállása nélkül fel tud venni új adatmezőket, vagy meg tudja változtatni a meglévőket (séma módosítás), azok megjelenési módját is beleértve.



3. ábra: A COURAGE projekt honlapja

12 Micsik András, Pataki Balázs, Kovács László et al.: A Magyar Tudományos Művek Tára 2.0 verziójának fejlesztése. Networkshop 2016

13 Micsik András: Besúgók és provokátorok - történelmünk kutatása kapcsolt adatokkal. Networkshop 2017

Utószó

Az MTA SZTAKI DSD, a kutatóintézet Elosztott Rendszerek Osztálya 25 éven keresztül töretlenül dolgozott a digitális könyvtárak és archívumok kutatás-fejlesztése területén és létrehozott egy egyedülálló tudás és technológiai megoldás portfoliót, mely jelenleg a teljes magyar felhasználói közösség szolgálatára áll. A jövőben szeretnénk tovább foglalkozni a tudásreprezentáció és tudásfúzió felmerülő elméleti és gyakorlati problémáival. A portfolióban nem csupán korábbi megoldásaink, illetve az elsajátított technológiák, know-how-k stb. reprezentálnak értéket, hanem az az időközben az osztályon kialakított képességünk, mely bonyolult műszaki-tudományos problémák kezelését holisztikusan szemlélve egyidejűleg képes akár felfedező, alkalmazott kutatási és fejlesztési tevékenységeket kombináltan végezni. Az osztály munkatársai e képesség birtokában bátran vállalkoznak bármely új kihívás esetén egy lehetséges, a gyakorlatban működő, korszerű és tudományosan is értékelhető megoldást megtalálni.