

DIGITALIZÁLÁS

Bartók István

az irodalomtudomány kandidátusa
MTA Irodalomtudományi Intézet
bar8062@helka.iif.hu

Horváth Iván

az irodalomtudomány doktora
ELTE Irodalomtörténeti Intézet – ivan@elte.hu

Mayer Gyula

PhD, MTA Ókortudományi Kutatócsoport
gam@cs.elte.hu

Tóth Tünde

PhD, egyetemi adjunktus, Pannon Egyetem
Magyar Ir.tud. Tsz, ttunde@almos.vein.hu

Golden Dániel

tudományos segédmunkatárs
MTA Filozófiai Kutatóintézet
Golden@webmail.phil-inst.hu

Káldos János

Országos Széchényi Könyvtár, MEK
kaldos@oszk.hu

Mártonfi Attila

MTA–ELTE Nagyszótári Kutatócsoport
wad@ludens.elte.hu

Vadai István

Szegedi Egyetem
vadai@tiszanet.hu

Vaskó Péter

szerkesztő, Filmvilág
vaskopeter@enternet.hu

az MTA Informatikai Irodalomtudomány Munkabizottságának tagjai

Nem-szövegek digitalizálása

Különböztessük meg azt az esetet, amikor szöveget digitalizálunk, attól, amikor másvalamit. Munkabizottságunknak természetesen csak az első esetről lehet álláspontja. Akadémiánk I. Osztályában viszont a második is előfordul: zeneművek és mozgóképek számjegyes rögzítései.

A nem-szöveg példája legyen a film. Felhasználási területe bővül. Az emberek kevesebbet olvasnak, és több filmet néznek. Megváltoznak a befogadói szokások.

A digitális feldolgozás nyomán közös képzetes térbe kerülnek át a különböző természetű anyagok, ugyanakkor éppen ez az új lehetőség okoz fejlődést az alkotásban, archiválásban és a terjesztésben. Az egyik fő probléma, hogy a film erősen eszközfüggő

médium, a mozgóképes tartalom kódolásához és dekódolásához bonyolult berendezésekre van szükség. Emiatt a rohamosan romló állagú filmes művelődési kincs túlélési esélyeit tekintve hátrányban van az elvont szimbólumokkal operáló, kis helyigényű, olcsón sokszorosítható írott szövegekkel szemben. Ráadásul a digitális feldolgozás csupán elvileg jelent egységes közeget, valójában a digitális rögzítés és feldolgozás területén a szabványok, kódolási és dekódolási eljárások nagy változatosságot mutatnak. Mivel a digitális képfeldolgozás gyorsan fejlődik, a változatosság egyhamar valószínűleg nem fog csökkenni.

Nehéz a mozgóképes tartalom tartalmi feltárása. A leírási módszerek javarészt az írott szövegekre alapozott könyvtári metódusok átvételéből származnak, és alapvetően

metaadatok leírását jelentik. A film esetében azonban nincs lehetőség a szövegeknél megszokott, a mutatókat többé-kevésbé helyettesíteni képes, ún. teljes szövegű keresésre, ami másutt hatékonyan egészíti ki a metaadatrendszerrel.

Csak mostanában indult el a kifejezetten filmek feltárását elősegítő eljárások kidolgozása, amely automatikus kép- és hangelemző eljárásokkal, mintafelismerő algoritmusokkal igyekszik szemantikai térképet adni a mozgókép tartalmáról.

A digitalizált mozgóképnek nem lebecsülendő közvetett művelődésszociológiai hatása is van. A film területén is jelentős változást hoz a kulturális javakhoz való könnyű internetes hozzáférés. Zenéket és filmeket a felhasználók az interneten testvériesen megosztanak egymással. Elképzelhető, hogy az új művelődéspolitikai-gazdasági-jogi dilemma megoldásában az amerikai eredetű *GNU/Linux* (www.gnu.org) mozgalom vagy az ennek tapasztalatait általánosító magyar *f-book*-törekvés (www.f-book.com), esetleg a legújabb javaslat, az amerikai *Creative Commons* (www.creativecommons.org) nyújt megoldást.

Szövegek digitalizálása

1. A szöveg jelek rendezett halmaza.
2. Jel az, amit egy jelkészlethalmaz egy elemébe leképezünk.
3. Pontos másolás történik akkor, amikor a forrásul szolgáló szöveg elemeit és az eredményül kapott szöveg elemeit a jelkészlethalmaznak rendre ugyanazokba az elemébe képezzük le.

Az utolsó meghatározás értelmében csak szöveget lehet pontosan lemásolni. Amit nem lehet lemásolni belőle, az nem is része a szövegnek. A szövegdigitalizálás csupán a pontos másolás egy fajtája. Kép-, mozgókép- és hangállományok digitalizálása pedig nem egyéb, mint szöveggé alakításuk. Csak szöveges adatbázisok vannak: nemcsak a

metaadatok, hanem a kép- és hangadatok is csak már szövegekké alakulva kerülhetnek be az adatbázisba.

A számítógépek korát ebben az értelemben nem a szövegek visszaszorulása, az ún. „Gutenberg-galaxis” vége jellemzi, hanem sokkal inkább a szövegeken alapuló kommunikáció uralkodóvá válása. A látható és hallható felszín gyakran kép, mozgókép, hang – a mélyben azonban szövegek áramlanak.

A 2. és 3. meghatározás szerint csak akkor lehet sikeres egy szöveg gépi rögzítése, ha történeti ismereteink segítségével azonosítani tudjuk a szóban forgó szövegre jellemző, többé-kevésbé egyedi jelkészlet-halmazt. (Vannak szokatlan esetek. Bebizonyították például, hogy Mallarmé *Kockadobás*-ában az átírás csak akkor tekinthető pontosnak, ha megőrzi az eredeti nyomdai betűmetszést és -méretet, mivel e tényezőknek épp ebben a költeményben jelentésük van.) Az ilyen nehézségekből az következik, hogy a gépi szövegrögzítés elméleti hozzáértést igényel. A tevékenységet nem egyszerű technikai műveletnek, hanem szöveggondozásnak kell tekinteni. Helytelen, hogy a papíron lévő dokumentumok elektronikus változatainak megalkotását gyakran műkedvelők vagy a dokumentumokat őrző gyűjtemények könyvtárosai vagy éppen tartalomszolgáltatásra szakosodott üzleti vállalkozások sebtében kiképzett alkalmazottai végzik. Megfigyelhető, hogy gyakran épp a szövegrögzítéshez értő tanszékek és kutatócsoportok nem jutnak támogatáshoz. A pénzt inkább olyanok kapják, akik nem rendelkeznek megfelelő szakértelemmel (<http://magyar-irodalom.elte.hu/biop/barbar/>).

Ez a probléma lényegében a szöveg fogalmának laikus értelmezéséből fakad. A döntéshozók bizonyára abban a hiszemben intézkednek digitalizálási munkálatokról, hogy tudják, miről van szó.

A géppel rögzítendő szövegre eredetileg jellemző jelkészlethalmaz azonosításával

többféleképpen megpróbálkozhatunk; itt a feladatok három lehetséges megközelítését tekintjük át.

Történeti kutatással megállapítjuk, hogy a szövegben mi számít jelnek és mi zajnak, majd rögzítjük a jeleket.

Gondoskodunk megfelelően gazdag mai gépi jelkészletről, mely alkalmas arra, hogy a történeti követelményeknek tökéletesen megfeleljen.

Zárójelbe tesszük az egész jelkészlet-kérdést, és a jó szerencsére bízunk magunkat. Ezt az utat kell követnünk akkor, hogyha nagyon különböző korokból és szerzőktől származó szövegeket kell igen gyorsan rögzítenünk, lehetőleg önműködő készülékek igénybevételével. A mai kor könyvtárosa olykor ezt az utat kénytelen járn.

Gépi szövegnyilvántartás és -kiadás: eredmények és kudarcok

Magyarországon az 1970–80-as években készült a régi magyar költészet gépi nyilvántartása, mely nagyon sok és nagyon különmemű (bibliográfiai, poétikai és irodalomtörténeti) adatot dolgozott fel, egyebek közt abból a célból, hogy majd ez lehessen az alapja a későbbi szövegrögzítésnek (<http://magyar-irodalom.elte.hu/repertorium/>). A Balassi-versek 1998-as hálózati kritikai kiadása (<http://magyar-irodalom.elte.hu/gepesk/bbom/>), majd József Attila prózai munkáié (1999, <http://magyar-irodalom.elte.hu/ja/>) nemzetközi összehasonlításban is úttörő vállalkozások voltak. Az 1980-as években újdonság volt a szöveges adatbázisban ún. relációs adatmodell használata. Új volt a háttértörténeti szempont: a szövegek sorsában bekövetkezett összes textológiai esemény (másolás, szedés, cenzúrázás, idézés stb.) láthatóvá és tanulmányozhatóvá tétele a befogadástörténész számára. Az 1990-es évek végének hálózati kritikai kiadásaiiban pedig új volt a programozott véletlenszerű olvasás módszere. Az egykori versenyelőny

mára a támogatások elmaradása nyomán megszűnt; a szövegnyilvántartás és a kiadások elavultak.

A klasszika filológia hazai művelői egyrészt igyekeznek a hazai és külföldi kutatók számára föltárni és közzétenni azokat a forrásokat, amelyek lelőhelyük, őrzési helyük, tárgyuk vagy egyéb miatt elsősorban a hazai kutatás látókörébe tartoznak, másrésztől bekapcsolódnak a szakok nemzetközi vérkeringésébe. Az első csoport jellegzetes példája a *Corpus Inscriptionum Graecarum Pannonicarum* (CIGP, <http://delfin.klte.hu/~history/sophia/felirat.html> vagy http://www.iif.hu/dokumentumok/niif_fuzetek/okortud/okor8.html), mely – mint elnevezése utal rá – a Pannonia provinciához köthető ókori, görög nyelvű feliratok több szempontú dokumentálására és feldolgozására vállalkozott. A magyar okortudományi kutatások teljes körű bibliográfiai feltárása a célja a *Bibliographia Studiorum Antiquorum in Hungaria* (BiSAH, <http://luna.btk.ppke.hu>) nevű vállalkozásnak, melynek keretében korszerű szabványok és eszközök (unicode <http://www.unicode.org/>, SQL <http://www.sql.org/>, linux <http://www.linux.org/>, apache <http://www.apache.org/>, PostgreSQL <http://www.postgresql.org/>) fölhasználásával olyan rendszer épült ki, melynek minden egyes eleme szabadon fölhasználható akár más, hasonló feladatok megoldására, akár az elkészült adatbázis továbbfejlesztésére. Nemzetközi együttműködésben, magyar kutató részvételével készül az oxfordi pontú *Electronic Text Corpus of Sumerian Literature* (ETCSL, <http://www-etcsl.orient.ox.ac.uk/>). (Egy technológiailag elavult, de még használható kísérlet egy fizetős szövegadatbank szabad eszközökkel való implementálására a Betatex = <http://clphx.btk.ppke.hu/classics/gam/betatex.html>). Több kutatóhelyen is készül olyan – (latex, <http://www.latex-project.org/>, <http://www.math.bme.hu/latex/>, illetve edmac ([833](http://www.</p>
</div>
<div data-bbox=)

homepages.ucl.ac.uk/~ucgadkw/edmac/) formában – kritikai kiadás, mely elég következetes tagolással jár ahhoz, hogy például valamilyen SGML-változatba (<http://www.w3.org/MarkUp/SGML/>, <http://xml.coverpages.org/sgml.html>) konvertálható legyen (Janus Pannonius; Hipereidész).

Mind a magyar, mind a klasszika filológiában a digitális kritikai kiadások elkészítését jelentősen hátráltatja, hogy azok technológiai és tartalmi szempontból egyaránt folyamatos karbantartást igényelnek, és ez stabil és legalább részben az adott feladathoz rendelt személyi feltételeket kívánna meg, ezek azonban a kutatóhelyeken manapság csak elvétve állnak fenn. Minden ilyen feladatnak – ha színvonalasan végzik – nagy az élők munkáigénye, azonban sem (a nyomtatottól eltérő) megjelenési formája, sem (a hozzá nem értők szemében) pusztán reprodukív tünő jellege nem jár olyan megbecsüléssel, amely a kutatók számára különösebben vonzóvá tenné ezt a foglalatosságot.

A tudományos megbecsülést megadhatja a nyomtatott megjelenés, különösen ha külföldi és tekintélyes a kiadvállalat; azonban másrésztől éppen ez jogi akadály lehet a korszerű, szabadon hozzáférhető digitális közzétételnek.

Gépi jelkészlet a szövegnyilvántartás és -kiadás számára

Anyelvészet legősibb történetének is már a digitalizálás az alapja. Hiszen a folytonosan változó akusztikai jelenségből, a beszédfolyamnak a diszkrét jegyekre: szótagokra vagy betűkre bontása talán a legkorábbi digitalizálási hullám az emberiség történetében. Ez természetesen független volt az elektronikus technológiáktól.

A különféle írásrendszerek jeleinek számítógépes reprezentálása azonban nem triviális probléma. Mindezt még súlyosbítja az a tényező is, hogy a számítógépes kutatások

legfőbb közege épp egy olyan nyelv, melynek írása szinte a legkevesebb jelet használja.

A mai nyelvek ma használatos írásainak jeleire az utóbbi évtizedekben többféle kódolási mechanizmust dolgoztak ki. A korábbi időkben különféle 8 bites kód táblák változtatásával törekedtek úrrá lenni a nehézségeken, de mivel ez egyszerre csupán legfeljebb 256 jel használatát tette lehetővé, számos problémát megoldatlanul hagyott. A két legfőbb: egyrészt egy nyelvre rendszerint számos versengő kódolás is létezett, így elkerülhetetlen volt a gyakori konverzió; másrészt azok a dokumentumok, amelyek (például többnyelvűségük okán) egymás mellett több kódtáblát használtak, nem teljesítették az egy jel = egy kód elvét, így a különféle szövegműveletek (keresés, csere stb.) jelentős bonyodalmat okoztak. A megoldást a Unicode jelentette. Ennek mára elterjedt szegmense 16 bites kódolást, tehát több mint 65 ezer jel egyidejű használatát teszi lehetővé, de elvben lehetőség van a 32 bites reprezentációra, azaz több mint 4 milliárd jel egyszerre történő használatára is.

A régebbi irodalommal foglalkozók igénylik, hogy a ma már nem használatos nyomdai, sőt kódexbetűket számítógéppel elő tudják állítani, a szövegek pontos, betűhű megjelenítéséről ugyanis sem a nyelvész, sem az irodalmár nem mondhat le.

A Unicode konzorcium azonban a különféle régi betűk kódolását egyelőre nem oldotta meg. Ezen jelek túlnyomó része nem magyarspecifikus, hiszen a környező nyelvek (különösen a német, illetve elsősorban a rövidítések, ligatúrák terén a latin) régebbi gyakorlatában szintűgy megtalálhatók voltak.

Az elszigetelt egyéni próbálkozások tanúbizonysága szerint a különleges betűk előállítása technikailag megoldható. Néhány hónapja megalakult munkabizottságunk elő fogja segíteni, hogy egységes, könnyen elérhető és általánosan használható, szabványosított karakterkészlet álljon a kutatók és olva-

sók rendelkezésére. A számítógépes nyelvészet Magyarországon az 1960-as évek óta számos nagy eredményt ért el. Méltatlan e történethez a magyar betűkódok helyzete.

*Az ipari szövegrögzítés műhelye:
a könyvtár*

A hagyományos könyvtári dokumentumfogalom nehezen illik rá a digitális információhordozóra.¹ A lezárt, időtálló hordozón létező dokumentumokhoz szokott könyvtárosnak szembe kell néznie a digitális információ olyan tulajdonságaival, mint a változékonyság, a fizikai adathordozók sérülékenysége, az információ kezeléséhez szükséges hardverek és szoftverek gyors elavulása. Emiatt lett távlati cél a hosszú távú megőrzést szolgáló ajánlások, szabványok, technológiák kialakítása. A könyvtárak támogatják az eszköz- és gépfüggetlen adatformátumokat, illetve olyan virtuális számítógépek kifejlesztését, amelyek képesek az eszközfüggetlen adatok olvasására. Sok könyvtárban muzeális információtechnológiai részleget is kialakítanak, melyekben elavult hardverek és szoftverek biztosítják az információ értelmezését. Fel kell készülniük nagy mennyiségű digitális objektum konvertálására és szállítására is.

A könyvtárak hagyományos feladata az írott örökség összegyűjtése, nyilvántartása, megőrzése és szolgáltatása. Az internet új feladatok elé állítja a könyvtárt. Az írásbeliség új formái jelentek meg (e-mail, fórum, chat, blog stb.). Ilyenkor mindig szembe kell nézni a helymeghatározás, azonosíthatóság, hitelesség kérdéseivel. A könyvtárak ezért aktívan támogatják a karbantartható egyetemes helymeghatározó rendszer (URN, <http://nbn.urn.hu/>) továbbá a digitális objektumok egyedi azonosítását lehetővé tevő rendszerek (például DOI, <http://www.doi.org/>) kifejlesztését és elterjedését, vala-

mint a digitális dokumentumokba beépülő metaadatokra és azonosító jelekre irányuló kezdeményezéseket (EXIF, XMP, IPTC stb.: <http://www.exif.org/>, <http://www.adobe.com/products/xmp/>, <http://www.iptc.org/>).

A digitális információ hatalmas mennyisége miatt annak összegyűjtése, nyilvántartása, megőrzése új módszerek bevezetését teszik szükségessé. Kísérletek folynak a világháló szakaszos archiválására (<http://www.archive.org>). Egyre több könyvtár vállalkozik az ún. nemzeti web-terek (<http://www.webarchiv.cz/>) vagy néhány fontosnak ítélt hálózati hely archiválására. Tizenkét nagy könyvtár közreműködésével 2004-ben létrejött az International Internet Preservation Consortium (IIPC, <http://www.netpreserve.org/>). A nagy adatmennyiség feldolgozása is új módszerek bevezetését teszi szükségessé. Biztató kísérletek folynak szöveg- és képlemező rendszerek kiépítésére, illetve automatikus metaadatrögzítő eszközök kifejlesztésére is.

A könyvtárak alapvető feladata lett a hatalmas mennyiségű hagyományos könyvtári anyag digitalizálása és hozzáférhetővé tétele. Napról napra újabb dokumentumok digitális változata válik elérhetővé. A könyvtári digitalizálásba bekapcsolódtak a piaci szféra olyan óriásai is, mint a Google (<http://books.google.com>) vagy az Open Content Alliance (<http://www.opencontentalliance.org/>). A technika lehetővé teszi, hogy a könyvtárakban őrzött anyag nagy része belátható időn belül digitális formában is elérhető legyen.

A könyvtár, különösen a mindenhol központi szerepet játszó nemzeti könyvtár helyzete és feladata jelentősen megváltozott. Digitális dokumentumok összegyűjtésével, illetve azok folyamatos, online közzétételével nagyhatású kulturális tömegmédium szerepét kezdik el betölteni. Az online olvasói létszám nagyságrenddel haladja meg a hagyományos látogatói létszámot. A digitalizálás és a technológiai feladatok miatt a könyvtárak egyre szorosabb kapcsolatba

¹ Információhordozón nem a fizikai adathordozót, hanem a digitális kódot értjük.

kerülnek a piaccal és a műszaki kutatással. A szakértők az elmúlt hónapokban vetették föl a könyvtárak és az internet új modelljének (WEB 2.0) kapcsolatán épülő új könyvtári modellt (Library 2.0).

Az egyre gyorsabb ütemű fejlődés sem tud megbirkózni a digitális vagy digitalizált szövegek tömegének pusztulásával. A történelem során az emberek eleddig a fontos

szövegeket kevés példányban, tartós szöveg-hordozón vagy akár megbízhatatlan szöveg-hordozón, de sok példányban igyekeztek elhelyezni. Most először fordul elő, hogy fontos szövegek kevés példányban, törékeny hordozón léteznek. A művelődés kincseit eddig békében őrző könyvtár, az új tömegmédium, nehezebb helyzetben van, mint valaha.

