Bakonyi Géza
József Attila Tudományegyetem Egyetemi Könyvtár

Információrögzítés és visszakeresés heterogén környezetben

A posztmodern, információs civilizáció megteremtésének jelenlegi szakaszában a társadalmat alkotó emberek, ezek érdekszövetségei olyan kulturális mikrokozmoszt építettek maguk köré, amely inkább hasonlít csapdákkal teli információs labirintushoz, mintsem egy védő-óvó otthonhoz. Kicsiben ugyanezzel néznek szembe a kultúra különböző médiumokon rögzített termékeit őrző és rendszerező könyvtárak is: heterogén, multimédia környezetben kellene rendet teremteniük, olvasóikat eligazítaniuk. Ebben a helyzetben a hagyományos könyvtári eszközök mellett sokszor a modern információs technológia eszközei is csődöt mondanak. Sőt, inkább csak újabb csapdákat, megtévesztő oldaljáratokat teremtenek az információs labirintusban. Vajon meg tudja-e teremteni ez a társadalom, benne a modern könyvtár azokat az eszközöket, amelyek Ariadné fonalaként segíthetnek megtalálni a helyes utat?

Olvasók és felhasználók

A Biblioteca Medicea Lorenziana (a világ első nyilvános könyvtára) olvasópadjainak elrendezése világosan tükrözi azt az információ-visszakeresési technológiát, amelyet a könyvtárak az évszázadok során kifejlesztettek, hogy megtalálhatóvá tegyék az olvasó által óhajtott művet az információs labirintusban. Ez a technológia (azaz az egyes padsorok oldalán kilógatott tábla a tárgykör és a tárgykörbe tartozó művek megjelölésével) elegendő volt a pár száz, később is csak ezernyi kötetet őrző könyvtár számára.

A könyvek megőrzésére vállalkozó könyvtár azonban mennyiségileg és minőségileg is új életszakaszba lépett: mára már nem egyszerűen a nyomtatott információ őrzését és megtalálhatóságát kell biztosítania. A mai könyvtár elsősorban és általánosságban a legkülönfélébb médiumokon megjelenő információforrásokat gyűjti, és az ezeken található információkat igyekszik hatékonyan rendszerezni és szolgáltatni. A két feladat ellátása között óriási különbség van mind a szervezeti felépítés, mind az alkalmazott szakemberek felkészültsége, mind az alkalmazott technológia tekintetében.

Ebben az összefoglalónak, problémafelvetőnek szánt tanulmányban elsősorban az információ-visszakeresési technológia új lehetőségeiről lesz szó. Hangsúlyozni szeretném, hogy nem egyszerűen a bibliográfiai leírásokhoz kötődő információkeresési technológiát vizsgálom, hanem a könyvtárak által feldolgozott különféle – akár hagyományos, akár digitális, elektronikus médiumon megjelenő – gyűjteményekre vonatkozó, illetve az e gyűjteményekben található információs visszakereshetőségi lehetőségeket általában.

A könyvtárakban a monográfiák, sorozatok és periodikumok bibliográfiai és tematikai (szakjelzetek, deszkriptorok, tárgyszavak stb.) feldolgozása mellett egyre hangsúlyosabb problémát jelent az eddig csak egyéb kategóriába sorolt más információforrások feldolgozása. Először a mikrofilm, mikrofilmlap és a videofilm tette próbára a könyvtári szakemberek felkészültségét, később a számítógépes adatbázisok (akár CD-ROM, akár online adatbázisok; akár faktográfiai, akár teljes szövegű adatbázisokról is legyen szó), majd az Interneten keresztül elérhető hipertext jellegű, különféle típusú objektumokat tartalmazó adatbázisok, archívumok megjelenése. Ezzel párhuzamosan az olvasókból felhasználók lettek, s ez nem egyszerűen csak megfogalmazás kérdése, inkább annak a felhasználói magatartásnak a megváltozását írja le, amikor a könyvtárban megjelenő egyén már nem egyszerűen csak egy dokumentum meglétét vagy meg nem létét kéri számon a könyvtárostól, hanem általában az általa sokszor kérdésként meg sem fogalmazott igényt kielégítő választ. A régi olvasót egyszerűen csak az érdekelte, hogy megtalálható-e *Plinius* műve a botanikáról, a mait viszont az érdekli, hogy mit olvashat a botanika és az irodalom kapcsolatáról, ha csak magyarul és angolul tud.

A továbbiakban röviden tekintsük át, hogy ezek a jelenségek miképpen mutatkoznak meg a könyvtári szolgáltatások egyes szintjein.

Keresés számítógépes könyvtári katalógusokban

Most nem annyira a számítógépes könyvtári katalógusok visszakereshetőségi problémáiról lesz szó, mégis a jelzett elméleti kérdések éppen ezeken az olvasói felületeken jelentkeznek a leglátványosabban, így először ezekről ejtünk szót. A könyvtári rendszerek általában kétféle megoldási javaslattal állnak elő a keresési ellentmondások feloldására.

Technikai megoldások

Az olvasóból felhasználóvá avanzsált egyénnek elsődlegesen azzal a problémával kell megküzdenie, amelyet az új technológia megjelenése generál. Eddig elég volt, ha tudott olvasni, írni, és tájékozott volt a saját szakterületén; mostanra viszont meg kell tanulnia a mentális protézisek használatát – azaz a számítógépes input és output kezelését, s az ebből következő viselkedésbeli változásokat is el kell sajátítania. (Gondoljunk csak arra, hogy mennyire más koncentrálást, lényegfelismerési eljárást igényel csak az, hogy a bibliográfiai leírásokat nem katalóguscédulán, katalógusok fiókjaiban kell olvasnunk, hanem színes vagy feketefehér képernyőn.)

A könyvtár nem zárkózhat el az elől, hogy segítse ebben az átállásban az olvasóját – hiszen akarva-akaratlan, éppen ő generálta ezt a problémát. A szakirodalom elég pontosan össze is foglalja az ebből eredő problémákat:

- az indexekben történő direkt és indirekt keresések nyilvánvaló felajánlása (a kulcsszavas és a böngésző keresés együttes biztosítása);
- az új technológiából következő új szókincs alkalmazása és értelmezése (nem biztos, hogy mindenki azonnal érti, mit jelent a találati halmaz, keresőkifejezés, nincs megfelelő rekord stb. fogalma);
- a direkt keresés eredménytelensége esetén értelmes hibaüzenet biztosítása (a „no records found" üzenet különböző nyelvi változatai legfeljebb azt eredményezik, hogy az olvasó negatív véleménye az adott könyvtárról, esetleg önmagáról tovább erősödik); fel kell hívni az olvasó figyelmét, hogy talán elgépelt valamit, nem vette figyelembe az ékezetes betűk használatát stb.;
- ugyanakkor ellenkező esetben, vagyis ha a találati halmaz túlságosan nagy, segíteni kell a keresési kérdés túl általános megfogalmazásának konkrétabbá tételében (felhívni a figyelmet a keresőkifejezés szűkítésének a lehetőségére és ennek demonstrálása);
- a túlságosan bonyolult, illetve túlságosan egyszerű „help" képernyők kerülése (nem szerencsés, ha egy konkrét segítségkérésre nagyon általános választ adunk, ahogy a fordítottja sem – *Karinthy*: „benyúltam a zsebembe és felsikoltottam: egy kéz volt ott" mintájára előállított „a jobb oldali kurzorral menj le a negyedik sorba, ott nyomd meg az F1 billentyűt, majd gépelj be egy x jelet, és utána nyomd meg az Enter vezérlőbillentyűt" típusú help szövegekre gondolok);
- az összetett keresések esetében a keresőkifejezés megfogalmazásának a biztosítása a Boole- és a közelségi operátorok pontos ismeretének elsajátítása nélkül (nem feltétlenül szükséges, hogy a felhasználó tökéletesen birtokolja a logikai kifejezések kifinomult használatának tudását abban az esetben is, ha *Dosztojevszkij* Bűn és bűnhődés című könyvének egy magyar nyelvű fordítását akarja olvasni);
- felhívni a figyelmet a keresőkifejezések pontos megfogalmazása, illetve balról vagy jobbról való csonkolása fogalmának a jelentőségére;
- biztosítani, hogy az egyszer megfogalmazott keresőkifejezés a továbbiakban is felhasználható legyen („history" és „set" típusú fájlok);
- a megjelenítési (beleértve a mentést és a nyomtatást is) formátumok egyértelmű kezelése, azaz a felhasználó ne kényszerüljön arra, hogy a „párhuzamos cím" és „főcím", az „egyedi szerző" és a „testületi szerző" közötti különbségek rejtelmeibe elmerüljön;
- kihasználni a grafikus felhasználói felület nyújtotta előnyöket (drag 'd drop, cut and paste stb.) anélkül, hogy a felhasználót az operációs rendszer szakértőjévé képeznénk ki.

A felhasználók oktatása

A könyvtárak, illetve a könyvtári informatikai rendszerek általában két úton igyekeznek megközelíteni a problémát, egyik esetben sem túlságos hatékonysággal. Egyrészt állandóan meg-megújuló kísérleteket tesznek az olvasók továbbképzésére, másrészt a felhasználói felületeket igyekeznek egységesíteni (amit sokan összetévesztenek a HTML-alapú help képernyők elburjánzásával).

Természetesen egyik megoldás sem lebecsülhető, s nem is mellőzhető. A felhasználók – az olvasók újabb és újabb generációjának könyvtárba tódulása (s ez nem csak az egyetemi könyvtárakra jellemző, hiszen a folyamatosan növekvő hallgatói létszám kiszolgálására lényegében nem épült új könyvtár, a régiek pedig elavult könyvtári térkialakí-

tással épültek, ha egyáltalán könyvtárnak építették) által sokszor reménytelennek látszó – oktatása elsősorban azt a célt szolgálja, hogy tudatosítsa bennük a modern információs technológiában rejlő lehetőségeket, megmutassa az ehhez csatlakozó módszerek hatékonyságát. A helyzetet nehezíti, hogy a felhasználók kényelemből, rossz beidegződésből nem a megfelelő keresési módszereket részesítik előnyben (egyszerűsített kulcsszavakban gondolkodnak, illetve ellenkezőleg: a címre való keresésnél gondosan beírják a cím összes szavát a releváns fogalmak megjelölése helyett stb.).

Mindez nem teszi feleslegessé a tájékoztató munkát, sőt annak megújulását teszi szükségessé.

Programok

Az EU több könyvtári programot is indított, amelynek középpontjában éppen az olvasók, könyvtárosok továbbképzésére vonatkozó oktatási módszer áll. Ilyen pl. az EDUCATE program (EnD-User Courses in information Access throught Communication Technology), amely – hasonlóan a többihez – a WWW nyújtotta lehetőségek kiaknázásával, a távoktatásban szerzett tapasztalatok kihasználásával az önképzésre épít. A HyperLib ennek a módszernek továbbfejlesztett változata. A program indítói érezték, hogy a HTML formátum nyújtotta lehetőségek erősen korlátozottak az egyszerűbb hipertext dokumentumokat leíró nyelv formális szabályai által, ezért azt tűzték ki célul, hogy az SGML (Standard Generalised Markup Language – ennek végtelenül egyszerű altípusa a HTML) szabványra építve a könyvtári kalauzok általános logikai leírását, struktúráját fogalmazzák meg. A leírás a WWW böngészők HTML-beállítottsága következtében nem terjedt el (SGML szöveg megjelenítésére jelenleg is csak két, könnyebben elérhető eszköz áll rendelkezésünkre: a SoftQuad Panorama és a HotMetal), de semmiképpen sem tanulságok nélkül való: tanulmányozása sok módszertani buktatón átvezethet minket.

Az „online help" minden hatékonysága mellett szükséges az olvasók személyes kontaktusra épülő továbbképzése. A könyvtárosok, túlterheltségükre hivatkozva, általában szívesen használnak különböző feliratokat (a szövegszerkesztők, nyomtatók sablon ikonjai csak támogatják ezt), de ezeknek a feliratoknak és falragaszoknak a túlburjánzása inkább csak elijeszti az olvasókat, s nem segíti őket. A személyes kontaktus csak akkor lenne hatékony, ha kis csoportos foglalkozás keretében valósulna meg, de erre valóban kevés lehetőség van. Ugyanakkor nem szabad lemondani róla: célszerű napi, heti állandó időpontokat megadni, és a foglalkozásokat kevés számú érdeklődő esetén is megtartani. A másik lehetőség az egyes olvasócsoportokat megcélzó kurzusok meghirdetése (egyetemi könyvtárakban bizonyos szakok, évfolyamok; közkönyvtárakban iskolai osztályok stb.). A nagy létszámú csoportoknak tartott tájékoztatásról sem szabad lemondani: a beiratkozásnál jelentkező nagyobb csoportok esetében alkalmazhatjuk ezt (egyetemi könyvtárakban az akadémiai év kezdetén).

Szabványok a tartalmi feldolgozásban

A számítógépes könyvtári katalógusok gyors elterjedése és fejlődése viszonylag rövid időn belül felszínre hozta azt az igényt, hogy a könyvtári dokumentumok feldolgozását illetően kidolgozzák azokat a szabványokat, amelyek alkalmazásával lehetséges az egyszer létrehozott rekordok átvétele (shared cataloguing), és egységes szempontok szerinti visszakereshetősége (union catalogue). A MARC formátum, az ALA karakterábrázolás és az LC tárgyszórendszer, szakjelzetek egyszerűnek tűnő, de hatékony megoldást nyújtottak mind a könyvtárosoknak, mind az olvasóknak. Egységes katalogizálási szabályzat nélkül persze ez csak látszólagos, hiszen egyrészt a különböző könyvtári hagyományok és szokások összehangolása, másrészt az ezekből következő kompromisszumok algoritmizálása már egyáltalán nem volt olyan gyors folyamat. A számítógépes könyvtári rendszerek nemzetközi elterjedése tovább bonyolította a helyzetet: az egyes nemzeti könyvtárak általában ragaszkodtak a saját nemzeti szabványaiknak megfelelő, a nemzeti karaktereket is megőrző számítógépes feldolgozáshoz. Ennek következtében az egységesnek szánt szabványok variációkra hullottak szét. Létrejött a USMARC, a CANMARC, a UKMARC, a MARBI és társaik, s természetesen a tárgyszó-, szakjelzetrendszerek sem jártak jobban.

A kelet-európai országok könyvtárainak gépesítése csak még zavarosabbá tette ezt a helyzetet: egy országon belül akár többféle szabványt is használnak, attól függően, hogy az adott könyvtár milyen országból származó rendszert vett meg.

Hasonló a helyzet Magyarországon is. A nagyobb egyetemi és megyei könyvtárakban olyan rendszereket használnak, amelyeknek a fő (sokszor az adatbázis belső) formátuma a USMARC. Ezek a könyvtárak általában ragaszkodnak is ehhez a MARC formátumhoz, hiszen az egyetemi és szakkönyvtárakban jelentős mennyiségű külföldi monográfiát dolgoznak fel, és ezek bibliográfiai leírásai általában megtalálhatók a nemzetközi hálózaton elérhető számítógépes katalógusokban, és minden nehézség (Z39.50, WebPAC megoldások) nélkül importálhatók az adott rendszerekbe. Más

rendszerek esetében a magyar szabványnak ajánlott HUNMARC adatsorának importálása nem jelent nehézséget, mivel e rendszerek belső adatbázis-formátuma vagy különben sem egyezik a MARC szegmentálással, vagy kellő részletességgel szegmentált, hogy bármilyen csereformátumhoz tudjon alkalmazkodni. A HUNMARC-alapú adatsort természetesen mindegyik képes kezelni, bár ez mindig több-kevesebb adatvesztességgel, adatbázis-inkonzisztenciával járhat. Különösen azokban a magyar könyvtárakban probléma ez, ahol a rendszereknek nem csak a katalogizálási és kölcsönzési modulját használják.

Jól demonstrálják ezt az állapotot a közelmúlt két nagy országos könyvtári projektjének kidolgozása és végrehajtása közben felmerült problémák.

KözElKat

A KözElKat projekt olyan egységes lekérdező felület létrehozását javasolta, amelynek segítségével a hozzá illeszkedő könyvtári katalógusok – a felhasználó számára látszólagosan – párhuzamosan lekérdezhetők. A KözElKat alapvetően nem a nemzetközi szabványként elfogadott Z39.50 adatbázis-lekérdezési szabványon alapul (bár ma már tudja ezt is), tekintettel arra, hogy az ehhez való ragaszkodás a legtöbb tagkönyvtárat kizárta volna a rendszerből. Ezek rendszerei ugyanis nem lettek volna az e szabványnak megfelelő szerver-kliens megoldásokkal megközelíthetők. Az így kialakított felület azonban az egyszerű lekérdezések (szerző, szerző/cím, kulcsszó) nagy százalékában megközelítőleg pontos eredményt ad az olvasóknak.

A különben is keresési bizonytalanságokat okozó esetekben viszont a KözElKat által adott megoldás felerősíti a mögötte álló rendszerek, könyvtári katalógusok inkonzisztenciáját, a különböző adatbázis-formátumokból, a különféle feldolgozási szabványokból, és az ezek eltérő szegmentumai (mezők, almezők, indikátorok) alapján létrehozott kulcsszavas és böngésző indexek alkalmazásából eredő eltéréseket. (Hangsúlyozom: ez a megállapítás nem a KözElKat, hanem a könyvtári szabványok nem egyértelmű alkalmazásainak a kritikája.) Ebből következik aztán, hogy olyan szerzők esetében, akiknél a név transzliterálása nem egyértelmű, vagy nem egyértelműen alkalmazott, igen eltérő eredményeket kaphatunk, ha a KözElKat felületét használva keresünk egyszerre több könyvtárban, vagy ha az adott könyvtárak katalógusát egyenként kérdezzük le.

A nevekre való keresés általában is problematikus, hiszen az egyes rendszerek behatóbb tanulmányozása alapján tisztában lehetünk azzal, hogy míg az egyik rendszerben a szerzői index kizárólag a személyi szerzők tárolására kinevezett mezők/almezők tartalmát indexeli, addig más rendszerek esetében az egyéb szerzőségi közlésben, sőt az utalókban található nevek is ebben az indexben kereshetők. Ennek következtében a *Reich Károly* grafikus nevére való keresés a szerző mezőben olyan találati halmazt fog eredményezni, amelyben megtalálhatók lesznek a Reich Károly szerzőségével, illetve illusztrálásával jelzett művek ugyanúgy, ahogy a róla szóló művek is. Ez még önmagában nem is lenne baj. A baj az, hogy minden egyes könyvtárnál a saját alkalmazás szerinti variációt kapjuk – tehát soha nem lehetünk biztosak abban, hogy mekkora is a tényleges találati halmaz. Megfelelően nagy adatbázisok bekapcsolása esetén (s ilyen azért szerencsére van néhány) ez persze nem jelent tényleges problémát a felhasználók számára. Szerencsés esetben minden olyan művet megtalálhat, amelyre szüksége van – legfeljebb e művek lelőhelyéről kap hiányos információt.

Súlyosabb a helyzet akkor, ha a szerző nevének transzliterálása, illetve a transzliterálás alkalmazása nem egyértelmű. Ez a helyzet a szláv nevek (*Dosztojevszkij*, Dostoevskij, Dostoyevsky, Dostoewskii stb.), a humanista szerzők (*Komensky*, Comenius, Komenius; lásd még *Prothasius de Boskowitz, Marsilius Ficinus, Galeottus Martius, Baptista Guarinus, Nicolaus Machiensis, Franciscus Maturantius, Gaspar Tribrachus* stb.), a híresebb latin auktorok eredeti latin és nemzeti hagyományokhoz kötött névváltozatainak keresőkérdésbe való illesztésénél is. A könyvtári szabványoknak éppen ilyenkor kellene hatékonyan működniük, de sajnos éppen ekkor mondanak a leglátványosabban csődöt.

MOKKA

A magyar osztott katalogizálási program (MOKKA) kidolgozása talán még több szabványosítási problémát hozott felszínre. Az ezekkel kapcsolatos viták középpontjában leginkább a HUNMARC-nak a USMARC-tól és a UNIMARC-tól eltérő mező- és almező-alkalmazása állt, hiszen ez a legtöbb tagkönyvtárat kétséges eredményű konvertálásra fogja kényszeríteni. Ennek ellenére én inkább nem ezzel a különben is sokszor körüljárt és még a közeljövőben többször körül is járandó kérdéssel foglalkozom, hanem három másikkal.

Rekordkapcsolatok

Látszólag nem nagy jelentőségű, csak technikai megoldást sürgető probléma. Valójában az egész program megbukhat rajta. A kérdés az volt: tekintettel arra, hogy a tagkönyvtárak egy része a saját

rendszerében is rekordkapcsolatok létrehozásával oldja meg a sorozatok és a többkötetes művek feldolgozását, nem lenne-e célszerű ennek a technikának az alkalmazása a központi katalógusban is. A kérdésre eddig nem igazán talált választ a MOKKA informatikai bizottsága, hiszen ez a megoldás feltételezné, hogy azok a rekordazonosítók, amelyeken a központi adatbázisban a rekordkapcsolatok létrehozása alapulna, azonosak minden egyes tagkönyvtár esetében is. Azaz a le-, illetve feltöltött, rekordkapcsolatokat tartalmazó rekordok és részrekordok kapcsolata nem jelent majd inkonzisztenciát az adatbázisokban. Mint említettem, sajnos megoldás az ilyen fokú együttműködésre jelenleg nem létezik.

Egységesített nevek

A MOKKA bibliográfiai szabványokkal foglalkozó bizottsága – teljes joggal – ragaszkodna hozzá, hogy az egységesített adatokat tartalmazó besorolási rekordok együtt mozogjanak a hozzájuk tartozó bibliográfiai rekordokkal. Azaz ha egy tagkönyvtár le- vagy feltölt egy bibliográfiai leírást, akkor azzal együtt a hozzá tartozó egységesített nevek (elsődlegesen a szerzői és a címbejegyzések) rekordjai is fel-, illetve letöltve legyenek. A probléma megint csak az, hogy mi fogja biztosítani az együvé tartozást, amikor a tagkönyvtárak eltérő számítástechnikai és informatikai elvek alapján működő rendszereket használnak. A legjobb megoldás talán az lenne, ha a tagkönyvtárak eltekintenének a bibliográfiai és a besorolási adatok együtt mozgásától. Ezzel azonban éppen az osztott katalogizálás egyik legnagyobb hasznától fosztanánk meg magunkat: lemondanánk az azonos elvek alapján létrehozott egységes besorolási alapokat tartalmazó fájlok létrehozásáról. S ezzel annak az esélyéről is, hogy az olvasók a közös lekérdezési felületek használatában valaha is százszázalékosan megbízzanak.

Tartalmi feltárás

A MOKKA-t létrehozó tagkönyvtárak természetesen nem mondhatnak le a tartalmi feltárás egységesítéséről. A szakjelzeteket illetően szembe kell nézni végre az ETO egységes használatának az elfogadásával – ugyanis a tagkönyvtárak által alkalmazott variációk végtelen száma teljesen lehetetlenné teszi, hogy a tudós vagy laikus felhasználó akár csak véletlenül is olyan szakjelzetet találjon, amelynek segítségével egy szakcsoportot le tudna kérdezni. A helyzetet nehezíti az is, hogy az ETO rengeteg olyan karaktert használ az alfogalmak jelölésére, amelyek többségének használata a legtöbb számítógépes adatbázis-kezelő esetében kérdéses, ha nem éppen tiltott használatú. Így egyes rendszerekben visszakereshetetlen a gondolatjel, a zárójel és az idézőjel. Más rendszerekben csak a kulcsszavas és a böngésző keresés kombinálásával érhetünk el legalább megközelítő eredményt.

Kliens-szerver megoldások

Az integrált könyvtári rendszerek már elég korán megtalálták a könyvtári igényeknek legmegfelelőbbnek látszó adatbázismodellt: az SQL nyelvre építő relációs adatbázis-kezelőket. Ezeknek egy része egész egyszerűen elmenti egy adattáblába a MARC adatsort, majd az egyes mezők és almezők tartalmát indexelhető adattáblákban megismétli; mások megpróbálták rugalmasabb adattábla-kiosztással ésszerűsíteni az adatbázisok kezelését.

Akármelyik megoldást is választják, a szegmentálási és az adattáblák közti relációk szülte kényszer a visszakeresésben is megmutatkozó hiányosságok megjelenéséhez vezetett. Ilyenek pl. a mezők meghatározott hosszából is eredő ismételhetőség behatároltsága az egyes adatmezők definiálásánál; a különböző adattáblák között létrehozható relációk lehetséges maximuma; a többvariációs táblák közti relációk kiváltására használt kereszttáblák. Ezek a korlátok elsősorban azért jelentenek problémát, mert eleve behatárolják a későbbi igények kielégítésének a lehetőségeit. Az igények ugyanis a számítástechnika, az információs technológia gyors fejlődése következtében gyorsan változnak; az alkalmazott technológia és a rá épülő szabványok viszont sokkal merevebbek annál, hogy ezt a változást megfelelően követni tudnák. Az IT fejlődésével cseppet sem lépést tartó, de az anyagi eszközöket birtokoló fenntartók így a könyvtári rendszereket nem tudták olyan gyorsan kicserélni, ahogy arra szükség lett volna. Az igényeket azonban nem lehetett elfojtani: így a hiányosságokat a szállítók a kliensoldal erősítésével igyekeztek elleplezni.

A problémát fokozza az is, hogy a felhasználók nagy előszeretettel alkalmaznak a strukturált adatbázisokban való visszakeresésre olyan technológiát, amely legfeljebb a teljes szövegű adatbázisokban hozhat megfelelő eredményt, azaz a kulcsszavas keresést részesítik előnyben.

A negyedik generációs könyvtári rendszerek elsősorban abban különböznek az előző generációtól, hogy a megoldás egy részét az alapvető adatbázis-funkciókat ellátó szerveroldalról áthelyezik a felhasználó kliensére. Így a felhasználó, részben jogosan, abban a hitben van, hogy az óhajtott fejlődés bekövetkezett. Részben jogosan, hangsú-

lyozom. A felhasználó meg tudja változtatni azokat a beállításokat, amelyekhez eddig nem volt joga – hiszen ez a jog eddig a szervert menedzselő „root" kizárólagos joga volt. Megmondhatja, hogy a bevitt adatokat milyen formátumban szeretné kezelni (azaz a MARC, a MARBI stb. melyik verzióját részesíti előnyben) – igaz, közben tudomásul kell vennie, hogy bizonyos (más formátumokhoz tartozó) mezők feletti ellenőrzési jogát elvesztette. Meghatározhatja a visszakeresés paramétereit: melyik mező, almező tartalmát akarja látni az indexben; milyen szabályok alapján kívánja alkalmazni a böngészést. Viszonzásképpen beletörődik, hogy ez a megjelenítés nem érinti az adatbázis tényleges szerkezetét – hiszen ez majd csak más területeken fog jelentkezni problémaként. A rendszer- és adatbázisgazda fog szenvedni tőle, s áttételesen az olvasók.

A fent leírt kliens-szerver modell persze elsősorban a Microsoft-alapú rendszerekre igaz. Emellett lehetséges olyan osztott rendszer kialakítása is, amikor a kliensek nem egymástól teljesen elszakítva működnek, hanem egy közös rendszerbe foglalva, az erőforrások ésszerű megosztásának elve alapján.

Új megoldás: MARC DTD

Egészen más megoldást sugall az a kísérlet, amely a MARC formátum SGML szabvány alapján való átírásában rejlik. Ezzel ugyanis lehetővé vált (persze jelenleg inkább csak elméleti szinten), hogy a bibliográfiai leírást ne mint mechanikus szegmentálást alkalmazzuk, hanem mint általános dokumentumleírási szabályt (DTD), amely nem korlátozza a dokumentum leírásában rejlő információt, sőt inkább felszabadítja. A MARC DTD létrehozása nagyon rokon törekvés a TEI program által létrehozott megoldásokkal, amelyek persze elsősorban az irodalomtudományi, és az ezzel a területtel kapcsolatos bibliográfiai leírásokra koncentráltak (lásd később).

Ha sikerül létrehozni azokat a szoftvereszközöket, amelyek a MARC DTD alapján képesek segíteni a könyvtáros szerkesztő munkáját, illetve az olvasó visszakeresését, akkor lehetségessé válik, hogy egy olyan értelmes dokumentumfeldolgozási eszközt kapjunk, amely nem mechanikus címleírások, hanem szemiotikai értelmezés alapján képes feldolgozni a dokumentumokat. A MARC DTD alapján működő böngésző pedig egységes elvek alapján (hiszen ugyanannak a DTD-nek az értelmezése alapján dolgozik) mintegy természetesen felkínálja és egységesíti a visszakeresési szempontokat.

Ezt a megoldást nagymértékben segítheti az újabban kidolgozott XML szabvány, amelynek használatával lényegesen leegyszerűsödhet a különben eléggé bonyolult SGML alkalmazás. Lehet, hogy a katalóguscédula új, digitális formában fog visszatérni?

Digitális könyvtárak visszakeresési problémái

A digitális könyvtárak (ebben a szövegkörnyezetben elsősorban azokat a szövegarchívumokat értjük e megnevezésen, amelyek egyrészt a – legtöbbször szépirodalmi, tudományos szövegek – hagyományos kiadása során keletkezett számítógépes változatát, másrészt e szövegek digitalizálása során keletkezett szövegváltozatokat gyűjtik egybe) megjelenése a kezdetekben egyáltalán nem jelzett visszakeresési problémákat. Egyszerű indexek, gyarapodásjegyzékek is megfelelőnek látszottak. A Magyar Elektronikus Könyvtár (MEK) még úttörőnek is számított annak a megoldásnak az alkalmazásával, amelyet MEK fejlécnek nevezünk. Ez, egy némiképp leegyszerűsített könyvtári leírást idézve, tükrözte a szöveggel kapcsolatos legfontosabb adatokat.

Ezek között az adatok között már kezdetben volt olyan, amely felvillantotta a későbbi problémákat: milyen szövegváltozat digitális megjelenéséről van szó, ki végezte a digitalizálást, milyen eszközöket használt, volt-e korrektúra stb. Alaposabb utánagondolást igényelne, hogy milyen egyéb paramétereket igényelne még a digitális szöveg leírása – a MEK-et kritizáló írások részben ezeket a kérdéseket is feszegetik.

Ugyanakkor az is világossá vált, hogy az egyre nagyobb számú dokumentumot őrző szövegarchívumok szinte kínálják a lehetőséget a teljes szövegű visszakeresésre. Csábító lehetőség a teljes magyar irodalomban visszakeresni egy szóra, egy kifejezésre, s megvizsgálni, hogy az mely íróknál, milyen művekben fordul elő. A teljes szövegű visszakeresés látszólag teljesebb lehetőséget kínál, mint a számítógépes katalógusok összetett kulcsszavas kereséseinél alkalmazott Boole-operátorok. Itt eléggé kifinomult csonkolási lehetőségekkel, közelségi operátorokkal is dolgozhatunk – ezek azonban valójában csak mechanikus keresési technikák alkalmazását jelentik, a tényleges szemantikai tartalomhoz nem sok közük van.

Szépirodalmi szövegek esetében ez bizony eléggé szegényes eszköz, s ami azt illeti, már a szociológiai, statisztikai adatfeldolgozások is rafináltabb tartalomelemzési eljárásokat dolgoztak ki. Ezeknek a tartalomelemzési eljárásoknak ma már romantikusnak ható előjátéka volt a világháború idején alkalmazott módszer: az újságok és a rádióadások hírei óriási szöveggyűjteményének a

feldolgozása a tényleges csapatmozgások, a fronthelyzet előrejelzése érdekében.

A mai számítógépes szövegarchívumok a törekvéseket sem hasznosítják az akkor kidolgozott módszereknek, bár egyes törekvések azért tetten érhetők. A Progetto 200 (az olasz költészet kezdeteit feldolgozó olasz szövegarchívum) például lehetővé teszi, hogy a szövegekben rímpárokra, sorvégződésekre is vissza lehessen keresni. Ez még nem különösebben kifinomult keresési lehetőség, de már eleve több, mint amit az egyszerű teljes szövegű keresés nyújthat.

Multimédia archívumok keresési problémái

Az SGML a lehető legjobban leegyszerűsített DTD szabványára, a HTML nyelvre épülő WWW oldalak elterjedése azzal a illúzióval kecsegtette és jelenleg is kecsegteti a felhasználókat, hogy a hálózati hiperlinkek létrehozásával képesek lesznek a nekik megfelelő információkat könnyedén és hatékonyan megkeresni. Ez a technológia lehetővé tette a felkészületlen és nem szakember felhasználó számára is, hogy a legképtelenebb kérdésekre is olyan találati halmazt kapjon, amely egyrészt a legkülönfélébb információhordozókon tárolt, másrészt akármilyen távoli szövegösszefüggésben megjelenő adatokat is tartalmazza. Ez a mennyiségi robbanás azt az illúziót kelti a felhasználóban, hogy minőségi szempontból is lényegi információhoz jut, hiszen nem feltételezheti, hogy egy több tízezres találati halmazban ne lelje meg a számára lényegi információt.

Ma már megjelentek azok a szoftvereszközök (pl. a shareware Copernicus), amelyek lehetővé teszik, hogy az általunk megfogalmazott keresőkérdésre ne csak egy, hanem számtalan Internet-keresőmotor (Internet-böngésző) egyidejű üzemeltetése révén kapjunk választ. Ennek a szoftvernek a piaci megfeleltetése az a számtalan cég, amely az Internet irdatlan és vegyes információtömegében éppen az ügyfélnek megfelelő és kielégítő információ gyors és profi meglelésével kecsegteti ügyfeleit. Ezek a technikák azonban csak látszólagos megoldást adnak a heterogén metainformációs rendszerekben való visszakeresésre. Az igazi megoldás még várat magára. Természetesen már ma is találunk kísérleteket, amelyek (illetve együttes alkalmazásuk) tényleges eredményeket ígérnek.

Ilyen megoldást sugall például az a technológia, amely a hagyományos (legtöbbször relációs) adatbázisokban tárolt információt összekapcsolja, illetve kiegészíti a „tudásbázisok" (knowledge bases) lexikális, enciklopédikus háttér-információval. Ennek a technológiának az alapja a „Találj meg engem" módszer („find-me system"). A módszer az ügyfél, a felhasználó meghatározott viselkedési típusát fogalmazza meg: „Videokazettát szeretnék kölcsönözni. Pontosan nem tudom, hogy mit, de valami olyasmire gondoltam, mint a Vissza a jövőbe. Esetleg lehet *Michael J. Fox* valamelyik másik filmje, de bármi érdekel az időutazással kapcsolatban."

A módszer e bizonytalan igénymegfogalmazásra igyekszik hatékony választ adni tudó technikát kidolgozni. A rendszer lényeges eleme a megfelelően szegmentált adatokat tartalmazó adatbázis, amely potenciálisan valahol őrzi azokat a megadandó paramétereknek megfelelő adatokat, amelyek az adott igényt kielégíthetik. A rendszer másik része pedig egy olyan adat- (illetve paraméter-) halmaz, amely az adatbázisban tárolt információ megtalálásához hozzásegít. A fenti kérdésre adott választ a Video Navigator elnevezésű program. A kérdés nem az adatbázist éri el, hanem a tudásbázist. Itt egy megfelelő algoritmus megállapítja, hogy a kérdés szerzőre, műfajra vagy kulcsszóra vonatkozik-e. A meghatározás után a tudásbázist a megfelelő szempontú index alapján lekérdezi, ennek alapján pontosítja a keresőkérdést, és csak eztán kérdezi le az adatbázist. Ez a lényegében egyszerű módszer elég jó megközelítéssel lehetővé teszi, hogy bizonytalan kérdésre kielégítő választ kaphassunk.

Metainformációs rendszerek

Ezek a rendszerek általában a metainformációs kapcsolatokra építenek. Azaz nem a különböző (relációs és teljes szövegű) adatbázisokban, illetve multimédia, hipertext kapcsolatokat felhasználó virtuális és digitális könyvtárakban tárolt adatelemeket összekapcsoló technológiát alkalmazzák, hanem magára az összekapcsolási mechanizmusra ajánlanak kész módszert, amelynek alkalmazásával a legkülönfélébb típusú információkat tartalmazó választ kaphatunk.

Általában kétféle típussal találkozhatunk: egy hagyományosabb, hierarchikus tudományfelosztáson alapuló módszerrel, és egy számítástechnikai, a programozói munkát előtérbe helyező (a logikai strukturalizmusra kacsingató) megoldással.

Az előbbi a Subject Tree, a Yahoo! vagy a HuDir által is alkalmazott tematikus megoldást részesíti előnyben, részben elébe menve a valódi keresőkérdés megfogalmazásának. Tulajdonképpen ezt a módszert egészíti ki az Interneten található információforrások MARC-alapú feldolgozása is. Hiszen a megfelelő szabvány alkalmazásával bizonyos adatsorokat eleve előnyös pozícióba helyez: az információforrás típusa (adatbázis, web,

gopher, ftp stb.), az információforrás helye (URL, illetve PURL), az adott könyvtárban alkalmazott tematikai besorolási technika (LC szakjelzet, ETO, természetes nyelvi deszkriptorok).

A másik típus a metainformációs rendszerek lekérdezési technikájának algoritmizálásával próbálkozik. Már az első szöveges adatbázisok megjelenésekor lehetett találkozni azzal a programozói célkitűzéssel, hogy a hagyományos civilizációs információfeldolgozást (azaz a könyvtári, bibliográfiai, kiadói, statisztikai stb. technikákat) számítástechnikai megoldásokkal helyettesítsék. Ez részben hamis szemléleten alapuló tevékenység volt, hiszen egy adott tevékenység algoritmizálását az adott tevékenység automatizálásával akarta helyettesíteni. Ez olyan tévedés, mint amikor *Kempelen* sakkozógépét a sakkozó helyettesítésére akarták használni, pedig annak zsenialitása az izommozgások mechanikus leképezésében állt. Ugyanakkor ez az útkeresés nagyon sok olyan érdekességre világított rá, amely a problémamegoldás szemiotikai struktúrájának fontos sajátosságaira hívta fel a figyelmet. (Lásd *Pierce, Sebeők* elméleti munkássága.)

Az egyik ilyen megoldás a Michigani Egyetem Digitális Könyvtárában alkalmazott módszer (UMDL), amely a következő típusú kérdésekre keresi a választ: „Van ebben a könyvtárban olyan dokumentum, amely tartalmazza az általam feltett kérdést kielégítő választ?" A rendszer négy komponensből áll:

- *Users*: zseniális ötlet, hogy a felhasználót és a felhasználóban megfogalmazandó kérdést a rendszer részeként tüntetik fel, hiszen ez az a mozzanat, amely mozgásba hozza a rendszert (számítástechnikai szempontból ez a feltehető kérdések előre megjósolt típusainak preferált algoritmizálását jelenti).
- *User Interface Agents:* ez az a felület, amelynek segítségével egyrészt az olvasó megfogalmazhatja és pontosíthatja a kérdését, másrészt a rendszer a megfelelő formális logikai nyelvre lefordíthatja azt.
- *Mediators*: azoknak a célorientált (task-specific) és hálózati szabványoknak az összessége, amelyek lehetővé teszik, hogy a formális nyelven megfogalmazott keresőkérdés eljusson az adat-, illetve tudásbázishoz.
- *Collection Interface Agents:* ez egy olyan számítástechnikai applikáció, amely egyrészt tartalmazza az adatbázisok (gyűjtemények) pontos leírását a megfelelő szegmentumokkal (pl. MARC, TEI-ajánlások), másrészt az adatbázisok (gyűjtemények) megfelelően szegmentált halmazát.

Hasonló, de egyszerűbb megoldás a heterogén adatbázisok egységes logikai kontextusban való lekérdezése.

Komplex visszakeresési módszerek

Az eddigiekből talán már levonható a következtetés: a heterogén, multimédia környezetben a komplex visszakeresés nemcsak új keresési technika kidolgozását teszi szükségessé, hanem az adatrögzítés és szegmentálás/kódolás új módszerét is.

A szöveges és nem szöveges információk általános rögzítésének és kódolásának ilyen újfajta lehetőségét nyújthatja az SGML. Az SGML olyan általános leíró nyelv (valójában metanyelv), amelynek segítségével egy adott típusú dokumentumcsoport, egy digitális objektum oly módon rögzíthető, hogy egyrészt megtarthatja az eredeti fizikai dokumentum jellemzőit, másrészt a legkülönfélébb szempontok alapján megjeleníthető, és így visszakereshetővé válik.

Az SGML ténylegesen egy hierarchikus szöveg- (ha úgy tetszik objektum-) modell kodifikálása: a hierarchikus jellegű modell leírásáé. A szakirodalom általában *Lotman* hierarchikus szövegmodelljével példázza ezt a szövegleírást: „... a szöveg leírásának nyelve hierarchia". A szabvány a szöveg metaadatainak leírását szabályozza: ezek a metaadatok nem a végrehajtandó utasítást rögzítik (mint pl. az SGML leszűkített változatában, a HTML dokumentumokban), hanem a szövegszegmentumok azon tulajdonságát, amely ezt rendszeresen kiváltja, feltételezi.

Lássunk egy példát.

Egy vers HTML-rögzítése általában így kezdődik:

```
<html>
<head><title>Harmincharmadik</title></head>
<body>
<h1 align=center>Harmincharmadik</h1>
<h6 align=center>Az nőtája: Bánja az Úristen</h6>
<h3 align=center>Kiben bűne ...</h3>
<pre>
   Bocsásd meg, Úristen, ifjúságomnak a vétkét
    Sok hitetlenségét, undok fertelmességét.
   .....
<pre>
```

Ez a rövid szöveg is nyilvánvalóvá teszi, hogy a HTML-leírásnak valójában semmi köze az eredeti szöveg hierarchikus modelljéhez. Egyetlen célja, hogy egy adott szöveg jellegzetes szegmentumait (cím, alcím, versszak stb.) különböző stílusjegyekkel megkülönböztesse, és megjelenési formáját megszabja. A szöveg szemantikai rétegeit ez a fajta leírás pedig egyáltalán nem érinti, s így az is világos, hogy egy HTML dokumentum visszakereshetősége erősen korlátozott. Legfeljebb (pl. perl scriptek segítségével) egyes jellegzetes elemeire lehet keresni (pl. cím, szerző), a kifinomultabb tartalmi feltárás már lehetetlen.

Ugyanennek a szövegnek az SGML szabvány alapján történő modellezése viszont már összetettebb rögzítést és visszakeresést tesz lehetővé:

```
<sgml>
....
<body>
<div1 name=„vers">
<head>
<title>Harmincharmadik</title>
<line.break>
<highkighted rendered=„italic"> Az nótája: Bánja az Úristen </highlited>
<line.break>Kiben bűne ....
<line.break> .....
<line.break>
<! – A nóta kottája -->
</head>
<poem id=„rpha185">
<stanza id=„bb03301">
<l n=„bb0330101">
<highlited rendered=„bold">B</highlited> Bocsásd meg, Úristen, ifjúságomnak a vétkét,</l>
<l n=„bb0330102">Sok hitetlenségét, undok fertelmességét,</l>
<l n=„bb0330103">..... </l>
</sgml>
```

E rövid részletből is kiolvashatók azok a szabályok, amelyek alkalmazásával egy SGML dokumentum képes visszaadni egy szöveg (objektum) hierarchikus struktúráját, az egyes elemekhez rendelhető stílusokat és értelmezéseket, illetve az egész SGML dokumentum kapcsolódását (lásd az azonosító – id – elemek használatát) más dokumentumokhoz, s egyben relációit az összes dokumentum hierarchiájában.

Az eddigiekben zárójelbe tett kiterjesztés fontosságára szeretném felhívni a figyelmet: az SGML nyelv nemcsak szövegek, hanem bármilyen objektum (kép, mozgókép, hangzó anyag, Internetforrások, MARC formátum stb.) leírására is alkalmazható.

Lássunk egy példát elektronikus leveleink feldolgozására.

```
<email>
<head>
<from><name>Bakonyi
Geza</name><nickname>bg</nickname><address>
bakonyi@bibl.u-szeged.hu</address><from>
<to><name>TMT
Szerkesztosege</name><address>tmt@omk.omikk.hu
</address></to>
<subject>Cikk kezirata</subject>
</head>
<body>
<p>Kedves Szanto Peter. A mellekletben kuldom a cikk keziratat.</p>
<attach ecoding=„mime" name=„cikk.doc"/>
</body>
</email>
```

Ez a fajta szövegrögzítés ugyanazt a részletes szegmentálást képes biztosítani a szövegnek, mint például egy relációs adatbázis táblái egy feldolgozandó adatsornak. Az egyes szövegelemek pedig perl scriptekkel, illetve Java-alkalmazásokkal hasonló hatékonysággal feldolgozhatók. Az így rögzített és kódolt szöveg pedig nemcsak más, de hasonló típusú szövegekre alkalmazható, hanem bármilyen adatbázissal, SGML szabvány alapján feldolgozott szövegarchívum szövegeivel összekapcsolható, azonos szempontok alapján visszakereshető.

Éppen ezt a tulajdonságát kívánjuk kihasználni a szegedi Egyetemi Könyvtárban, amikor létrehoztuk az Eruditio adatbázist.

Eruditio

Olyan „tudásbázis"-ról van szó, amely négy különböző adatbázist és egy szövegarchívumot kapcsol össze oly módon, hogy szükség esetén lehetővé válik a bármelyik résznél feltett keresőkérdés kiterjesztése távoli adatbázisra is. A visszakeresést biztosító felületet HTML-oldalak adják, amelyek biztosítják a relációs adatbázis (Postgres) tábláiban való SQL kereséshez való illeszkedést, illetve a szövegarchívumban való teljes szövegű keresést. (Itt jegyzem meg, hogy sokan egészen téves elképzeléseket fogalmaznak meg a teljes szövegű kereséssel kapcsolatban. Általában olyan „vészhelyzetre" tartogatott megoldásnak tartják, amelyet akkor kell alkalmazni, amikor minden más keresési módszer csődöt mond. Ennek szellemében sokszor olvashatjuk azt az igényt integrált könyvtári rendszerek tendereiben, hogy az indexekre építő keresés mellett legyen lehetőség szabad szöveges keresésre is. Pedig egy teljesen szabad szöveges keresés inkább csak hátráltatja a hatékony keresést, semmint segítené. Jól érzékeltetik ezt az internetes böngészők [AltaVista, Heuréka, Excite, Lycos stb.], amelyek a legártatlanabb kérdésre is tízezres nagyságú találati halmazt állítanak össze. Ezek hatékony használatához is szükséges egy kifinomultabb logikai szintaxis alkalmazásának elsajátítása, s még akkor sem lehetünk biztosak abban, hogy valóban a nekünk szükséges választ kaptuk meg. A szabad szöveges keresés valójában éppen annyira nem szabad, mint ahogy az irodalomban egy időben sokat emlegetett szabad verselés sem a szabályok nélküliség értelmében volt szabad. Így a teljes szövegekben való visszakeresés is csak akkor igazán hatékony, ha meg tud kapaszkodni a szöveg szegmentumaiban és sajátosságaiban, s ezeket a keresés paramétereiként tudja hasznosítani. Éppen ez utóbbiban segíthet az SGML-alapú rögzítés.)

A szövegek (jelen esetben régi magyar könyvtárak leltárai) rögzítése eleve számítógéppel (különböző szövegszerkesztőkkel) történik. A legtöbb

esetben nemcsak a számozott leltár listáját tartalmazzák, hanem magyarázó és értelmező megjegyzéseket, sőt akár tanulmányokat is. A számozott listák kijelölése az előbb említett SGML-leírás alapján világosan megkülönböztethető az egyéb szövegektől, s az egyes sorokban található címleírások is értelmezhetők. Ezeknek az általunk originálisnak elnevezett címleírásoknak a feloldása a már feloldott címek bibliográfiai leírásait tartalmazó adatbázisban történik. A szövegek sajátságai (azonosító szám, tulajdonos stb.) egyéb SGML elemek alkalmazásával megy végbe. Az SGML szegmentumokat egy Java program értelmezi és jeleníti meg. Az egyes dokumentumok leírási szabályait elegendő volt egyszer kitalálni, és az úgynevezett Document Type Definition (DTD) fájlban rögzíteni.

DTD

Mint már említettük, az SGML szabvány elsődleges feladata az, hogy szintaktikai szabályokat biztosítson a szöveg hierarchikusan rendeződő elemeinek formális leírásához. Ennek megfelelően a szabvány egyik leglényegesebb eleme a DTD, amely nem más, mint a szövegben feltételezhető hierarchikus szegmentumrendszernek és jelölésének a leírása.

A DTD a rögzíteni és kódolni kívánt szövegtípus (ez lehet akármi: az elektronikus levél dokumentumtípusától a drámáig, az egyszerű képtől a filmig) általánosított szövegmodelljét fogalmazza meg. Azaz lényegében egy struktúra leírását kísérli meg e struktúra elemeinek és ezen elemek sajátosságainak a leírásával. Természetesen a DTD leírásának is vannak korlátai. Elsősorban az, hogy viszonylag statikus leírást ad, a generatív és szemantikai igényű szövegleírás finomságait nem képes visszaadni. Ez persze nem jelenti azt, hogy ne lehetne kiegészíteni egyéb, szövegen belüli és kívüli, önállóan deklarált egységekkel. Például egy, az Isteni Színjáték szövegstruktúráját leíró DTD elején deklarálhatjuk, hogy a szöveghez csatlakozik a reneszánsz szövegkiadások Pokol ábrázolásainak digitalizált képi anyaga, a kommentárok önállóan feldolgozott gyűjteménye, a Színjáték BRS/Search adatbázisa stb. Ezek mindegyike saját DTD-vel rendelkezhet, és a kapcsolatot ennek elején hasonlóképpen deklarálhatják.

Az SGML nyelv bevezető ismertetései általában egyszerű példákkal szemléltetik ezt a sokszor rendkívül bonyolult leírásgyűjteményt. Ezek egyikét követve az alábbiakban egy éttermi étkezést sajátos érvényű dokumentumtípusként felfogva, DTD-jének fiktív vázlatát ismertetjük.

```
<! Doctype étkezés – DTD éttermi étkezések leírása
számára -->
<! Entity menükártya SYSTEM
„http://www.nagyzaba.hu/menu/menu.html>
```

Ez a bevezető rész egyrészt deklarálja, hogy a DTD milyen dokumentumtípusra érvényes, másrészt a szegmentumok felsorolása előtt felhívja a figyelmet, hogy a leírást kiegészíti egy másik dokumentumtípus, amelynek feltehetően saját DTD-je van.

```
<! Element étkezés (ebéd) + -- egy ebéd minden személy számára>
<! Element ebéd (előétel?, hús, desszert, fizető)
+(ital)>
```

Ezek a sorok az adott dokumentum szövegmodelljének fő struktúráját mutatják. Az étkezés jelen esetben csak ebédből áll, s annak értéke lehet egy vagy több, a vendégek számától függően. Az ebéd viszonylag egyszerű: előétel (a ? arra utal, hogy ez el is maradhat, azaz értéke nulla is lehet), hús, desszert, s természetesen az éttermi ebéd kötelező része a fizető (tehát az is, hogy ki fogja ezt az egészet kifizetni). Az ebédet italrendelés egészíti ki, tetszés szerinti számban (ami nemcsak azt jelenti, hogy a kedves vendég ihat, mint a gödény, hanem azt is, hogy az ital elem többször többfajta értéket is felvehet).

```
<! Element előétel (leves | saláta) >
<! Element leves EMPTY>
<! Element saláta EMPTY>
<! Attlist saláta típus NÉV #köteleződresszing (francia |
ezersziget | kapros) >
```

Azt hiszem most bárki magától is tud következtetni arra, hogy az egyes jelölések milyen tartalomra utalnak: az előétel (ha kérte a kedves vendég) leves vagy saláta lehet; ha valaki mégis kér salátát, akkor kötelezően meg kell adnia, hogy milyen salátát milyen dresszinggel kér. Látható az is, hogy a függőleges vonal a logikai vagy jele.

```
<! Element hús >
<! Attlist hús típus (főtt | sült | rántott) „körítés" ( burgonyapüré | sült krumpli | rizs) >
<! Element desszert (sütemény | palacsinta | fagylalt) >
<! Element (sütemény | palacsinta | fagylalt) EMPTY >
<! Attlist sütemény típus CDATA >
```

Itt az újdonság a CDATA, amely azt jelöli, hogy a sütemény nem bomlik további alstruktúrákra (elemekre), csak meg kell mondani a nevét (azaz csak adatot kell szolgáltatni).

```
<! Attlist palacsinta meleg (meleg | lángoló | hideg)
#IMPLIED >
```

Az IMPLIED az opcionálisra utal: megmondhatjuk, hogy milyen palacsintát kérünk, de ha nem, akkor meleget kapunk (pontosabban ki vagyunk szolgáltatva a pincérnek).

```
<! Element drink (víz | sör | bambi) >
<! Element (víz | sör | bambi) EMPTY >
<! Attlist víz típus (csap) #FIXED csap >
<! Element sör szám >
<! Attlist bambi típus (normál | diétás) #CURRENT >
```

Az utolsó sorok tartogatnak néhány érdekességet. Persze ez a fajta determináltság nem korlátozó jellegű (mint például a relációs adatbáziskezelők esetében a táblák kialakításának szabályai), egyszerűen csak szabályozó szerepük van. Tehát: inni nem kötelező, de ha valaki sört rendel, akkor elegendő a korsók számát mutatni; ha bambit inna, akkor az első rendelésnél mondja meg, milyen fajtát akar, biztos lehet benne, hogy később is ezt kapja; ha meg botor módon vizet kér, akkor akármilyen fajtát is szeretne, csak csapvizet fog kapni.

A példa ugyan nem teljesen szabályos, de remélhetőleg jól reprezentálja, hogy mi is a DTD szerepe, és talán három fontos részegysége is világosan áttekinthető: az alkotóelem (element), a tulajdonság (attribute) és az entitás (entity).

A továbbiakban azonban nem ennek részletezéséről lesz szó, hiszen ez a tanulmány nem annyira az egyes technológiákról, az egyes szabványok kifejtéséről szól, hanem az alkalmazásuk által nyitott lehetőségekről.

TEI

Az SGML szabvány (1986-ban nyilvánították nemzetközi szabvánnyá, ISO 8879 nyilvántartási jellel) egyik jelentős felhasználása született meg 1988-ban, amikor is három nemzetközi tudományos társaság összefogásával olyan tudományos program indult útjára (ez volt a Text Encoding Initiative), amelynek célja az volt, hogy ajánlást dolgozzon ki a természetes nyelvi szövegek számítógépes rögzítésének módszerére. Bevallott álmuk az volt, hogy az egységes rögzítési módszer alkalmazásával megvalósulhasson a „falak nélküli" világkönyvtár eszméje, amelynek anyaga bárhol is legyen a világon, egyformán kódolt és visszakereshető, feldolgozható.

Az ajánlást többféle szövegtípusra is kidolgozták, ami azt jelentette, hogy az egyes szövegtípusoknak megfelelő dokumentumleírásokat (DTD) hoztak létre, így a prózai, verses, drámai szövegekre is. A megalkotott DTD-k alapvetően négy feladatnak igyekeztek megfelelni.

- *A szövegrögzítés minél tökéletesebb tükröztetése.* Alapszinten egyszerűen csak a fontosabb szegmentumok rögzítéséről van szó: prózai művek esetében például kötetek, fejezetek, paragrafusok; verses művek esetében kötetek, ciklusok, versszakok, verssorok; drámai művek esetében felvonások, jelenetek, párbeszédek, szereplők stb. Természetesen ennél részletesebb szövegrögzítési szegmentumok alkalmazására is lehetőség van. Így lehetséges ugyanazon szöveg különböző szövegkiadásainak egyidejű tükröztetése (eltérő laptördelés, szövegváltozat stb.).
- *A genetikus szövegrögzítés*re való törekvés ugyancsak jelen van, ha nem is teljességében, hiszen ennek megjelenítése szétfeszítené az SGML-alapú ajánlás nagyrészt statikus, hierarchikus szerkezetét. Azonban így is lehetőség nyílik a szövegváltozatok többszintű ábrázolására (a nyelvi elemek variációi, legyen szó akár törlésről, akár szó- és mondatvariációkról), a kommentárok és jegyzetek közvetlenül a szöveghez való kapcsolása, időrendi variációk egymásra épülése stb.
- *A szemantikai-értelmező szövegkódolás*ra ugyancsak találunk törekvést, azonban éppen ez az a szint, amely csak töredékesen van jelen az ajánlásban. Ennek ellenére az ajánlás lehetőséget ad a nyelvi, stilisztikai és retorikai jelentésrétegek egyes elemeinek a kiemelésére és csoportosítására. A metaforák és metonímiák különböző típusait például jól nyomon követhetjük a megfelelő tagolások (ilyen a szegmentum- és a referenciajelölés) és attribútumok alkalmazásával.
- A TEI-ajánlás alapján *számítógépes eljárással rögzített és kódolt szöveg* komplex megoldást jelent, s az így feldolgozott szövegek, illetve objektumok összessége már ténylegesen módot ad arra, hogy egy tudományos intézet vagy könyvtár hálózati környezetében olyan „tudásbázisokat" lehessen létrehozni, amelyek messze túlmutatnak a hagyományos szegmentált és teljes szövegű adatbázisok, hagyományos és hipertext szövegarchívumok heterogén konglomerációja által nyújtott szolgáltatások határain.

Azt is el kell azonban ismerni, hogy a számítógépes és hálózati technológia jelen állapota és eszközei sok helyen meghaladták az SGML-alapú TEI-ajánlások korlátait, s azokat ma már részben nehézkésnek érezhetjük. Ezen akar segíteni az Extensible Markup Language (XML) ajánlása, amely részben leegyszerűsíti, részben kiterjeszti az előbbiek eszközeit, megoldásait.

Az ajánlás még nagyon új, lényegében csak ebben az évben kezdett elterjedni, jelentek meg az első alkalmazások és az XML-alapú szoftverek, Java-megoldások. Ennek ellenére már az első pillanatban sikeres terméknek tűnt, s nagyon hamar megjelentek az első XML-alapú TEI-ajánlások is (bibliaszövegek, Shakespeare-drámák).

XML

Az XML leíró nyelv szabályai végtelenül egyszerűek. Elegendő, ha a szöveges objektum jól

formált (well-formed) és érvényes (valid). Ez semmi mást nem jelent, mint az XML specifikációban rögzített néhány egyszerű szabály betartását, és azt, hogy minden egyes nyitó elemhez záró elemnek kell tartoznia (azaz ha valamit megnyitottunk, zárjuk le – az SGML szövegben ez nem kötelező).

Az XML specifikáció ugyan deklarálja, hogy az SGML specifikációval kompatíbilis akar maradni, de annak bonyolultságát lényegesen csökkenti. Erre azt a módszert találták ki, hogy a szövegrögzítés három fontos elemét elválasztották egymástól: a tükröztető szövegrögzítést, a stíluslapokat és a külső kapcsolódásokat. Az utóbbi két elem szövegen kívülre helyezésével megnyílt az út az előtt, hogy a szövegrögzítést nagymértékben közelítsék a szemantikai, értelmező jellegű jelöléshez és kódoláshoz. Ezen segít az is, hogy bevezették a „stand-alone" típusú deklarációt, azaz egy szöveges objektum rögzítése akkor is jól formált és érvényes lehet, ha nem csatlakozik hozzá DTD. Ez nagyon megkönnyítette az egyes szövegek rögzítését, hiszen egy-egy DTD megalkotása, de egy, már elkészített alkalmazása is elég bonyolult feladat.

XML/TEI alapú alkalmazások

Az ajánlások alapján lehetővé válna a Magyar Elektronikus Könyvtárhoz hasonló szövegarchívumok hatékony feldolgozása – anélkül, hogy ez a munka jelentősen megterhelné az archívumok gondozóit. Elsősorban arra van szükség, hogy az archívum által őrzött dokumentumtípusoknak megfelelő XML-alapú, egyszerűsített DTD-ket kell kidolgozni, majd a szövegeket ennek megfelelően konvertálni. A továbbiakban a feldolgozás szintjének megfelelően a DTD szkeletonokat bárki tetszés szerint (az ajánlásnak megfelelően) kiegészítheti, stílusokkal és kapcsolatokkal (linkekkel) bővítheti. Ugyancsak fontos eleme a szövegfeldolgozásnak az elektronikus címlap, amely az elektronikus szövegváltozat legfontosabb paramétereit rögzíti, s a további felhasználás jogi és szövegkritikai lehetőségeit világosan meghatározza. Ez hasonló a MEK által kidolgozott fejléchez, de egyes elemei és szegmentumai a nemzetközi ajánlásnak felelnek meg, s így részletesebbek is.

A másik lehetőség a már említett „tudásbázisok" kialakítása. A szegedi Egyetemi Könyvtárban két ilyen kísérleti programot indítottunk el: az egyik egy olasz költő, *Giacomo Leopardi* verseit, a másik *Dante* Isteni Színjátékát dolgozza fel (ez utóbbi kapcsolódik a Commedia TEI-ajánlás alapú olasz feldolgozását szorgalmazó olasz programhoz). A program célja az irodalmi szövegek XML/TEI alapú tükröztető rögzítése, amely figyelembe veszi a szövegek olasz és magyar nyelvű változatait, a különböző kritikai kiadások jegyzetapparátusát, és a magyar, illetve olasz nyelvű kommentárok szövegeit. A szövegekhez külső kapcsolódású, ugyancsak XML-alapú szövegrétegek fognak kapcsolódni (pl. ikonográfiai, jelképeket, szinteket feldolgozó szövegek, multimédia anyagok).

Az XML/TEI alapú szövegfeldolgozás természetesen nem fog minden problémát megoldani, de alkalmazása valószínűleg jelentősen segíteni fog abban, hogy a multimédia, heterogén jellegű hálózati környezet tisztább képet mutasson, és így az itt tárolt információk értelmes módon váljanak visszakereshetővé.

Irodalom

CIOTTI, F.: Il testo elettronico: memorizzazione, codifica ed edizione informatica del testo, in Macchine per leggere. Tradizioni e nuove tecnologie per comprendere i testi, atti del convegno (Firenze 19 novembre 1993), Spoleto, Cisam-Fondazione E. Franceschini,1994.

CIOTTI, F.: Testi elettronici e biblioteche virtuali: problemi teorici e tecnologie. = Schede Umanistiche, 2. köt. 2. sz. 1995.

GOLDEN D.–TÓTH T.–TURI L.: Virtuális örökkévalóság: objektumok a digitális könyvtárban. = TMT, 45. köt. 8–9. sz. 1998. p. 299–314. http://www. neumann-haz.hu/tanulmany/index.htm

HORVÁTH I.: Bölcsészet a bábeli könyvtárban. = 2000, 1997. május, p. 61–63.

The Second International Workshop on Next Generation Information Technologies and Systems (NGITS '95), 27–29 June 1995, Hotel Carlton, Naharia, Israel. http://sunsite.ust.hk/dblp/db/conf/ngits/ ngits95.html

TURI L.: Számítógép az irodalomtudományban (szakdolgozat). ELTE Tanárképző Főiskolai Kar, 1992. MEK Társadalomtudományi olvasó.

Szabványok, leírások

SGML Solutions SGML Overview.: http://www.chubsoft.demon.co.uk/files/sgml_ov.htm

Introduction to the SGML PRIMER.: http://www.sq.com/sgmlinfo/primintr.html

SGML Open Web Site: Index Page.: http://www.sgmlopen.org/sgml/

SGML Syntax Summary.: http://www.tiac.net/users/bingham/sgmlsyn/contents.htm

The SGML/XML Web Page.: http://www.sil.org/sgml/sgml.html

SoftQuad: SGML Resources.: http://www.softquad.com/htmlsgml/

TEI A 15: TEI Application Page.: http://www-tei.uic.edu/orgs/tei/app/topics.html

SGML.: http://telemat.die.unifi.it/book/Internet/Sgml/indsgml.htm

Writing HTML, SGML, TEI, etc.: http://sunrise.sote.hu/kszerv/kszerve/idk/idk/html.html

On SGML and HTML.: http://www.sch.bme.hu/misc/howto/html40/intro/sgmltut.html
XML: Overview.: http://sunsite.unc.edu/pub/sun-info/standards/xml/pres/9707ja/sld02000.htm
XML.com.: http://www.xml.com/
Text Encoding Initiative Home Page.: http://www-tei.uic.edu/orgs/tei/
CRILet – Codifica SGML e Text Encoding Inititive.: http://crilet.let.uniroma1.it/sgml/sgml.htm
Duecento.: http://www.silab.it/frox/200/pwhomita.htm
TEI P3 DTDs, Declarations, & Entity Sets.: http://www.uic.edu/orgs/tei/p3/dtd/dtd.html
James Tauber's XML INFO:Example XML Documents.: http://www.jtauber.com/xml/examples.html
Presenting XML.: http://www.mcp.com/info/1-57521/1-57521-334-6/

Beérkezett: 1998. VI. 2-án.

Rendezvénynaptár

2. európai konferencia a digitális könyvtárak kutatásáról és korszerű technológiájáról
Heraklion, 1998. szeptember 19–23.
Szervező: Rena Kalaitzaki
University of Crete, Computer Science Department
Tel.: +30 81 393504
Fax: +30 81 393501
E-mail: ecdl@cc.uch.gr

Kutatások és fejlett technológiák a digitális könyvtárak számára
Heraklion (Görögország),
1998. szeptember 21–23.
Szervező: University of Crete
Mrs. Rena Kalaitzaki
Computer Science Department
Ampelokipi, Heraklion
Tel.: +30 81 393504
Fax: +30 81 393501
E-mail: ecdl@cc.uch.gr
URL: http://www.ics.forth.gr/2EuroDL

Compfair '98, 11. Nemzetközi Számítástechnikai és Telekommunikációs Szakkiállítás és Szakvásár
Budapest, 1998. október 13–17.
Szervező: Compexpo Kft.
1053 Budapest
Kálvin tér 5.
Tel.: 317 6760
Fax: 317 0436
E-mail: ihring@compexpo.datanet.hu

EPIDOS (European Patent Information and Documentation Systems) éves konferencia
Jéna, 1998. október 20–22.
Szervező: European Patent Office
Dienststelle Wien
Schottenfeldgasse 29
Postfach 82, A-1072 Wien
Tel.: +43 1 521 26-0
Fax: +43 1 521 26 54 91
URL: www.european-patent-office.org

10. nemzetközi kémiai információs konferencia
Nimes (Franciaország), 1998. október 18–21.
Szervező: Infonortics Ltd.
Chemical Information Conference
15 Market Place
Tetbury
Glos, GLS8 8DD, UK
Tel.: +44 1666 505772
Fax: +44 1666 505774
E-mail: chemical@informatics.com

A jövő könyvtárai és az emberi kommunikáció. 20. IATUL-konferencia
Chania (Kréta, Görögország), 1999. május 16–22.
Szervező: Technical University of Crete
The Library
University Campus, 73100 Chania
Tel.: +30 821 37281
Fax: +30 821 28460
E-mail: anthi@library.tuc.gr