Egy szegény elektronikus könyvtáros panaszai

A digitális szövegformátumok problémái

Drótos László
kondrot@gold.uni-miskolc.hu

Elhangzott az MKE Elektronikus Könyvtár Szekciójának az OSZK-ban 2001. március 1-jén tartott rendezvényén, melynek központi témája a szövegdigitalizálás volt.

Ha már előadásom címéül szegény Kosztolányi versciklusának címét torzítottam el, még egy Csokonai parafrázist is megengedek magamnak: *„Az is bolond, aki elektronikus könyvtárossá lesz manapság"*. Annyira kiszámíthatatlan ugyanis a számítógépes dokumentumok jövője, a sokféle szabvány, kvázi szabvány és házi szabvány vetélkedése, hogy lehetetlen okos döntést hozni annak, aki digitális könyvek archiválásával és szolgáltatásával akar foglalkozni. Forrong az egész terület, az elektronikus publikálás egyre nagyobb üzlet és sok új játékost vonz, akik gyakran még azt a néhány játékszabályt is felrúgják, ami nagy nehezen kialakult. A HTML például az elmúlt években teljesen összekuszálódott (vagy ahogy egy amerikai cikk szerzője találóan fogalmazott: „balkanizálódott"), a kiutat ígérő XML lassan terjed, viszont havonta jelennek meg új, semmivel sem kompatibilis e-book formátumok, miközben a konzervatívabb folyóirat-és könyvkiadók legszívesebben a nyomtatotthoz leginkább hasonló PDF-ben publikálnak, ám ugyanakkor szinte mindenki a Word-öt veszi elő, ha szöveget akar írni számítógépen.

Ahhoz, hogy valamennyire a jövőbe tudjunk látni, a múltba érdemes előbb visszanézni, és ha áttekintjük a gépek és szövegszerkesztő programok elmúlt 20-30 éves történetét, azt vesszük észre, hogy a számítógépes szövegek tárolási és megjelenési módját bizony leginkább a mindenkori műszaki (főleg hardver) lehetőségek határozták meg, és nem az, hogy mi lett volna az emberek számára az ideális megoldás (pl. a drága háttértárak és memóriák miatt lett először csak 6, majd 7, később 8 bites az ASCII szabvány, pedig a 12 vagy 16 bites karakter-kódolás az optimális az emberiség számára, ezért fogunk még évtizedekig szenvedni a magyar ékezetekkel; vagy például állítólag a 80 oszlopos lyukkártyák miatt lettek 80 karakteresek az első

képernyők és nyomtatók, pedig az emberi szemnek az 50-60 karakteres sorhosszúság a kényelmes, ahogy ez a nyomtatott könyveknél is van.). Várhatóan továbbra is a technikai változások fogják leginkább befolyásolni az elektronikus dokumentumok sorsát. Hogy a mindenkori „tökéletes megoldás" mennyire függ a műszaki lehetőségektől ezen a téren, azt jól mutatja a *Ted Nelson* nevéhez kötődő *Xanadu projekt,* mely több mint 30 év alatt sem hozott gyakorlati eredményt a tökéletes hipertext rendszer megalkotása terén, mert az elképzeléseket és az elkészült szoftvereket állandóan hozzá kellett igazítani az újabb és újabb számítástechnikai lehetőségekhez, míg végül a kilencvenes évek elején a – sokkal tökéletlenebb – Web megjelenésével és gyors elterjedésével az egész Xanadu fejlesztés értelmét vesztette. A napjainkban zajló e-book forradalom is hasonló fordulatot hozhat: hiába írtak össze annyi minden okos dolgot az elmúlt években arról, hogy milyen is az ideális formátum az elektronikus szövegekhez és hiába születtek a bonyolult, mindenre kiterjedő szabványok, indultak be az ezekre épülő különféle kezdeményezések, könnyen lehet, hogy pár év múlva az emberek egyszerű, hordozható kis célgépeken fognak könyveket olvasni, melyek alig tudnak többet, mint a hagyományos nyomtatott könyvek, igaz nem is drágábbak és a használatukhoz sem kell többet tudni. (Analógia: a 70-es és 80-as évek csúcstechnikájú, precíziós HiFi lemezjátszói helyett ma a jóval gyengébb hangélményt nyújtó MP3 lejátszók terjednek el, olcsóságuk és egyszerűségük miatt.)

Persze, hogy milyen is volna a tökéletes elektronikus könyv, az nagyban függ attól, hogy ki, mit ért ez alatt a fogalom alatt. A formátumokról folytatott viták mögött gyakran a definíció hiánya áll. Mindenféle értelmezés előfordul: az alapvetően nyomtatáshoz formázott, csak éppen számítógépen tárolt és terjesztett könyvektől kezdve, a csak interneten vagy sepciális e-book olvasókon böngészhető dokumentumokon át, a tudományos célokra is alkalmas, tökéletesen feltárt szerkezetű teljes szövegű adatbázisokig. Az én számomra az ideális elektronikus szöveg olyan, mint a „folyadék": könnyen önthető át egyik edényből a másikba, amelynek automatikusan felveszi az alakját, egyszerűen szétosztható vagy egyesíthető, és olcsón, gyorsan továbbítható egyik helyről a másikra, akár a legprimitívebb csatornákon át is.

Néhány konkrét tulajdonság, amelyek a számítógépes formátumú dokumentumoktól jogosan elvárhatók lennének:

Az elektronikus dokumentum legyen...

- böngészhető, átfutható

mindenféle külön művelet (pl. teljes letöltés, visszakódolás, konvertálás) nélkül azonnal bele tudjunk olvasni, lehetséges legyen a szöveg tetszőleges pontjára ugrani, illetve felmérni a terjedelmét, tetszőleges és kényelmesen olvasható külalakot lehessen beállítani hozzá, a vakok által használt programokkal is felolvastatható legyen...

- kereshető

a teljes szövegben kifinomult és gyors keresési lehetőségre van szükség (szekvenciálisan, illetve indexelt adatbázisként), de nemcsak az egyedi dokumentumok szintjén, hanem a dokumentumok tetszőleges méretű halmazán is, akár a felhasználó saját gépén, akár az egész interneten...

- átformálható, konvertálható

az elektronikus szövegnek – hosszú ideig – a legkülönbözőbb alkalmazásokba, mindenféle célra, többféle platformra egyszerűen és rugalmasan átalakíthatónak kell lennie, és miközben csak az elméletileg is elkerülhetetlen információvesztés következhet be...

- módosítható, javítható, jegyzetelhető

a módosíthatóság alapvető különbség a nyomtatott és az elektronikus könyv között, ezt a lehetőséget

mindenképpen biztosítani kell, legfeljebb azzal a megkötéssel, hogy – indokolt esetekben – az eredeti változat mindig előhívható marad...

- másolható, idézhető

az olcsó vagy ingyenes másolhatóság szintén hatalmas előny és a elektronikus könyvek népszerűségének legfőbb oka, azért a magáncélú (a copyright tiszteletben tartásával történő) másolás tiltása értelmetlen, legfeljebb – indokolt esetben – a korlátozott mértékű másolás vagy a tényleges lemásolás nélküli beillesztés lehetősége fogadható el...

- nyomtatható

a papírról való olvasás még nagyon sokáig fennmarad, ezért az elektronikus dokumentumoknál biztosítani kell a jó minőségű nyomtatás lehetőségét, részben vagy egészben a felhasználó által preferált formátumban...

- hivatkozható, azonosítható, hitelesíthető

az egyik legnagyobb hiányosság jelenleg, hogy az elektronikus dokumentumok lelőhelye, önazonossága és hitelessége nem garantálható, ami elsősorban a tudományos szakirodalom terén nehezíti meg az átmenetet a nyomtatott publikálásról az elektronikus felé...

- hordozható, birtokolható, átadható

egy további tulajdonságcsoport, amivel a hagyományos könyv magától értetődően rendelkezik, de az e-book megoldásoknál – főleg jogi okokból – egyre inkább veszélybe kerül: egyes kiadók egy konkrét gépre, személyre, meghatározott időtartamra korlátozzák az olvasás lehetőségét.

Mint sejthető, egyelőre nincs olyan ideális szabvány vagy rendszer, amely a fenti kritériumoknak mind maradéktalanul megfelel. Ez még nem lenne olyan nagy baj, mert legfeljebb több formátummal kell együtt élni és mindig azt választani, ami az adott dokumentumhoz és felhasználási formához a leginkább megfelel. (Ezt az elvet követtük, mi is, a MEK-nél az elmúlt 7 évben és nem bántuk meg.)

Az viszont aggasztó, hogy a jelenlegi trendek épp ellentétes irányba mennek: nemhogy letisztulna a formátumok kavalkádja és kikristályosodnának az ideálishoz legközelebb álló megoldások, inkább csak a káosz nő.

Néhány konkrét formátum és probléma

- A Microsoft Word minden újabb verziója egyre nagyobb inkompatibilitásokat mutat, még a saját korábbi változataival is. A Word formátumot annak idején egyáltalán nem gondolták át, így ma már semmi nem garantálható, ha valaki egy Word dokumentumot egy másik gépen akar megnézni vagy egy másik verzióba akar konvertálni (a Unicode karakterek bevezetése miatt az ékezetes betűkkel is nagy bajok vannak). Pedig a Word sok szempontból megfelelne a felsorolt kívánalmaknak: jók a formázási, másolási, módosítási, keresési és nyomtatási lehetőségei, és ma már ingyenes plug-in is van hozzá a népszerűbb böngésző programokhoz, így on-line is olvasható, és a StarOffice megjelenése óta – elvben – a Linux világban is jelen van.
- Az általános szövegcsere formátumnak tekintett RTF szabvány sem megbízható már, a különböző szoftverek olyan egyéni feltételeket tesznek az RTF-fájlokba, amiket azután már egy másik alkalmazás nem, vagy csak egészen másként mutat meg. Ráadásul nincs (ingyenes) RTF megjelenítő, úgyhogy tényleg csak közvetítőeszköznek jó egyes alkalmazások között.
- A HTML hatalmas változásokon ment át az elmúlt évek során és a Netscape-Explorer vetélkedés sokat ártott neki, bár már régebben is jellemző volt rá a szereptévesztés, mert a fejlesztői

univerzális eszközzé próbálták tenni, ahelyett, hogy megmaradtak volna az eredeti célnál: egy platformfüggetlen egyszerű hipertext formátumnál. A HTML jól böngészhető és kereshető, de pl. hiányzik belőle a jegyzetelés lehetősége, gyakran csak azzal a HTML szerkesztővel módosítható és formázható át, amelyikkel készült, s a sok darabból álló dokumentumok letöltése és nyomtatása is nehézkes.

- Általában az SGML-t tekintik az univerzális csodaszernek az elektronikus szövegek problémáinak megoldására. De az SGML csak egy keretet ad, amit nem könnyű konkrét formába önteni, olcsó és tömeges alkalmazássá fejleszteni. Nem is terjedt el igazából sosem („Sounds Good Maybe Later"), csak 1996-ban sikerült egy ígéretes megvalósítást létrehozni: az XML-t. Az XML rugalmasan konvertálható és kereshető elektronikus könyveket ígér, viszont a fő hangsúlyt a dokumentumok szerkezetére és nem a képernyőn és nyomtatón való megjelenítésre helyezi, ami pedig a felhasználók számára legalább ugyanolyan fontos. Az SGML vagy XML elemeket leíró DTD-k helyes definiálása olyan szakértelmet igényel, amit kevesen fognak elsajátítani, s ez csak tovább növeli a rossz minőségű elektronikus dokumentumok számát. Az egyes szakterületekre kidolgozott és egységesített DTD-k kisebb-nagyobb „szigeteket" hoznak létre az internetes dokumentum-tengerben.
- A PostScript és a belőle kifejlesztett PDF elsősorban a nyomtatásra szánt szövegek ideális formátuma. A PostScript állományokhoz csak gyenge nézegető programok vannak, ezeket igazából egy PostScript nyomtatóra való kiküldésre szánták. A PDF-hez ugyan már van egy jó, ingyenes böngésző és plug-in (Acrobat Reader) mindenféle operációs rendszerre, de a papírhoz való túlságosan erős kötődés ezen is nagyon látszik: a dokumentumot az olvasó már nem formázhatja át, alig lehet másolni és keresni, és a felolvasó programok sem boldogulnak vele. Az Acrobat újabb verziója pedig már saját maga korábbi változataival sem kompatibilis, egyes .pdf állományok csak kis kockákat vagy üres lapokat tartalmaznak, ha nem a megfelelő Readerrel nézzük meg őket, sőt már a nyomtatás sem megbízható néha.
- A nyomdai kiadványszerkesztésre készült szoftverek, mint amilyen a Quark Express, a Ventura, vagy a TeX, a nagyközönség számára élvezhetetlenek, speciális hardvert és (a TeX kivételével) drága szoftvert igényelnek, valamint komoly szakértelmet és még így is a legkülönbözőbb kompatibilitási problémák adódnak. Ezek is elsősorban nyomtatásra szánt formátumok, a velük kiszedett szöveget papírra írva kell terjeszteni.
- Az elmúlt évben hirtelen megszaporodó e-book „szabványok" egyfajta visszalépést jelentenek, bár kétségtelen, hogy többségüknél végre igyekeztek figyelembe venni az elektronikus szövegek olvasási szokásait, az ergonómiai szempontokat, a felhasználók várható igényeit. Mégis, az Open E-book, a Rocket E-book, a Softbook, az Adobe E-book (korábban Glassbook) és társaik minden korábbinál jobban „befagyasztják" a szöveget az adott formátumba és csak néhány alapfunkciót engednek meg az olvasónak (részben a másolástól és az olvasáson kívüli egyéb felhasználástól való félelemből, részben amiatt, mert többségüket eredetileg is kis teljesítményű noteszgépekre vagy speciális célgépekre írták). Mivel az e-book piacon rendkívül éles a verseny, a cégek és a „szabványok" csak kérészéletűek. Ezért a mostanában megjelenő elektronikus könyvek az értük beszedhető gyors nyereségen kívül más, tartósabb értéket nem hordoznak.
- Végezetül szólni kell még az optikai lemezeken kiadott könyvekről, elektronikus dokumentumokról is. A CD-ROM kiadványokhoz jó né-

hány szerkesztő programot fejlesztettek ki az évek során, melyek között vannak teljesen „elvarázsolt", más szoftverrel gyakorlatilag konvertálhatatlan formátumok (pl. Toolbook), és SGML-szerű, így más környezetbe is átvihető, hosszabb életű rendszerek (pl. Folio). Mindenképpen az off-line hordozókon megjelent dokumentumok vannak kitéve leginkább az elavulás és emiatt az olvashatatlanná válás veszélyének (pl. az ABCD esete).

Befejezésként még néhány szó a Magyar Elektronikus Könyvtár (MEK) terveiről, abból az apropóból, hogy remélhetőleg az idén elkezdjük a 2.0-s verzió építését. Szeretnénk minden dokumentumból legalább egy on-line böngészhető (HTML, XML, PDF) és egy összecsomagolt, letölthető (az előbbiek mellett RFT, Word, TeX, PostScript, Open E-book, vagy egyéb) verziót is szolgáltatni. Azoknál a dokumentumoknál, ahol fontos a szerkezet is (pl. bibliográfiák, szótárak, lexikonok), SGL/XML kódolást tervezünk. Ehhez az átalakításhoz az egész állományt át kell nézni (részben selejtezni is) és elkészíteni a szükséges formátumokat azoknál a dokumentumoknál, amelyeknél nincs meg mindkét verzió. (Azért már most is gyakran vannak alternatív formátumok ugyanabból a műből a MEK-ben. Nemrég tettem fel például Illyés Gyulától a „Puszták népé"-t HTML, PDF Word 6 és RTF formátumban, amiket egy finnországi segítőnk készített.) Igyekszünk a legelterjedtebb formátumok optimális „keverékét" megtalálni, hogy legalább ez ne legyen akadálya annak, hogy – Ranganathan 1931-es könyvtári alaptörvényeit a mai korra kiterjesztve – *„minden elektronikus könyv megtalálja a maga olvasóját"* és minden olvasó megtalálja a maga elektronikus könyvét.

A LEGJOBB PUBLIKÁCIÓ(K) 2000-BEN

Évek óta szokás szerkesztőségünkben, hogy a legjobbnak ítélt tanulmány szerzőjét Nívó-díjjal jutalmazzuk. 2000-ben a sok remek cikk közül nagyon nehéz volt választani. Nem is sikerült egyetlen szerző mellett döntenünk, ezért alakult úgy, hogy két szerzőnek adtuk oda e címet, mégpedig *Géró Katalin*nak és *Murányi Péter*nek. Mindkettőjük tanulmánya a folyóirat 2000. 1–2-es számában jelent meg:

GÉRÓ KATALIN:
KNOWLEDGE MANAGEMENT – MÚLÓ HÓBORT AVAGY A JÖVŐNK?

MURÁNYI PÉTER:
CD-ROM-OK (ÉS EGYÉB SZÁMÍTÓGÉPES FORRÁSOK) ÚJ REFERENSZ MÁSODFOKÚ BIBLIOGRÁFIÁKBAN

A szerzőknek szívből gratulálunk!

A szerkesztőség