

Az EHM (EPA–HUMANUS–MATARKA) közös kereső létrejötte

A Miskolci Egyetem Könyvtár, Levéltár, Múzeum által 2009-ben beadott TÁMOP 3.2.4 „Tudásdepó-Expressz” – *A könyvtári hálózat nem formális és informális képzési szerepének erősítése az élethosszig tartó tanulás érdekében* pályázat 194 millió forint támogatást kapott. A program kezdete: 2009. október 1., a projekt vége: 2011. június 30. volt. A projekt igen sok célt tűzött maga elé és sok eredményt ért el, de ezek között – országos szempontból mindenképp a legfontosabb – az EPA–HUMANUS–MATARKA folyóirat-adatbázisok egy ponton való lekérdezése.

A három adatbázis

EPA Elektronikus Periodika Archívum (<http://epa.oszk.hu>)

Az EPA 2004 óta működik, a Magyar Elektronikus Könyvtár (MEK, <http://mek.oszk.hu>) kezdeményezésére jött létre azzal a céllal, hogy – a MEK monografikus gyűjteménye mellett – a magyar vonatkozású elektronikus folyótató kiadványok feltárásával, nyilvántartásával valamint stabil digitális folyóirat-archívum építésével foglalkozzon. Az EPA-archívum gyűjtőkörébe tartoznak a nyilvánosan és ingyenesen online szolgáltatható, magyar nyelvű vagy magyar kiadású, heti vagy annál hosszabb periodicitású (a határon túli kiadványok kivételével), oktatási, tudományos vagy kulturális szempontból hasznos időszaki kiadványok. Az archiválás alapja a „szám”. A támogatott formátumok: HTML, XML, PDF, RTF, JPG, PNG, ASCII text.

Az EPA három különböző nyilvántartást egyesít egyetlen katalógusban. Az *archivált* hozzáférésű kiadványok teljes szövegű állományait saját szerverén tárolja, és onnan nyilvánosan online szolgáltatja. A *távoli* hozzáférésű kiadványok teljes szövegű állományaira csak URL-lel hivatkozik a katalógus. Az *offline* hozzáférésű kiadványok teljes szövegű digitális állományainak csak fizikai elérhetőségéről található információ a leírásban (ezek a források online nem érhetők el).

HUMANUS – humán tudományok adatbázisa (<http://www.oszk.hu/humanus>)

A 2008 februárjában átadott HUMANUS célja a humán tudományok területén megjelenő, magyar vonatkozású nyomtatott és elektronikus időszaki kiadványokban, valamint a nyomtatott és elektronikus tanulmánykötetekben (köztük konferencia-anyagokban) megjelent publikációknak a bibliográfiai számbavétele. A HUMANUS konzorciumi formában működik. Tagjai lehetnek mindazon könyvtárak, kiadók, szerkesztőségek, kutatóintézetek, amelyek részt kívánnak vállalni a HUMANUS rendeltetésének megvalósításában.

MATARKA – Magyar Folyóiratok Tartalomjegyzékeinek Kereshető Adatbázisa (www.matarka.hu)

A MATARKA magyar szakfolyóiratok tartalomjegyzégeit dolgozza fel könyvtárak és folyóirat szerkesztőségek összefogásával a Miskolci Egyetem Könyvtár, Levéltár, Múzeum vezetésével. 2002-ben kezdődött a fejlesztése, 2003-tól ismert MATARKA név alatt. A felvett adatok adatbázisba kerülnek, és a szerzők, illetve a címben levő szavak szerint visszakereshetők, az egyes számok külön-külön böngészhetők. A feldolgozás egyrészt kézi adatbevitellel történik, másrészt digitális formában már rendelkezésre álló adatok konvertálásával. A cikkek mintegy 15 százaléka teljes szöveggel közvetlenül is elérhető, a többitől cikkmásolat rendelhető.

A projekt indítása

A közös kereső programozására betervezett pénzüsszeg szerencsére nem volt olyan magas, hogy közbeszerzési eljárást kellett volna indítani. Elég volt három ajánlatot bekérni. Nyilvánvaló volt, hogy olyan szoftverfejlesztőt kell találnunk, aki a három adatbázis közül legalább az egyiket jól ismeri, és készített már könyvtári szoftvert. A MATARKA egyik programozója régebben e sorok írója volt, de 2005 óta ilyen jellegű tevékenységet nem folytat, a MATARKA jelenlegi programozójának pedig megosztott a munkája és többgyermekes családanya. Az EPA programozója nemrég ment nyugdíjba, nem lett volna tanácsos olyan fejlesztőt keresni, akinek kötődése a könyvtárügyhöz fokozatosan csökken. Így egyértelművé vált, hogy az amúgy is legösszetettebb rendszer, a HUMANUS fejlesztője, Zawiasa Róbert, a Bodza Bt. tulajdonosa jöhet szóba. Az Országos Széchényi Könyvtárnak egyéb szolgáltatásait is kezeli, a Szegedi Tudományegyetem könyvtárának számos alkalmazását programozza. Nagyon jól ismeri a könyvtárügyben használatos szabványokat (MARC, Dublin Core, OAI). Természetesen közvetlenül csak arra lehetett megkérni, hogy árajánlatkérésünkre adjon pozitív választ, vagyis lehetőleg legyen a három ajánlattevő között. Szerencsénk volt, az ő ajánlata lett a legkedvezőbb, így 2010. január 29-én megtörtént vele a szerződéskötés.

A szerződésben a munka teljesítésére 11 hónapot kötöttünk ki, 2010. február 1-től december 31-ig tartott a fejlesztési időszak. Zawiasa Róbert a teljes időszak alatt kapcsolatot tartott Csáki Zoltánnal, az EPA könyvtáros szakmai rendszergazdájával, Tamás Kincsovvel, az OSZK Humántudományi Bibliográfiai Osztály vezetőjével, és velem, a cikk szerzőjével, aki a TÁMOP-projektünk MATARKA-s részeit koordináltam.

A munkamenet

Csáki Zoltán bemutatta neki az EPA adatszerkezetét és rendelkezésére bocsátotta az úgynevezett EPAX munkakörnyezetben létrehozott tartalomjegyzékeket. A MEK Egyesület megbízásából ugyanis már régebben elkészült az EPA tartalomjegyzék DTD, amit később XSD formátumba alakították át, és azóta is így használnak. Erre épül az XML fájlokat feldolgozó, weben adminisztrálható mun-

környezet, amelyet az EPA és az XML betűszavak keresztezéséből EPAX-nak neveztek el.

Az EPAX-szal összefüggő tevékenység célja az EPA-struktúrából hiányzó „cikk szint” elérhetővé tétele a felhasználók számára. A tartalomjegyzék feldolgozás csak egy eszköz ehhez a tevékenységhez. Az EPA nem tekinti feladatának a tartalomjegyzékek leírását, csak abban az esetben, ha azok külső forrásból nem érhetőek el. Ideális esetben az EPA az összes tartalomjegyzék adatot külső forrásból gyűjti be (pl. a MATARKA-ból), és csak a teljes szövegű hivatkozással egészíti ki. Az EPA teljes szövegű hivatkozások nélküli tartalomjegyzékeket vesz át, és teljes szövegű hivatkozással kiegészített tartalomjegyzékeket ad vissza.

Az 1. ábra egy XML-ben elkészült EPA tartalomjegyzéket mutat be. Zawiasa Róbert közös keresője ezeket az EPAX fájlokat dolgozza fel. A távoli és offline tételek adataival az EHM projekt nem foglalkozik, csak az EPA-ban ténylegesen archiváltakkal.

```
- <Pack xsi:noNamespaceSchemaLocation="http://mek.oszk.hu/mekdtd/epax/epax.xsd">
  <TOC>
    <Head>
      <dc_title>A Herman Ottó Múzeum Évkönyve</dc_title>
      <Issue year="1980" number="1" type="pdf">19. (1980.)</Issue>
      <host>epa.oszk.hu</host>
      <dc_identifier>02030</dc_identifier>
      <ev_szam_Path>00017/pdf</ev_szam_Path>
    </Head>
  - <Contents>
    - <Section>
      <SectionTitle>Régészet - Archaeology</SectionTitle>
      <Article>
        <Language>hu</Language>
        <Link>HOM_Evkonyv_19_009-041.pdf</Link>
        - <Author invert="no">
          <FamilyName>Sándorfi</FamilyName>
          <GivenName>György</GivenName>
        </Author>
        <Title>Motte típusú várak a történelmi Borsod vármegyében</Title>
        <Range>9</Range>
      </Article>
    </Section>
  - <Section>
    <SectionTitle>Régészet - Archaeology</SectionTitle>
    <Article>
      <Language>en</Language>
      <Link>HOM_Evkonyv_19_039-041a.pdf</Link>
      - <Author invert="no">
        <FamilyName>Sándorfi</FamilyName>
        <GivenName>György</GivenName>
      </Author>
    </Article>
  </Contents>
</Pack>
```

1 ábra: Egy EPA tartalomjegyzék XML változatának kezdete

A MATARKA adatszerkezetét és működését az én leírásaim alapján ismerte meg Zawiasa Róbert. A HUMANUS adatbázis a Bodza Bt. terméke, ezt külön nem

kellett megismernie. Szerencsére a HUMANUS teljesen be tudja fogadni mind a MATARKA, mind az EPA leírásait. A MATARKA MySQL adatbázisát közvetlen lekérdezés után konvertálja az EHM, nincs közbeépített interfész (XML vagy OAI). A közvetlen adatkapcsolat új konverziós eljárás kidolgozását igényelte, de nagyon megérte létrehozni, mert így kompromisszumok nélkül a lehető legteljesebb és leggyorsabb megoldást sikerült megvalósítani a MATARKA → EHM adatáramlás számára.

A HUMANUS fejlesztése idején már kialakított generikus cikk-kezelési eljárásoknak köszönhetően a funkcionalitás (keresés, rendezés, csoportosítás, megjelenítés, letöltés stb.) nagy része megvolt, de ezeket persze számos ponton többé-kevésbé finomítani kellett.

Igazán új EHM-fejlesztés a duplumszűrés mechanizmusa, amely más szempontból is különleges: nem előre az adatbetöltés idején történik, mint például a MOKKA esetében, hanem a találati halmaz lekeresésekor. Ennek számos előnye létezik és még messze nincs kihasználva: pl. az eredeti rekordok változatlan formában megmaradnak a rendszerben, nem „torzulnak el” az összevonás folytán, így később olyan keresések, elemzések is végezhetőek, amelyek az összevonó MPLLA-típusú rendszerekben nem.

A kereső szolgáltatást RSS támogatás (legfrissebb tételek) egészíti ki, illetve két nem webes felület: nyilvános Z39.50 és kérésre engedélyezett OAI szolgáltatás.

Egyéb munkálatok az adatbázisok fejlesztésében, bővítésében

MATARKA – A TÁMOP-projekt támogatásával az adatbázis több mint 600 000 ezer rekorddal bővült, 277 új folyóirat feldolgozása kezdődött el, illetve a meglévő folyóiratok retrospektív feldolgozására került sor. Minden folyóirat metaadatai közé bekerült az OSZK Nektár katalógusában alkalmazott Cool URI és a fűzet számozás is többféle formában (megfelelő mezőkben kiírva: 3. évf. 2. sz., vagy csak számjeggyel: 3, 2) érhető el.

EPA – A projektben az adatbázis tartalmi bővítésének támogatása is szerepelt. Az adatbeviteli munkákat különböző kiadványlisták alapján végezték az EPA munkatársai. Az adatfeldolgozásban hét fő vett részt 2010 júliusától 2011. június végéig. A támogatásnak köszönhetően, 99 672 tétellel bővült a teljes szöveggel elérhető, nyilvántartott cikkek száma.

HUMANUS – A Humántudományi Bibliográfiai Osztály munkatársai a MATARKA és a HUMANUS meglévő állományait a HUMANUS gyűjtőköre szerinti megfeleltették, vagyis a hiányokat, valamint a többszörös feldolgozásokat számba vették, a hiányokat felszámolták, a MATARKA állományából érkező rekordokat a HUMANUS-on belüli honosították, vagyis tartalmilag mélyebben feltárták, és olyan, a HUMANUS gyűjtőköréhez tartozó kiadványokat dolgoztak fel, amelyek mindeddig sem a MATARKA-ban, sem a HUMANUS-ban nem voltak analitikusan feldolgozva.

Az EHM jellemzői

Az EHM közös kereső három keresési lehetőséget kínál. Az *egyszerű keresés*-nél csak egy kereső mező jelenik meg (mint a Google-nél), és a beírt kereső kifejezést a rekord minden adatában (szerző, cím stb.) lekérdezi a program. Az *összetett keresés*-nél mód van a kereső index kiválasztására és összetett, vagyis több indexet is lekérdező, kombinált keresések hajthatók végre. A *szakértői keresés* szakembereknek (pl. könyvtárosoknak) való. A szakértői keresés ellen szól az, hogy bizonyos indexek csak a HUMANUS-ban léteznek (pl. tárgyszó), tehát ha tárgyszavakban keres a felhasználó, akkor csak HUMANUS-találatokat fog kapni. Vagyis valójában nincs szüksége a közös kereső használatára, maradhatna a HUMANUS-ban.

A közös (duplum, triplum) rekordokat kétféleképpen kezeli a szoftver. Ha sikerült a duplumot azonosítani, akkor csak a legtöbb adattartalommal rendelkező rekord jelenik meg, ha ez nem sikerült, akkor mind a két vagy három rekord listázása megtörténik. A rekordok mellett a képernyő jobb oldalán mindig látszik az, hogy az adott rekord melyik adatbázisból származik és a felhasználó rögtön továbbléphet az eredeti adatbázisba is, ha kíván.

A 2. ábrán látható, hogyan jelöli meg az EHM a forrást. Az e+h+m, e+m, h+m elnevezések értelemszerűen arra utalnak, hogy több adatbázisban is előfordul az adott rekord, a *humanus*, *matarka*, *epa* jelölés alkalmazása esetén a rekord csak egy adatbázisban szerepel.

A találati listák több szempont szerint rendezhetők: számozás, szerző, cím, dátum, dokumentumtípus, forrás. A találati listákból a kívánt tételek kiválaszthatók és a kért formátumban letölthetők.



2. ábra

Letöltés [X]

Letöltendő:
5 kijelölt tétel

Formátum:

- HTML PDF PostScript dokumentum
- szöveges MARC UTF-8 kódolással
- ISO-2709 MARC UTF-8 kódolással
- ISO-2709 MARC ANSEL kódolással
- RIS RefWorks bibliográfia
- A feldolgozandó forrásdokumentumok jegyzéke [HTML]

Levélben?
e-mail cím:

3. ábra: Letöltési párbeszédablak

A legfontosabb statisztikai adatok 2011 júniusában, a projekt befejeződéskor:

EPA:	218 455 rekord
HUMANUS:	376 606 rekord
MATARKA:	1 600 158 rekord
MKszle:	53 214 rekord

Összesen: 2 248 433 rekord

Néhány példa a keresésre

Az EHM közös kereső hasznossága különösen azokon a területeken érzékelhető, amelyeket mind a három vagy a háromból legalább két adatbázis részletesen dolgoz fel. Ilyen többek között a történelem. Vegyünk tehát történelmi példákat:





<i>1. példa:</i> Bach-korszak		<i>2. példa:</i> K. Lengyel Zsolt	
MATARKA találatok száma:	32	MATARKA találatok száma:	24
HUMANUS találatok száma:	17	HUMANUS találatok száma:	22
EPA cikkszinten önállóan nem kérdezhető le.		EPA cikkszinten önállóan nem kérdezhető le.	
EHM találatok száma:	48	EHM találatok száma:	47

Az EHM találatok megoszlása:

e+m	8
e+h+m	2
h+m	2
epa	4
matarka	19
humanus	13

Az EHM találatok megoszlása:

e+m	5
e+h+m	1
epa	8
matarka	12
humanus	21

34. Hungarologie und Ungarn-Bild in Deutschland : politische, methodische und organisatorische Probleme nach 1990 / Lengyel K. Zsolt In: Das Ungarnbild in Deutschland und das Deutschlandbild in Ungarn : Materialien des wissenschaftlichen Symposiums am 26. und 27. Mai 1995 in Hamburg / hrsg. von Holger Fischer. – München : Südosteurop.-Ges., 1996, p. 75-95.	
35. Thomas von Bogyays Hungarologie im Exil 1945-1994 / Lengyel K. Zsolt In: Wissenschaftsbeziehungen und ihr Beitrag zur Modernisierung : das deutsch-ungarische Beispiel / hrsg. von Holger Fischer unter red. Mitarb. von Mirja Juelich. – München : Oldenbourg, 2005, p. 483-565.	
36. Az emelkedettség szószólója : búcsú Szabó T. Ádámától / K. Lengyel Zsolt In: Európai utas. – 6. (1996) 1., p. 14-18.	
37. 45 éves a Müncheni Magyar Intézet / K. Lengyel Zsolt In: Európai utas. – 19. (2008) 1., p. 47-52.	

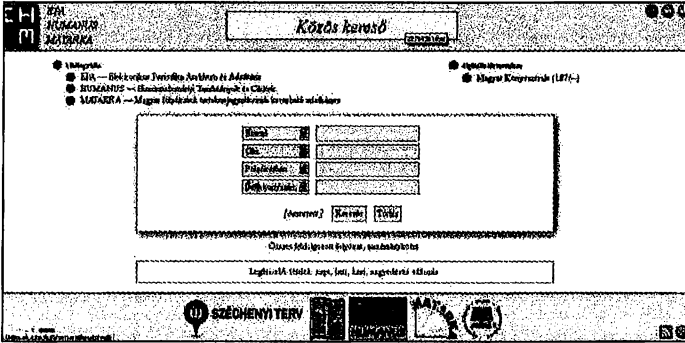
4. ábra: Találati lista részlete

A felület

Az EHM-felület nyitó képernyőjét és az EHM logóját a Miskolci Egyetem könyvtárának informatikus könyvtárosa, Kiss Andrea tervezte meg. A keresőfelület, a találati listák felépítése hasonlít a HUMANUS webes megjelenéséhez: a találatok ugyanúgy több szempontból rendezhetők, csoportosíthatók, a találati listák többféle formátumban letölthetők. Legnagyobb különbség az, hogy a képernyő nincs kettéosztva.



5. ábra: Az EHM logója



6. ábra: Az EHM webes kezdőlapja

Összefoglaló

Az EHM közös kereső mind a három szolgáltatáshoz képest óriási továbblépést jelent, ami a keresést illeti, hiszen a három adatbázist egyszerre kérdezi le, a felhasználónak elég egy felületet használnia. Ugyanakkor természetesen az egyes adatbázisok önmagukban olyan, egyéb szolgáltatásokkal is rendelkeznek, amelyet az EHM természetesen nem kínálhat: a tartalomjegyzékek böngészése az EPA-nál és a MATARKA-nál jobban megoldott, cikkmásolatküldés csak a MATARKA-nál létezik, a HUMANUS-nál finomabb a rekordkapcsolatok megjelenítése.

Az EHM közös keresőt egyelőre kevesen ismerik, ezért e beszámolóciikk szerzőjének fő célja az, hogy a könyvtáros közösség a szolgáltatásról tudomást szerezzen, a jövőben eredményesen használja. Bízom benne, hogy újabb pályázatok hozzásegítenek bennünket ahhoz, hogy a három rendszer még komolyabb színvonalú további integrációja is megvalósulhasson.

Burmeister Erzsébet