Egy NDA-kompatibilis keresőmotor

Kiss Gergő, Kovács László, Micsik András
{gergo.kiss, laszlo.kovacs, micsik}@sztaki.hu
Elosztott Rendszerek Osztály
MTA SZTAKI
Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutató Intézet

A Nemzeti Digitális Adattár (NDA) a magyar kormány által vezérelt kezdeményezés, amely a magyar kulturális értékeket kívánja digitális formában elérhetővé tenni. Az elérés szó ebben az esetben az információhoz való hozzáférés legtágabb jelentését takarja, azaz a digitalizálás, a katalogizálás, a kereshetőség, stb. mind ide értendő. Ezáltal az NDA a XXI. század tudás-alapú társadalma számára elengedhetetlen, új technikai infrastruktúra megteremtésén is munkálkodik.

Az NDA honlapján [1] olvasható ismertetők alapján az NDA – neve ellenére – nem egy archívum, hanem egy program, amely a résztvevő intézmények közötti kooperációt emeli magasabb szintre. Hasonlóan olvashatjuk, hogy "Az NDA hangsúlyozza az önszerveződés, a közösségi együttműködések, az egyéni és intézményi kooperációk fontosságát…".

E gondolatok jegyében döntöttek úgy a SZTAKI Elosztott Rendszerek Osztályának munkatársai, hogy a korábbi, sikeres HEKTÁR projekt [2] eredményeit felhasználva, egy második keresőszolgáltatást valósítanak meg (http://nda.sztaki.hu). Ezzel is demonstráljuk, hogy az NDA architektúrája egyszerűen lehetővé tesz több, eltérő felhasználói szolgáltatás megvalósítását, melyek egymás alternatívájaként, vagy egymást kiegészítve tudnak működni.

Az NDA számítógép-hálózati szempontból nézve egy OAI [3] alapú decentralizált, elosztott rendszer, amely a következő elemekből épül fel:

- Adatszolgáltató, vagy OAI szerver, amelyet általában az adatvagyon gazdája üzemeltet, és az adatvagyon elemeit leíró metaadat rekordok begyűjtését teszi lehetővé
- Szolgáltatásgazda, vagy OAI kliens, amely a begyűjtött metaadatrekordok alapján különféle szolgáltatásokat valósít meg a nagyközönség felé vagy az NDA-n belül.
- Protokollgazda, amely a hálózat elemei közötti kommunikációhoz, illetve a metaadatrekordok cseréjéhez szükséges egységes protokollokat, formátumokat és sémákat definiálja, és ezeket a definíciókat nyilvánosságra hozza.

Az NDA szolgáltatás megvalósítása

A HEKTÁR projekt (amely az első nyilvános OAI alapú közös kereső volt Magyarországon) eredményei alapján elhatároztuk, hogy a létrehozott szoftverrendszert hozzáigazítjuk az NDA-hoz. A legnagyobb eltérés az NDA és a HEKTÁR között (a keresés szempontjából) az eltérő metaadatsémák használata. Amíg a HEKTÁR az OAI által kötelezőként megszabott oai_dc sémán alapult, addig az NDA-ban több különböző metaadatséma is használható, és új metaadatsémák bevezetése folyamatosan várható.

Metaadatsémák

Napjainkra elfogadott az a gyakorlat, hogy az általános metaadatsémákat egyes területek számára specializálják oly módon, hogy a speciális sémák és az alapul szolgáló sémák közötti kapcsolat megmaradjon, és a speciális metaadatok egy tágabb, kevésbé speciális kontextusban is értelmezhetőek

maradjanak [9]. Az a rendszer, amelyet a Dublin Core nyelvtanának is lehet nevezni, rögzített alapsémákat (névtereket) és ún. alkalmazási profilokat különböztet meg [7]. Az előbbi egy szabványosító szervezet által kiadott eredeti metaadatséma, míg az utóbbi egy speciális terület vagy alkalmazás számára létrehozott módosított, finomított séma. Definíció szerint az alkalmazási profil egy vagy több névtérből gyűjt össze elemeket, de nem vezethet be új elemeket. A kiválasztott elemeket az alkalmazási profil az alábbi szabályok szerint módosíthatja:

- finomíthatja annak értelmezését, de csak az eredeti értelmezés szűkítésével: például a formátum elem finomítása rádióműsorok esetén a műsor hosszának megadása lehet;
- módosíthatja az elem előfordulási módjait: kötelező, ismételhető stb.;
- megszabhatja az elem értékkészletét egy adott szókészlettel (pl. riport, magazin, hírműsor, zene) vagy egy ún. kódolási sémával (pl. dátum formátuma).

Az NDA metaadatsémái is nagyjából ennek az elvnek alapján épülnek fel, tehát az alap Dublin Core elemek finomított változataiból állítanak össze sémákat annak érdekében, hogy a nyomtatott anyagokat, hanganyagokat, képeket, filmeket, stb. pontosabban le lehessen írni.

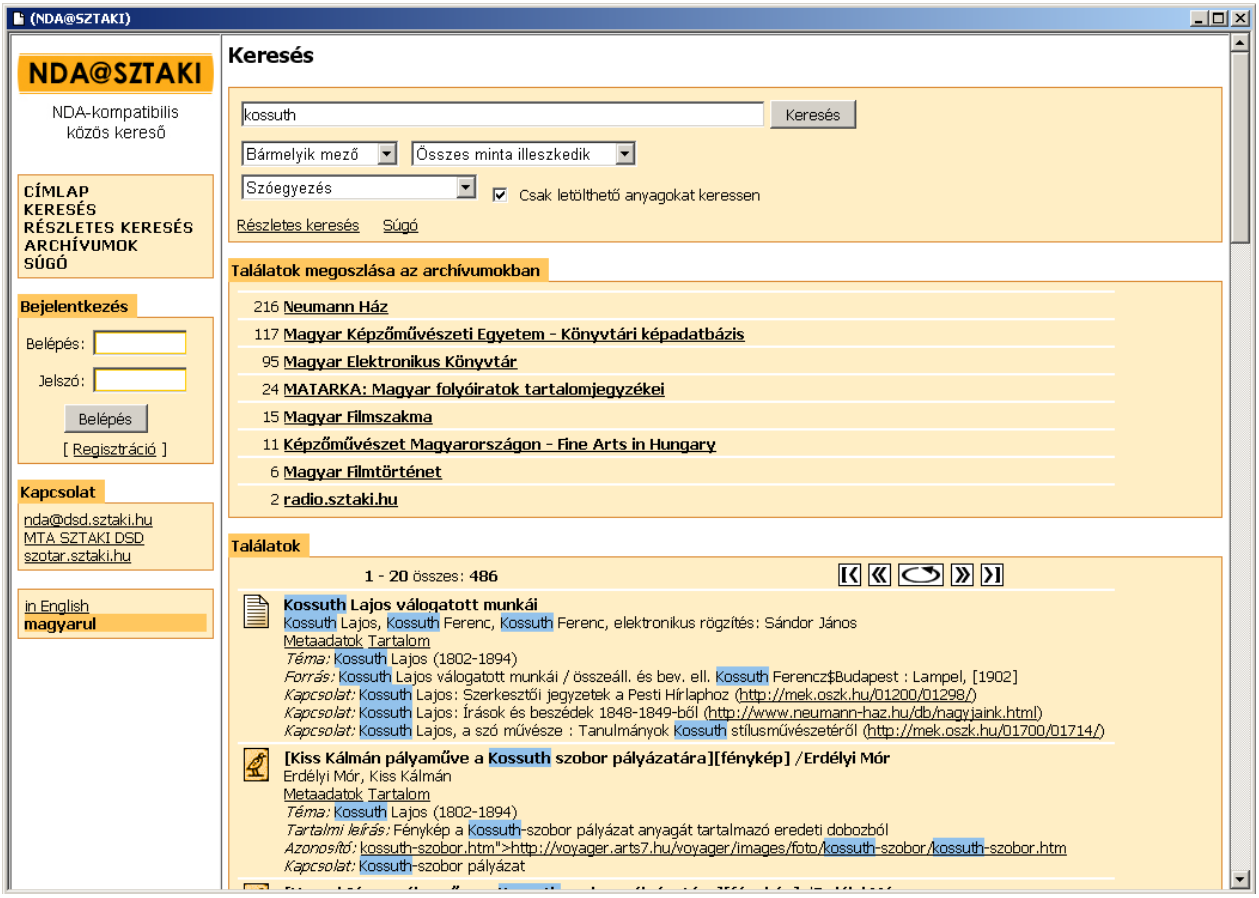
A CORES projekt keretén belül az MTA SZTAKI részvételével egy olyan átfogó metaadatséma-nyilvántartás készült el, amely a fenti modell alapján rendszerezi és összefogja a különböző metaadatsémakat, és ezen felül támogatja új sémák készítését is [4,5]. Mivel a CORES nyilvántartást is szerzők üzemeltetik (http://cores.dsd.sztaki.hu), természetesnek látszott, hogy az NDA metaadatsémáit is ebben a nyilvántartásban tároljuk. Egy olyan dinamikus keresőrendszert terveztünk, amely a metaadatsémák definícióit egy külső nyilvántartásból (jelen esetben CORES) veszi. A sémák tehát nincsenek az alkalmazásba statikusan beépítve, hanem automatikusan mindig a legfrissebb sémák alapján működik a rendszer.

Begyűjtés

A szolgáltatás motorja a háttérben meghúzódó begyűjtő (szüretelő, harvester) modul, amely a regisztrált adattárakat (repository) naponta felkeresi, és az OAI-PMH protokoll alapján az új vagy módosított metaadatrekordokat az OAI szervertől (data provider) lekéri. Az újonnan érkező metaadatrekordokat a modul a rekordban definiált metaadatséma alapján értelmezi, illetve saját fejlesztésű heurisztikák alapján megpróbál egyes mezőket, bejegyzéseket szinkronizálni. Ez azt is jelenti, hogy egyes adattárakból érkező metaadatrekordok esetén előfeldolgozást kell alkalmazni annak érdekében, hogy a kereshetőség szempontjából a többi adattal konzisztensek legyenek. Egyszerűbb példa erre a Language elemben használt nem szabványos jelölések egységes formára hozása (Magyar, magyar, hu, hu_hu, HUN, stb. helyett egységesen hun). Közbeeső lépésként a modul előállítja a generikusan használható címet, URL-t (ha van), dokumentumtípust, stb. amelyek a keresésnél és listázásnál játszanak fontos szerepet. Végül a rekordokat a speciálisan a több metaadatsémán alapuló keresésre optimalizált adatbázisba illeszti a begyűjtő modul.

Felhasználói felület

Az NDA@SZTAKI szolgáltatás 2004 májusától üzemel (http://nda.sztaki.hu), és folyamatosan gyarapodó választéka 2005 áprilisra megközelítette a félmillió metaadatrekordot. Napi forgalma átlagosan 5000 lekérés. A szolgáltatás egyformán működik magyar és angol nyelven is.



1. ábra: A szolgáltatás alap keresőfelülete

A felület lehetőséget ad a regisztrált archívumok böngészésére, az archívumok gyűjteményeit (set) is áttekinthetjük, valamint a tételek listáját is többféle rendezés és gyorskeresés alapján böngészhetjük. Az alap keresőfelület (1. ábra) egyszerű, de hatékony keresési módot próbál nyújtani: legtöbb esetben elég csak a keresett szavakat beírni, a keresési opciókat változatlanul hagyhatjuk.

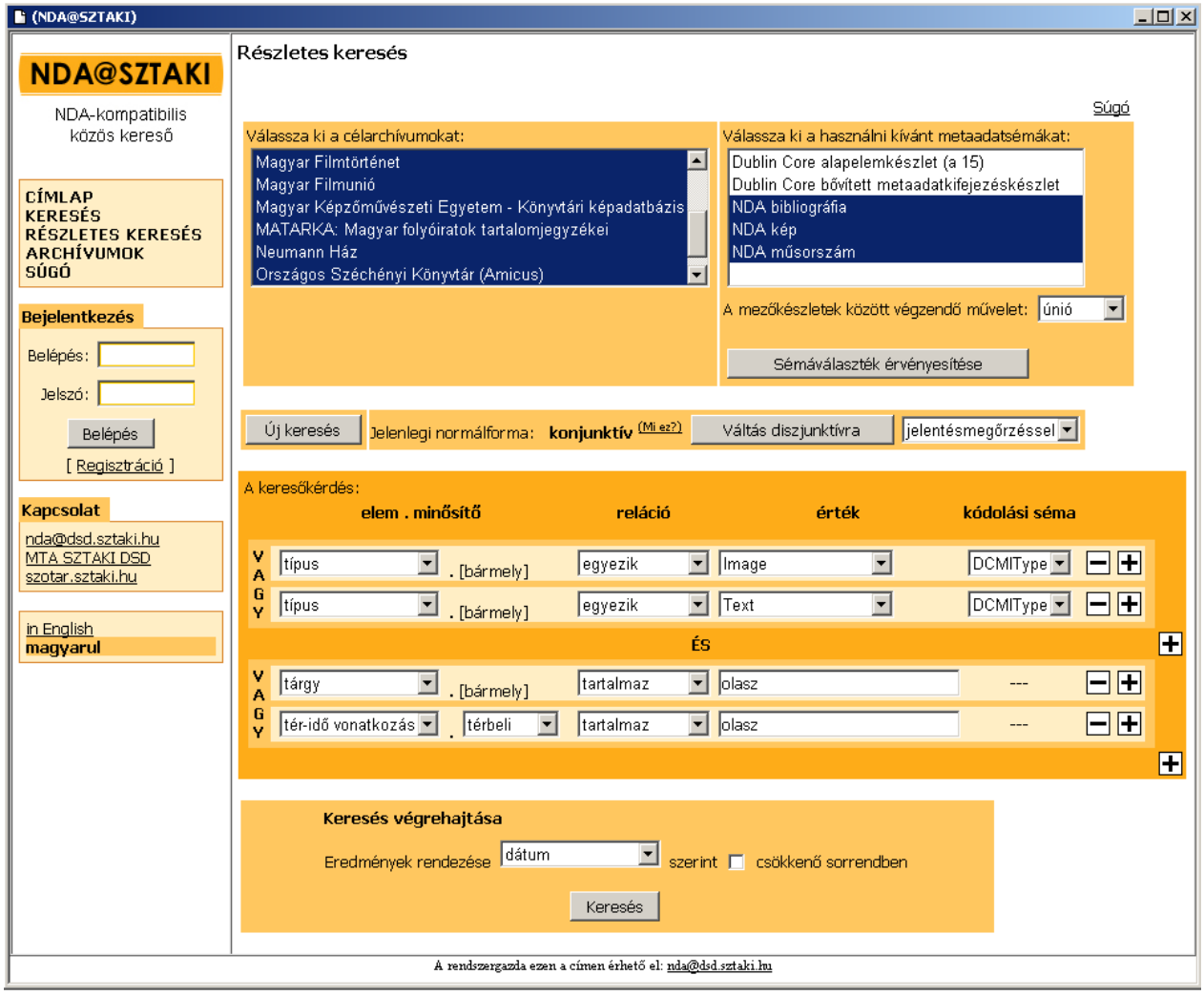
A gazdag metaadatsémák alapján történő keresés képességeit demonstráljuk a „Részletes keresés” menüpont alatt (2. ábra). Ez a felület tetszőleges logikai kifejezés építésére ad lehetőséget egyszerű és átlátható módon. A keresés lépései abban az esetben, ha minden beállítási lehetőséget kihasználunk:

- Azon archívumok kiválasztása, melyekben keresni akarunk
- Metaadatsémák kiválasztása, melyeket felhasználunk a keresőkérdésben
- A keresőkérdés összeállítása
- Az eredménylista rendezésének kiválasztása
- A keresés végrehajtása, az eredmények böngészése
- Visszakapcsolás a keresőkérdés szerkesztőbe, a keresőkérdés finomítása és újrafuttatása
- A keresőkérdés elmentés a személyes adatok közé, így az később is végrehajtható lesz

A keresőkérdés a szerkesztőben konjunktív vagy diszjunktív logikai normálformában áll elő, azaz a kifejezés *(A vagy B…) és (C vagy D…)*…illetve *(A és B…) vagy (C és D…)*…alakú. ahol A,B,C,D egyszerű (atomi) kifejezések, mint például *“szerző neve János (tartalmazás)”*, *“módosítás dátuma korábbi, mint 2005”*, stb. Egy atomi kifejezés a következő részekből áll:

- Metaadat séma elem
- A választott séma elem minősítője (ha van)
- Az egyezés formája (pl. tartalmazás, prefix tartalmazás, dátum esetén korábbi, későbbi, stb.)
- A keresett szavak, értékek (pl. dátum)
- Az érték kódolási sémája (ha van), például DCMIType a dokumentum típusának megadására

Az atomi kifejezést beállító mezők mindig az adott kontextusnak megfelelően változnak (pl. minősítők, az egyezés formája). Negációra is az egyezés formájának beállításával van lehetőség (pl. nem tartalmaz).



2. ábra: A keresés tetszőleges metaadatséma-elemek alapján

A kifejezés szerkesztő grafikailag hangsúlyozza a kereső kifejezés szerkezetét, atomi kifejezéseket a plusz és mínusz ikonokkal tudunk hozzáadni illetve törölni. A felhasználók korábban összeállított és elmentett keresőkérdéseiket is felhasználhatják szerkesztés közben, mint egy atomi kifejezést beszúrva a jelenlegi keresőkérdésbe.

A logikában járatos felhasználók számára hasznos funkció az átváltás a konjunktív és diszjunktív logikai normálformák között. Ezt kétféleképpen lehet megtenni: egyszerűen az ÉS és VAGY operandusok

felcserélésével, vagy pedig jelentésmegőrzéssel, mely esetben a szerver kiszámolja a jelenlegi kifejezésnek ekvivalens logikai kifejezést a másik normálformában.

A felhasználás 4 havi megfigyelése során azt tapasztaltuk, hogy a felhasználói tevékenység legnagyobb részét az archívumok és gyűjtemények böngészése (20%) és a metaadatrekordok olvasása (12%) teszi ki. A keresés eredménye 5%-ban származott egyszerű keresésből, és 0,3%-ban részletes keresésből (az összes hozzáféréshez képest). A hozzáférések 29%-a Google találatlistából érkezett, mivel a Google rendszeresen indexeli szolgáltatásunkat. A hozzáférések 23%-ában a felhasználók egy linket követtek, amely a dokumentum tartalomra vagy más hivatkozott weblapra mutatott.

Összefoglalás

Az NDA kezdeményezése alapvető fontosságú a magyar kulturális javak elérhetővé tételében. Az NDA elgondolása abszolút korszerű, mint nemzeti szintű, önszervező, OAI-alapú elosztott digitális archívum rendszer, és példát teremt hasonló jellegű megoldások számára a turizmusban, közigazgatásban és egészségügyben.

Egy ilyen nagyszabású megközelítés esetén a metaadatsémák módosulása, evolúciója természetes jelenség. Az itt bemutatott szolgáltatás élő példa arra, hogyan lehet a metaadatséma nyilvántartásokat a metaadat tárolókkal összekapcsolni. Ez a megoldás a CORES és a HEKTÁR projektek eredményeit is egyesíti, lehetőséget nyújt a begyűjtött metaadat rekordok automatikus ellenőrzésére, valamint a keresési felületet automatikusan a metaadatsémák legfrissebb változatához lehet igazítani. A bemutatott újszerű keresőfelület tisztán HTML és Javascript alapon teszi lehetővé logikai keresőkifejezések egyszerű és áttekinthető szerkesztését. A keresőfelület további specialitása, hogy támogatja több metaadatséma egyidejű felhasználását.

Hivatkozások

[1] Nemzeti Digitális Adattár honlap: http://www.nda.hu

[2] Kiss Gergő, Kovács László, Micsik András, Moldován István: „HEKTÁR: Hazai elektronikus könyvtári rendszerek összekapcsolása”, Networkshop 2004, Győr

[3] Open Archives Initiative, http://www.openarchives.org

[4] Rachel Heery, Pete Johnston, Csaba Fülöp, András Micsik: „Metadata schema registries in the partially Semantic Web: the CORES experience”, 2003 Dublin Core Conference, Seattle, Washington USA, http://www.siderean.com/dc2003/102_Paper29.pdf

[5] Fülöp Csaba, Kovács László, Micsik András: „Metaadatsémák nyilvántartása szemantikus web alapon”, Networkshop 2004, Győr

[6] Diane Hillmann: „Using Dublin Core”, http://www.dublincore.org/documents/usageguide/

[7] Rachel Heery, Manjula Patel: “Application Profiles: mixing and matching metadata schemas”, *Ariadne* 25. (2000 September), http://www.ariadne.ac.uk/issue25/app-profiles/

[8] Rachel Heery, Pete Johnston, Dave Beckett, Damian Steer: “The MEG Registry and SCART: complementary tools for creation, discovery and re-use of metadata schemas”, Proceedings of the International Conference on Dublin Core and Metadata for e-Communities, 2002.

[9] Fülöp Csaba, Kovács László, Micsik András: “A metaadatsémák és a szemantikus web: egységesítés és specializáció a metaadatok világában”, Tudományos és Műszaki Tájékoztatás 51. évfolyam (2004) 7. szám