TMT 49. évf. 2002. 5. sz.

Sándor Ákos — Hegyi Ádám
SZTE Egyetemi Könyvtár

Folyóirat indexelése Zebrával

Nagy mennyiségű szöveg tartalmi feltárása az XML megjelenésével gyorsabbá és hatékonyabbá vált. Ennek lehetőségeit kihasználva valósult meg egy irodalmi folyóirat indexelése Szegeden, az Egyetemi Könyvtárban. Az alábbiakban a kivitelezés lépéseit ismertetjük röviden.

A kitűzött cél

A folyóiratok tartalmi feltárása a hagyományos könyvtári munkafolyamat egyik nehéz feladata. A számítástechnika elterjedésével olyan lehetőségek is előtérbe kerültek, amelyek az egyes munkafolyamatok automatizálásán túl lehetővé tették a bonyolultabb tartalmi összefüggéseken alapuló visszakeresést. Gondoljunk csak arra, hogy egyes nyomdai szövegformázások gyakran szemantikai jelentéssel bírnak, amelyek feltárása nehéz munka volt. Most viszont lehetőségünk nyílt arra, hogy akár tipográfiai megjelenítésre való visszakeresést is megvalósítsunk. Közismert, hogy nagy mennyiségű szöveg tartalmi feltárása cédulázással hosszú és bonyolult feladat. Periodikumok esetében ezért a tartalmi feltárás szintje egy-egy tanulmány, cikk leírására korlátozódik. Digitalizált formában viszont lehetőség adódik arra, hogy összetettebb tartalmi szempontoknak megfelelő keresést hajtsunk végre. Ennek alapját egy teljes szövegű adatbázis-kezelő adhatja.

A *Szegedi Egyetemi Könyvtárban* elkészült a *Széphalom* című folyóirat 1927–1929 között megjelent számainak digitalizált változata. A megvalósítás keretét a *Nemzeti Kulturális Alapprogram*tól nyert pályázat biztosította. A kivitelezés több lépésből állt, amelyről *Bakonyi Géza* részben már beszámolt a Networkshop 2000 című konferencián.*

Adatbázis készítése

Az elmúlt egy év alatt elkészült a folyóirat teljes szövegű adatbázisa.** Célunk a kivitelezés során az volt, hogy weben keresztül legyen elérhető a Széphalom című folyóirat teljes szövege, és az ebben való keresését is tegyük lehetővé. Rendelkezésünkre álltak már az 1927–1929 közötti számok XML-ben elkészített változatai *(1. ábra)*, illetve minden egyes oldal PDF-ben tárolt formában.

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- edited with XML Spy v3.5 (http://www.xmlspy.com)
    by Geza Bakonyi (SZTE) -->
<!--DTD generated by XML Spy v3.5
    (http://www.xmlspy.coru) -->
<!ELEMENT cim (#PCDATA)>
<!ATLIST cim
        type CDATA #REQUIRED
        >
<!ELEMENT csillag (#PCDATA)>
<!ELEMENT foszoveg (italic)>
<!ELEMENT italic (#PCDATA)>
<!ELEMENT szephalom (cim, szoveg)>
<!ELEMENT szoveg (foszoveg) | csillag)+>
```

1. ábra **A *Széphalom* XML-részlete**

A megvalósítás során több problémát kellett megoldanunk. Első lépésben az XML szövegek adatbázisba építését kellett megoldani, majd a visszakereshetőséget biztosítani. Miután ez elkészült, a weben való megjelenítés problémája merült fel.

Az indexelés

Az adatbázis építéséhez szükségünk volt egy teljes szövegű adatbázis-kezelő szoftverre. Választásunk az *Index Data* cég *Zebra* nevű szoftverére *(http://www.indexdata.dk)* esett. Ez a szoftver nonprofit szervezetek számára, mint amilyen az Egyetemi Könyvtár, ingyenes. A választás azért esett

*BAKONYI Géza: Tartalomszolgáltatás – egy folyóirat digitális feldolgozása. Előadás, Networkshop 2000. *http://nws.iif.hu/NwScd/docs/eloadas/29/index.htm*

**A folyóirat jelenleg csak részben érhető el a *http://www.bibl.u-szeged.hu/szep/index.htm* címen.

rá, mert képes XML elemek (tagek) indexelésére. Ezenkívül, mivel eredendően könyvtári használatra készült, képes a *Z39.50* szabványt, valamint a GILS metaadatkészletet is kezelni. Egyik hátránya viszont, hogy csak egy adatbázist lehet vele kezelni. Ha szükségessé válik több folyóirat szöveges adatbázisba vitele, meg kell vásárolni a Zebra üzleti változatát.

A Zebra adatbázis-kezelőhöz tartozik az úgynevezett *YAZ-kliens*, amelyre a Zebra telepítésekor és az adatbázisban való keresésekor van szükség.

Az XML-ben készült szövegfájlok csak szemantikai szempontok alapján készült elemeket tartalmaztak, amelyekre a visszakeresést a Zebrával oldottuk meg. A Zebrában megtalálható egyik attribútumhalmazt kibővítettük azokkal az XML elemekkel, amelyekre az indexelést meg akartuk valósítani: szerző, cím, kiemelt szöveg, vers, versszak, versszor, jegyzet, lábjegyzet, dátum, főszöveg.

A Zebrában ezáltal kereshetővé váltak az XML elemekkel tárolt szövegrészek, de ezek a weben nem jeleníthetők meg. A megjelenítéshez szükséges a YAZ-kliens és egy PERL script, amelyek a webes keresést lehetővé teszik.

Webes felület készítése

A legfontosabb annak a problémának a megoldása volt, hogyan lehet XML fájlokat az elterjedt böngészők által egyszerűen és gyorsan megjeleníteni. A Zebra adatbázis-kezelő ugyan képes arra, hogy az XML elemeket indexelje, és keressen is bennük, de a találatokat bonyolult, nehezen átlátható eredménylistában jeleníti meg, ugyanis az XML elemek közötti találatot úgy mutatja meg, hogy a teljes XML struktúrát is a találatok közé sorolja. Az átláthatóság érdekében ezért kellett egy *PERL script*et írni, amely a HTML-be alakítást végzi el. Az adatbázisban való keresés tehát több összetett lépésből áll.

Az adatbázisban való keresés egy HTML űrlap kitöltésével történik. Itt lehetőség van Boole-operátorok használatára. Azokra az XML elemekre lehet keresni, amelyeket a Zebrában indexeltünk (cím, verssor, verscím stb.), és a keresés az operátorok által egyszerre több mezőre is megvalósítható. A keresőkérdés kérdezése eleve csonkolva történik. A találatok megjelenítésekor a szöveg teljes egészében betöltődik, amelyben egy fejlécben a kiadási, terjedelmi adatokat kiemeltük. Innen elérhető az adott folyóiratoldal PDF formátumú verziója is. A találatok élénk színű kiemeléssel vannak jelölve. Több találat esetén egyszerűen görgetni lehet az adott szöveget. A különböző szemantikai jelentésű találatokat eltérő HTML formázási elemekkel jelenítjük meg. Így például a szerzőt 14 pontos, félkövér, dőlt karakterekkel.

A találati oldal szerkezete *layer*ekkel van megoldva. Ennek megfelelően a megjelenítendő szövegek közötti görgetés úgy valósul meg, hogy a láthatóvá tett layer rész folyamatosan csúszik a szöveg felett. Ezáltal a szöveg görgethető. Az egyszerű böngészéshez ezért ajánlott minél magasabb verziószámú böngészőt használni.

Az űrlapon elküldött kérdéseket a webszerver kapja meg, amelyen egy PERL Script értelmezi őket. Az átalakított kérdéseket a Z39.50 szabványnak megfelelő keresőkérdés formájában kapja meg az YAZ-kliens, amely továbbítja azt a Zebrának. A Zebra elvégzi a keresést, és visszaküldi a keresőkérdést egy Z39.50-es szabvány szerinti eredményhalmazban a YAZ számára. E műveletek közben a Zebra *Z-szerver*ként, a YAZ *Z-kliens*ként működik. A találathalmazban a teljes XML fájl benne van. Ennek élvezhető olvasását a PERL script valósítja meg azzal, hogy a stíluslapokon definiált formázásokat HTML elemekkel helyettesíti. A találati halmaz ezzel böngészőprogramokkal megjeleníthetővé vált.

Az adatbázis működését a könyvtár webszervere biztosítja, amelyen a Zebra és a „Széphalom" adatbázis található.

A megvalósítás problémái

A kivitelezés során több kisebb problémával is találkoztunk. Egyik ilyen volt, hogy az XML fájlok szerkesztésük során *UNICODE* karakterkészletben készültek. Amikor ezeket Linux alatt néztük, több értelmezhetetlen karaktert is láttunk egy-egy fájlban. Ezek eltüntetésére szükséges volt az XML dokumentumok konvertálása sima szöveges fájlba (.xml.txt). Így csak olyan karakterek maradtak, amelyek kezelhetővé váltak a Linux számára is. Gondot jelentett az is, hogy hogyan lehet definiálni a Zebra adott attribútumhalmazában olyan XML elemeket, amelyeket mi akarunk felvenni. Ennek a megoldása lett a Zebrában található egyik attribútumhalmaz kibővítése.

Összegzés

Jelenleg a Széphalom című folyóirat 1927–1929 közötti példányainak teljes szövege elérhető az interneten, amelyekben megadott szempontok alapján lehet keresni, bár hozzá kell tenni, hogy az indexelt állomány feltöltése még nem teljes egészében történt meg. Ezzel lehetővé vált több, az irodalomtörténet szempontjából érdekes kérdés egyszerű megválaszolása is, mint például hogy hányszor, milyen szövegkörnyezetben, milyen értelemben használta egy-egy költő – mondjuk – a forradalom kifejezést, hiszen akár egy műben, akár egy évfolyamban lehetővé váltak az ilyen típusú keresések.

Beérkezett: 2001. XI. 14-én.

Rendezvénynaptár

Könyvtárak és egyesületek a változó világban: új technológiák és együttműködési formák CRIMEA 2002, 9. nemzetközi konferencia
Sudak (Ukrajna), 2002. június 8–16.
Információ:
Tel.: +7 095 924-9458, +7 095 923-9998
Fax: +7 095 921-9862, +7 095 925-9750
E-mail: CRIMEA2002@gpntb.ru
URL: http://www.iliac.org/crimea2002
http://www.gpntb.ru/win/inter-events/crimea2002

Nemzetközi konferencia az élethosszig tartó tanulásról
Yeppoon (Ausztrália), 2002. június 16–19.
Szervező: Lifelong Learning
Conference Secretariat
Central Queensland University Library
CQ Mail Centre, Rockhampton
Queensland, Australia
Tel.: +61 7 4923 2198 • Fax: +61 7 4930 6436
E-mail: lifelong-learning-conference@cqu.edu.au
URL: http://www.library.cqu.edu.au/conference

Az IFLA 68. konferenciája
Glasgow, 2002. augusztus 18–24.
Szervező: Mrs. Bodil Wöhnert
Centralbiblioteket i Esbjerg
Nřrregade 19
DK 6700 Esbjerg
Tel.: +45 76 16 19 61 • Fax: +45 76 16 20 03
E-mail: bow@esbjergkommune.dk
URL: http://www.ifla.org

Informatika a felsőoktatásban 2002. konferencia
Debrecen, 2002. augusztus 28–30.
Szervező: Karácsony Gyöngyi informatikus könyvtáros
Debreceni Egyetem Egyetemi és Nemzeti Könyvtár
Kenézy Könyvtára
4012 Debrecen Pf. 18
Nagyerdei krt. 98.
Tel.: (52) 489-400/4934 • Tel./fax: (52) 413-847
E-mail: gyongyi@clib.dote.hu
URL: http://www.date.hu/if2002/