A könyvdigitalizálás egyes kérdései

A digitális kultúra kialakulása történelmi léptékű társadalmi-gazdasági-kulturális korszakváltást jelent, és az emberiség életében legalább akkora változást idéz elő, mint amekkorát évezredekkel ezelőtt a szóbeliségről az írásbeliségre való áttérés okozott.

Az írás-olvasás önmagában kevés lett volna ahhoz, hogy történelmi korszak alkotójává, társadalomformáló erővé váljon. Ehhez az írásbeliség széleskörű elterjedése kellett – ez teremtette meg az elvont gondolatok szabatos kifejezésének, lejegyzésének módját és egyúttal a leírt szöveg állandóságát, történelemfelettiségét.[1] *Hajnal István* „írásgondolatnak", „írásgondolkodásnak" nevezi azt a gondolkodás- és érintkezéstechnikát, amelyet az írásbeliség alakított ki a gondolatok objektiválása terén. A gondolatok írásban való rögzítése egy idő után elért arra a szintre, hogy az írás már nemcsak eszközként, hanem általánossá vált kifejezési módszerként funkcionált a tudás lejegyzésére szolgáló folyamatban.[2]

Az írásbeliség tette lehetővé, hogy térben és időben eltávolodjék egymástól az információ *kibocsátója* és *befogadója*, illetve a *megismerő* és a *megismerés tárgya*. Az írásbeliség, majd később a könyvkultúra terjedése teremtette meg azt a kommunikációs modellt, amely az absztrakt gondolkodást a tudásátadás egyik legfontosabb momentumává tette.

A kódextől a digitalizált könyvig

Ha alaposan átgondoljuk, csak részben adhatunk igazat annak a széles körben elterjedt nézetnek, miszerint a kultúraközvetítésben Gutenberg találmánya volt az a mérföldkő, amelynek köszönhetően a *nyomtatott könyv* – az európai kultúra eme szimbóluma – félévezredes hódító útjára indult. A könyvnyomtatás ugyanis nem változtatta meg alapvetően sem az információrögzítési módot, sem a tudás áthagyományozására szolgáló legfontosabb információhordozó logikai és fizikai sajátosságait.

Az írásbeliség nyomtatáshoz kötődő korszakában a szöveges információ rögzítése közel hasonlóan megy végbe, akár kézzel, akár nyomdagéppel történik. Az írásjelek az adott írásrendszerre jellemző sorrendben kerülnek az információhordozóra – amely ebben a korszakban jellemzően a *papír.* A latin ábécét használó nyelvek balról jobbra, a sémi írások jobbról balra haladó, vízszintes sorokba rögzítik a betűket – akár kézzel írják, akár nyomdai úton állítják elő a szöveget. A fonetikus ábécét használó nyelvek a kézírásban és a nyomtatásban gyakorlatilag ugyanazokat a betűket és diakritikus írásjeleket használják, ezért – ha eltekintünk az írásjelek formai kialakításától, amely a nyomtatásban szükségszerűen vezetett el a betűformák szabványosításáig –, a nyomtatás és a kézírás többé-kevésbé hasonló írásképet eredményez.

Az európai kultúra átörökítését szolgáló legfontosabb információhordozó a *kódex*, amelynek a formája a Kr.u. I-III. századbeli kialakulása óta lényegében nem változott. A pergamenből készült kódexeket éppúgy, mint a mai könyveket, kötéstáblák közé kötött (fél-, negyed-, nyolcad-, tizenhatodrét), hajtogatott és középen összefűzött lapokra felvágott ívek alkotják. Sem Gutenberg, sem a későbbi modern nyomdatechnika számottevően nem módosította sem az egyes lapok formátumát, sem a használati és olvasási módot. A szöveg a kézbe vehető könyvben az egymás után következő lapok mindkét oldalán, a lapszélektől majdnem egyenlő távolságot üresen hagyó, többnyire álló téglalap formátumú *szövegtükör*ben, szabad szemmel olvasható.

Az információhordozó szerkezetében a digitális korszak hozta meg az igazán jelentős változásokat. *Roger Chartier* "Les métamorhposes du livre" című munkájában arra hívja fel a figyelmet, hogy az informatikai forradalomnak a nyomtatott kultúrára gyakorolt számos hatása közül az a változás a legfontosabb, amely az információhordozó szerkezetében és formájában, illetve a szöveg létrehozására szolgáló technikában következett be.[3] Az analóg világban bizonyos szövegek és a szövegeket tartalmazó tárgyak – vagyis az információhordozók – között kétséget kizáró volt a kapcsolat. Egy napilap vagy egy magánlevél az írásbeli kultúra egyértelműen definiált terméke volt, amelynek értékét és kezelési módját mindenki ismerte – így elmondható, hogy a műfajba sorolás pontosan kijelölte azt a helyet, amelyet az adott információhordozó az elsődleges valóságra épülő rendszerben betöltött.

A virtuális közegben viszont a legkülönbözőbb típusú és minőségű szövegek azonos módon, ugyanazon az eszközön, ugyanabban a formátumban tárulnak elénk, mesterséges egységbe vonva olyan tartalmakat, amelyekkel korábban különböző célból, a magán- és közélet elkülönült helyszínein, eltérő interpretációs közegben találkozhattunk. Az ugyanarról a hordozóról – a képernyőről – folyamatosan áradó, testetlen és végtelen szövegfolyam minden korábbi értéket kiemel a megszokott keretek közül. Mindennek az lett a következménye, hogy az az írásbeli kultúra, amely az analóg világban közvetlenül észlelhető tárgyakban öltött testet, a digitális közegben megkérdőjeleződött, és gyakorlatilag érvényét vesztette.

Az írás kialakulása és a kódex „feltalálása” óta eltelt évezredek alatt – az írásbeli kultúra érvényességi körén belül – lényegileg nem változott meg sem az *írás*, sem az *olvasás antropológiája*. A digitális korszak ebben is alapvető változást hozott. Bár továbbra is a kezünket használjuk az írásjelek lejegyzésére, már nem íróeszközt kézbe véve, hanem billentyűket lenyomva írunk. További jelentős változás, hogy – a papírral ellentétben – az írás nem az általunk létrehozott, állandó formában rögzítődik a számítógépen, hanem valahol a virtuális térben alkot egy bármikor, bármilyenné megváltoztatható karaktersorozatot. Olvasni pedig egyre gyakrabban nem a lapozható, kódex formátumú, nyomtatott könyvekből, hanem a többféle médiumot megjeleníteni képes képernyőn olvasunk.

A digitalizálás további számottevő változást eredményezett a szöveg immanens logikájában. A lineáris szerkezetű, szekvenciális szövegre a *deduktív logika*, a kereszthivatkozásokkal nyitott szerkezetűvé váló hipertext szövegre viszont a *relációs logika* jellemző.

Könyvkultúra, digitális kultúra

A könyv nemcsak kényelmes használhatósága és tartós mivolta, hanem főként a tartalma, illetve az olvasóknak a tartalomhoz fűződő, bonyolult viszonyrendszere miatt vívott ki fontos helyet magának az európai gyökerekre épülő kultúrákban. A gazdaságnak és a társadalomnak egyre több írni-olvasni tudó emberre volt szüksége, akik e tudásukat már nemcsak a munkájukban, hanem a szabadidejükben is kamatoztatták. Amint jelentősebb számú népesség vált könyvolvasóvá, mind több és többféle olvasnivalóra mutatkozott igény és szükséglet, így egyre bővült a könyv-, majd a folyóirat-kiadás, a nyomdaipar, illetve a könyvterjesztés.

Bár a könyvnyomtatás mind szélesebb rétegeket vont be az olvasás kommunikációs aktusába, a mindenkori *olvasók* arányához képest továbbra is kevesen váltak kitüntetett szerepű *szerzővé* – akárcsak az írásbeliség korábbi korszakaiban. Annak ellenére, hogy a könyvkultúra terjedése határozott demokratikus jegyeket mutatott föl, mégis megteremtette a *kiválasztottak*, a műveik kinyomtatására méltó, felmagasztosult *szerzők* mítoszát, ezáltal egyfajta kirekesztettségbe kényszerítve a társadalom nagy többségét. Ez utóbbiak számára ebben a szereposztásban egyetlen pozitív lehetőség: a *művelt, értő olvasóvá* nevelődés kínálkozott, akire "fogyasztóként", vásárlóként bizton számíthatott szerző és kiadója.

A társadalmi munkamegosztásban kialakult egy professzionális szervezet – előbb a *nyomda*, majd a *kiadó* –, amely hivatásszerűen foglalkozott a műveknek az olvasóközönséghez való eljuttatásával. A tudományos élet nem fejlődhetett volna igazán a rangos kiadóknál megjelenő, filológiailag gondozott szövegeket tartalmazó, megkülönböztetett értéket hordozó, hiteles forrásként kezelt könyvek nélkül, amelyekre tudományos körökben hivatkozni illett, sőt kellett. Az a „hozzáadott érték”, amely a jó nevű kiadók által nyújtott, megbízható minőségben nyilvánult meg a kézirat előkészítésétől kezdve a színvonalas nyomdatechnológia alkalmazásáig, a könyvvásárlók körében kialakította azt a *bizalmi hozzáállást*, amely az európai kultúra átörökítésében alapvető fontossággal bírt. Mindezen évszázados folyamatok következtében a szerző, a kiadó és az olvasó között kialakult a szövegek megbízhatóságára épülő *fiduciárius* viszony, amely az olvasó számára egyfajta minőségi biztosítékot szavatolva hivatkozási alapot teremtett.

A bizalom más vonatkozásban is megnyilvánult a könyv és olvasója között. Igen kevés kivételtől eltekintve bízni lehetett abban, hogy egy hivatásos kiadó jogszerűen kezelte a kéziratot. Ha a vásárló kifizette egy példány vételárát, akkor a birtokába került *példányra* vonatkozóan bizonyos korlátozott jogokra is szert tett: például kölcsönadhatta másnak elolvasni, vagy eladhatta az antikváriumban. A kinyomtatott példányokból

nemcsak magánemberek, hanem könyvtárak is vásárolhattak, ezeket a példányokat a beiratkozott olvasók legálisan kikölcsönözhették.

Az internetes közegben ez a bizalmi viszony is megkérdőjeleződött: a felhasználó gyakran nem tudja, használhatja-e legálisan, hivatkozhatja-e megbízható forrásként az ott talált szöveget? Sajnos, igen gyakran abban sem lehetünk biztosak, szerzői jogi szempontból jogszerűen került-e a hálózatra a szóban forgó mű, amelyet letöltve nem válunk-e egy jogszerűtlen cselekmény részeseivé akaratunkon kívül.

A könyv az elsődleges és a másodlagos valóságban

Hosszú évszázadokon keresztül a könyv egyrészt a benne rögzített tartalmat jelentette, másrészt azonban a fizikai valóságban létező tárgyként is funkcionált.

A könyv fogalmához hozzátartozott, hogy egy bizonyos *terjedelmet* (lapszámot), illetve *példányszámot* el kellett érnie, különben nem minősült könyvnek. A 80-as években kiadott szabvány így határozta meg a könyv fogalmát: „... olyan 48 oldalnál nagyobb terjedelmű nyomdatermék, amely két fedőlapból, valamint meghatározott sorrendben egymást követő [...], a gerincen tartósan összeerősített belső lapokból áll, és olvasható szöveget, ill. illusztrációt tartalmaz."[1]

A nyomtatott könyv a szerző, illetve a kiadó nevével és külső adottságaival is hordoz egyfajta üzenetet. E téren szintén kialakult egy szokásrendszer: drága papíron, bőrbe kötve csak az értékes, időtálló szöveget érdemes kinyomtatni; a „filléres" könyv a külső megjelenésével is elárulja, hogy nem örök igazságokat közlő tartalmat, hanem feledhető, könnyed szórakozást kínál az olvasónak.

A nyomtatott művek fizikai megjelenése azonban nemcsak minőségjelzőként funkcionál. A könyv tárgyi mivoltát meghatározó, az elsődleges valóságban érzékelhető tulajdonságai – mint a könyvtest formátuma és vastagsága, a borító kialakítása, a gerinc kiképzése – üzenethordozóként működnek a könyv használati funkcióiról.[2]

Az elsődleges fizikai valóságban létező könyv csak a maga tárgyi valóságában tudott eljutni az olvasókhoz, ezért a *könyvterjesztés* fontos eleme a szerző és az olvasó közötti értéklánc.

Újabban beszélünk ugyan „elektronikus", „digitalizált" stb. könyvről; de tudván tudjuk, hogy a kifejezésből csak a jelző valóságos, a jelzett tárgyra nem illik a „könyv" megnevezés. Az e-könyv tartalma elválik a hordozójától; a szöveg „testetlenül" lebeg valahol a virtuális térben. A digitális könyv nincs jelen a fizikai valóságban, és a könyv térbeli

[1] MI 5602-83

[2] Az okfejtés alátámasztására szolgáló néhány példa: zsebszótár, útikönyv, művészeti album stb.

kiterjedésével ellentétben, a képernyőn csak két dimenzióban jelenítődik meg. Mindez azt sugallja, hogy a szöveg állandóan létezik valahol, akkor is, ha az emberi érzékszervek számára az adott pillanatban elérhetetlen.

Az elsődleges valóságban létező írásbeli kultúra termékeit a rendelkezésre álló fogalmi készletekkel le lehetett írni – a digitalizált művekre viszont nincsenek adekvát kifejezéseink. *„Elektronikus könyv"*-nek nevezzük a nyomtatott könyv digitalizált változatát, amelyet kézbe sem lehet venni, *„digitális könyvtár"*-nak az elektronikus dokumentumok gyűjteményét, ahova be sem lehet menni, és *„lapozunk"* a képernyőn, ahol nincsenek is lapok... Több évtizede nem sikerült adekvát megnevezéseket kitalálnunk sem az új információhordozókra, sem a virtuális térben lévő szövegekre és más információegységekre.

A digitalizálás fogalmi keretei

Annak ellenére, hogy az eredmény mindkét esetben egy *digitális állomány*, a feldolgozási folyamat eltérő sajátosságai, illetve a szerzői jogi előírások miatt meg kell különböztetnünk egymástól a *digitális formában létrejövő (born digital)*, illetve a *digitalizált (digitized)* dokumentumokat. A digitális dokumentumok egyre nagyobb hányada eleve valamilyen számítógépes eljárással készül, tehát *digitális formában jön létre.* A digitalizálás során viszont a *korábban más hordozón megjelent műveket* valamilyen *digitalizáló eszközzel átkódoljuk* a számítógép nyelvére, illetve *rögzítjük* egy számítógéppel olvasható, adattároló eszközre. Az eredeti mű hordozója lehet papír, bakelit lemez, celluloid szalag stb., a rögzített információ lehet szöveg, hang, álló- vagy mozgókép, illetve ezek együttese.

A cikkünk tárgyát képező *könyvdigitalizálás* azoknak az eszközöknek, módszereknek, eljárásoknak az összességét jelenti, amelyek segítségével az analóg eljárással nyomtatott dokumentumról a számítógép által kezelhető, digitális jelek sorozata jön létre.[3] A digitalizálás során az analóg jeleket valamilyen digitalizáló eszközzel alakítják át a számítógép által olvasható jelekké (kódokká). Más szavakkal úgy is mondhatjuk, hogy *a digitalizálás eredménye az analóg nyomat számítógépes reprezentációja*.

A digitalizáló eszközök a digitalizálás tárgyát képező forrásművek információtartalmának csak egy részét képesek bináris kódokra áttenni, így bizonyos értelemben a digitalizált állomány információtartalma az eredeti forrásénál kevesebb. Más vonatkozásban viszont – a forrásmű információtartalmán túl – a digitális változathoz olyan további funkciókat is rendelhetünk, amelyek az analóg változathoz képest értéktöbbletet eredményeznek. Erre jó példa lehet egy madártani könyv, amelynek az egyes fajok hangjával kiegészített digitális változatából sokkal könnyebb a

[3] Nyomdatechnikai értelemben az analóg eljárás azt jelenti, hogy az adott felületen a nyomóformát egyidejűleg alakítják ki – szemben a digitális módszerrel, melynek során a nyomóforma pontonként (esetleg soronként) készül.

madárfajok felismerését elsajátítani, mint a nyomtatott könyvekben olvasható hangutánzó szavak alapján.

A digitalizálási folyamat bemeneti (input) oldalán az *eredeti* mű (a *forrásmű)* – kimeneti (output) oldalán pedig a *számítógépes reprezentáció* (a *digitalizált állomány)* áll.

Ha a digitális változat tulajdonságait az eredeti műhöz viszonyítjuk, három szintet különböztethetünk meg:

A *reproduktív szint* a forrásmű formai és tartalmi jegyeit egyaránt tükrözteti (az esetleges hibákkal, eltérésekkel együtt). A digitalizált változat az eredeti művel gyakorlatilag egyező hatást vált ki, azzal szinte egyenértékű. Ebbe a csoportba elsősorban a fakszimile állományok (képfájlok) tartoznak.

A *reprezentatív szint* a forrásmű tartalmát helyezi előtérbe, de alapvetően nem változtatja meg a szöveg lineáris olvasatát. Ezen a szinten az analóg szövegből digitalizált szöveget állítunk elő, amelynek információtartalma a számítógép nyújtotta szokásos eszközökkel könnyebben kereshető.

Az *interpretatív szinten* az eredeti forrás tartalmához hozzáadódik a feldolgozást végző szakemberek tudása és tapasztalata, melynek eredményeként új minőség jön létre. Az eredeti művet kiegészítő elemek (amelyek lehetnek magyarázatok, mutatók, hipertext hivatkozások, vagy a szövegtől eltérő műfajú elemek: hang- és videofájlok stb.) megbontják az eredeti szöveg lineáris egységét.

Ha a fent vázolt három szintet összevetjük a digitalizálás leggyakoribb forrásául szolgáló hagyományos könyvekkel, a következő eltéréseket állapíthatjuk meg. Az első szinten nincs lényegi különbség a nyomtatott könyv, valamint a csak képként megtekinthető és lapozható digitális állomány között. A második szint olyan keresési lehetőségeket kínál föl, amelyeket a nyomtatott könyv legföljebb csak részben tud nyújtani. A harmadik szinten a forrásmű szövege új dimenzióba kerül: a lineáris olvasatot megtöri a hivatkozásként beillesztett számtalan új elem, amelynek következtében a digitalizált könyv nem lesz többé homogén összetevőkből felépülő, egységes, lezárt egész, hanem egy nyitott struktúrájú, heterogén alkotóelemekből álló halmazzá válik, amelynek pontos határait már nem is lehet megvonni a reá mutató, illetve a belőle kilépő hipertext kapcsolatok rendszerén belül.

A nyomtatott könyv – amelynek tartalma bármilyen sokrétűen van strukturálva, indexelve – belső tulajdonságainál fogva *statikus*. Az előre kitalált szerkezeti felépítést, az oldaltükröt, a tartalomjegyzéket, indexeket, hivatkozásokat, utalókat a nyomtatás után már nem lehet megváltoztatni, az esetleges hibákat nem lehet kijavítani. Ugyanez igaz a sokszorosítási eljárással készülő CD-ROM-okon[4] publikált művekre is. A hálózaton keresztül elérhető művek viszont többé-kevésbé dinamikusak,

[4] A CD-ROM neve (Compact Disc - Read Only Memory) éppen arra utal, hogy a rajta lévő információkat csak olvasni lehet, szerkeszteni, megváltoztatni nem.

hiszen a szolgáltató szervereken tárolt állományok képernyőn való megjelenése a kliens oldali számítógép beállításától, illetve a felhasználó által futtatott programoktól is függ.

A szövegdigitalizálás módszerei

A szövegek digitalizálására használatos eszköztár gyakorlatilag a számítógép-billentyűzetre és a szkennerre korlátozódik. Ebből adódóan a nyomtatott szövegek digitalizálására szolgáló két fő módszer a *begépelés*, illetve a *szkennelés*. A kétfajta művelet eredményeként létrejövő állomány közötti lényegi különbség: amíg a gépelés eredményeként számítógéppel olvasható szöveg jön létre, addig a szkennelés eredménye egy képfájl.

Digitalizált szöveg előállítása

A *szöveg leírása* során a billentyűzeten keresztül rögzített karakterek (betűk, jelek, szóközök stb.) mindegyike külön-külön kódot kap, amelyek egymásutánja egy karakterláncot alkot. Az egymás után következő karakter-kódok alapján lehet az ún. teljes szövegű keresés során visszakeresni az ily módon leírt szöveget.

A *szkennelés* eredményeként a digitalizált oldal *képe*, az eredeti oldal hű leképezése jön létre. A szkenner által létrehozott képfájl a hagyományos nyomdatechnikában ismert fakszimile másolatokra hasonlít. A szkennelt képfájlt látva az emberi agy felismeri a szöveget, a számítógép viszont a képen látható információkat nem képes szövegként értelmezni.

Amennyiben a digitalizálás célja *számítógéppel olvasható szöveg* előállítása, akkor szükség van a szkennelt képek konvertálására, vagyis a képi elemekként tárolt információk karakterekre történő kódolására. E célra speciális szövegfelismerő szoftvereket[5] fejlesztettek ki, amelyek a képfájlon végighaladva a képpontok eloszlását hasonlítják össze azzal a mintázattal, amelyet a program az adott karakterkészletről tárol. A képfájlban található pontok és a memóriában tárolt karakterkészlet összevetésének eredményeként egy szöveg-imitáció áll elő. A szövegfelismertetés következő fázisa a karakterláncok értelmes szavakká alakítása.

A gyakorlatban az OCR technológia alkalmazásának legmunkaigényesebb fázisa a számítógép által előállított szöveg korrektúrázása és javítása. Érdemes tudni, hogy a karakterfelismerő szoftverek a lézernyomtatóval, famentes papírra, folyó szövegként kinyomtatott, mai helyesírású szövegekre vannak optimalizálva. Amennyiben a digitalizálandó dokumentum nem felel meg ezeknek a kritériumoknak, a karakterfelismertetés során jelentős minőségromlás áll be. Arra a kérdésre, hogy mikor melyik digitalizálási módszert érdemesebb alkalmazni, nincs általános szabály, de a tapasztalat szerint a régies

[5] Optical Character Recognition, OCR

helyesírású, vagy sok idegen szót, vagy különleges tipográfiai elemeket (például sok dőlt betűt, vagy hasábokra tördelést) tartalmazó szöveget érdemesebb begépeltetni, mintsem a karakterfelismerés után korrektúráztatni.

A digitalizálási folyamat célrendszere

Maga a digitalizálás nem túlságosan bonyolult folyamat, előkészítése azonban igen nagy körültekintést igényel. A megvalósítás előtt végig kell gondolni azokat a legfőbb szempontokat, amelyek segítségével pontosan meg lehet határozni a digitalizálás célrendszerét.

A digitalizálás legfontosabb indítékai általában a következők:

- *értékmentés, állományvédelem, állagmegóvás* – amely többnyire az elöregedett hordozók tartalmának átmentése, illetve az értékes eredeti dokumentumok állapotának megőrzése érdekében történik;
- *archiválás*, amelynek célja a digitalizált állomány hosszú távú megőrzése;
- *nyilvános szolgáltatás* esetén a digitalizálási cél lehet a nyomtatott formában egyáltalán nem, vagy csak nehezen hozzáférhető, de közérdeklődésre számot tartó könyvek és más dokumentumok hozzáférhetővé tétele;
- *jövedelemszerzés*, amely irányulhat a digitalizált változat értékesítésére vagy a digitalizált tartalom által fölkeltett érdeklődés reklámpiaci értékesítésére;
- *reprodukálás*, melynek során az eredeti dokumentumot újra publikálható minőségben digitalizálják;
- *on-demand szolgáltatás*, amelynek keretében konkrét megrendelésre digitalizálnak.

A döntéshozatalt meghatározó fontosabb szempontok

Először a *felhasználói célcsoportot* kell meghatározni – ennek alapján tisztázhatók a *szerzői jogi törvény* feltételei. Az alábbi három fő célcsoport, illetve cél közül lehet választani:

- magánszemély, aki saját magának digitalizál;
- intézmény, amely belső célokra digitalizál;
- tartalomszolgáltató, amely a nagyközönség számára nyújtandó szolgáltatás érdekében végzi a digitalizálást.

Szerzői jogi szempontból a tartalomszolgáltatásnak igen szigorú előfeltételei vannak: addig nem szabad, illetve nem érdemes a digitalizáláshoz hozzákezdeni, amíg a szerzői jogi feltételek nem rendezettek. A szerzői jogi szabályok szerint *a digitalizálás a mű többszörözésének minősül*, amelynek engedélyezése a szerző kizárólagos joga – ezért minden esetben először azt kell megvizsgálni: a szerzői jog által védett műről van-e szó? Ha nem, akkor nincs akadálya a digitalizálásnak. Ha igen, akkor fel kell kutatni a szerző(ke)t (illetve a jogtulajdonosokat), akikkel felhasználási szerződést kell kötni.

A digitalizálás döntési folyamatában fontos szempont a digitalizált mű közzétételi *időtartamának* a meghatározása. Más technológiát kell választani a *hosszú távú* megőrzés, illetve a *rövid távra* tervezett szolgáltatás esetében.

A *prioritási sorrend* a digitalizálás céljainak ismeretében fogalmazható meg. A döntés során meg kell határozni, hogy a *legértékesebb*, a *legnagyobb érdeklődésre számot tartó*, a *legkutatottabb*, a *legveszélyeztetettebb* stb. dokumentumok részesülnek-e előnyben, de a digitalizálandó forrásmű kiválasztásában további *tudományos, gyakorlati, üzleti* stb. szempontok egyaránt érvényesülhetnek.

A döntési folyamat egyik legnehezebb kérdése a *szelekció*, vagyis a digitalizálandó forrásmű kiválasztása, amely az egész tartalomszolgáltatási rendszer minőségét, a szolgáltatást igénybe vevők körét, a szükséges erőforrások nagyságát, a hosszú távú tervezést és minden további fontos összetevő mibenlétét meghatározza.

A tartalomszolgáltatás minőségét meghatározó szempontok közül a legfontosabbak:

- a digitalizált szöveg minősége, a megengedett hibák aránya;
- a letöltést, nyomtatást, másolást stb. lehetővé tevő megoldások alkalmazása;
- a szöveg egyes elemeinek kereshetővé tétele;
- a közzétételre szolgáló adathordozó típusa;
- a digitalizált mű azonosító adatainak megadása.

A tudományos irodalom reprezentációjára szolgáló formátumok és jelölőrendszerek

A szöveget három szinten: *formai* (layout), *logikai* (szintaktikai) és *tartalmi* (szemantikai) megközelítésben lehet értelmezni. A tudományos irodalom digitális reprezentációjára legalkalmasabb feldolgozási módszer kiválasztásához lényeges tudni, hogy vannak olyan szövegformátumok, amelyek csak a formai adottságokat, mások pedig a szintaktikai és szemantikai elemeket is tudják kezelni. Az előző fejezetben leírt szövegdigitalizálási eljárások közül a szkennelés eredményeként létrejövő képfájlt sem logikai, sem tartalmi szinten nem lehet értelmezni – ehhez a szöveg számítógépes kódolására van szükség.

Az interneten található szövegfájl-formátumok közül a leggyakoribb a HTML, a PDF, az XML, a képfájlok közül pedig a JPG és a TIFF. A ma leginkább elterjedt HTML formátum a szövegnek csak a formai sajátosságait tudja kezelni, így nem alkalmas a szövegelemek *minősített keresésére*, sem a számítógépes hardver- és szoftvereszközök adottságaitól független, széleskörű felhasználásra.

A szövegszerkesztő programok és a HTML-t kezelő Web-böngészők az ún. teljes szövegű (full text) keresésre alkalmasak; ekkor a számítógép karakterről karakterre hasonlítja össze a keresőkérdést a szöveggel, és csak a megegyező karakterláncot értelmezi találatként. A „minősített

keresés” során viszont – még a dokumentum digitalizálása előtt – egy előre kidolgozott séma alapján megjelölik azokat a szövegelemeket, amelyeket kereshetővé akarnak tenni. A számítógép a szövegben elhelyezett jelölők alapján találja meg a meghatározott elemeket. Ha például egy szövegben fontos az összes név kereshetősége, minden név elé beillesztik a <*name*> jelölőt, így megtalálhatóvá válik az összes Kiss István, Nagy Pista, Julcsi stb. név. Ha azonban külön-külön akarják kezelni a vezeték-, a kereszt- és a beceneveket, három jelölőt alkalmaznak: <*familyname*> <*forename*> <*nickname*>.[6]

A szemantikai információk visszakereshetővé tételére fejlesztették ki az SGML szabványt[7], amelyet 1986-ban fogadtak el. Az SGML-t azért hívják jelölő nyelvnek, mert a szabvány segítségével a szöveg minden fontosnak ítélt elemét meg lehet jelölni, és a jelölés alapján visszakereshetővé lehet tenni.

Az SGML alkalmazását megelőzően ki kell dolgozni a tartalmi elemek jelölésének módját, rögzíteni kell a különböző információtípusok közötti kapcsolatokat, valamint a dokumentum struktúrájára vonatkozó szabályszerűségeket. Azt is előre meg kell határozni, a dokumentumban mely elemek kötelezőek és melyek opcionálisak. A dokumentum struktúrájára jellemző szabályokat előre meg kell fogalmazni, és le kell írni a dokumentum-típus definícióban (Document Type Definition – DTD). Az SGML alkalmazásokban a DTD nem más, mint az egyes szövegtípusok (ez lehet például szabadalmi leírás, vers, dráma stb.) *szövegmodellje*.

Az SGML állományok nem tartalmazzák a dokumentumok formai jegyeit. Az egyes dokumentumtípusok megjelenítésével kapcsolatos valamennyi fontos információt részben a DTD fájlokban, részben a külön definiálandó stíluslapokban kell megadni. A dokumentumok megjelenítésére külön szabvány[8] szolgál.[9]

Az elmúlt két évtized során számos tudományterületre és annak jellemző dokumentumtípusaira kidolgozták a speciális SGML alkalmazásokat, a világot mégis csak az 1998-ban napvilágot látott XML[10] változat hódította meg, amely érvényesíti az SGML előnyeit, de igyekszik kiküszöbölni annak hátrányait. Annak ellenére, hogy sokkal több előkészületet igényel, és számottevő az élőmunka-ráfordítás igénye, a nagy értékű tudományos munkák digitális feldolgozása során érdemes az SGML szabványt, vagy annak legújabb „leszármazottját”, az XML-t alkalmazni.

---

[6] Például: <*familyname*>Kiss, <*familyname*>Nagy, <*nickname*>Julcsi stb.

[7] Standard Generalized Markup Language, ISO 8879:1986

[8] DSSSL Document Style and Semantics Specification Language, ISO/IEC 10179:1996

[9] Egy rövid példa arra, hogyan működik a dokumentum-formázás az SGML szabványcsaláddal. A DTD táblában definiáltuk a ’vers’ dokumentumtípust; ezt a szabvány előírta konvenció alapján jelöljük az SGML fájlban. A stíluslapon meghatározzuk, hogy a képernyőn a felhasználó gépének beállításától függően látható virtuális „lap” függőleges optikai középvonalához igazodjanak a címek, a verssorok – így a képernyőn a nyomtatásban megszokotthoz hasonló látványban lesz részünk. Ha olyan stíluslapot alkalmaznánk, amelyen a „vers” nincs definiálva, nem tudnánk ezt a „tipográfiai” hatást elérni.

[10] Extensible Markup Language – http://www.w3.org/XML/

A digitalizált állomány megőrzésének kérdései

Az informatikai hardver- és szoftvereszközök rendkívül gyorsan elavulnak, ezért a ma rendelkezésre álló digitalizálási eljárások eredményeként létrejövő számítógépes állományok várható élettartama igen rövid. A gyors technológiai avulás következtében a digitalizálás egyik kulcskérdése a megőrzés, illetve a tartalomszolgáltatás tervezett időtartama.

A digitalizált állomány megőrzése részben a fizikai, részben a technikai környezettel szemben támaszt követelményeket. Fizikailag biztosítani kell a tárolóeszközök védelmét a valós és virtuális veszélyek ellen (tűz- és vízkár, betörés- és vírusvédelem stb.), technikailag pedig karban kell tartani a tárolóeszközöket (beleértve az adatellenőrzést, és szükség esetén az egyik hordozóról a másikra való átírást). Kívánatos a dokumentumok azonosító adatainak, a metaadatoknak időnkénti ellenőrzése és karbantartása.

A szerzői jogok védelme a digitalizált művek esetében

A jelentős ráfordítással digitalizált művek illegális felhasználása ellen a tartalomszolgáltatóknak részben a saját érdekükben, részben a jogtulajdonosok érdekében védeniük kell a szellemi alkotásokat.

A digitálisan hozzáférhető állományok szerzői jogvédelmére a hagyományos eszközök nem alkalmasak, ezért e célra informatikai megoldásokat fejlesztettek ki. A digitális tartalmakhoz való hozzáférést lehetővé tevő, illetve a hozzáférést szabályozó technikai, műszaki, hardver- és szoftvereszközök összefoglaló neve: *digitális jogkezelés* (Digital Rights Management, DRM).

A különböző DRM-technológiák a szerzői jog által védett digitális tartalom meghatározására, azonosítására szolgálnak, és biztosítják a törvény által előírt szabályok betartását, illetve betartatását. A DRM a jogvédelem alatt álló digitális tartalmak illegális terjesztése ellen kifejlesztett olyan műszaki eljárások komplex rendszere, amely *korlátozza*, illetve *megakadályozza* a jogvédelem alatt álló tartalmakhoz a *jogosulatlan hozzáférést*, illetve biztosítja a felhasználás *engedélyezését,* a jogosulttól a felhasználóig a *tartalomátvitelt* és a felhasználási díj *elszámolását*.

*

Európa kulturális és tudományos tudáskincsét mostanáig főként a nyomtatott források őrizték meg az egymást követő generációk számára. Az európai a kulturális vagyon és a társadalmi emlékezet számottevő hányadát e források alkotják, ezért digitalizálásuk létfontosságú Európa kulturális sokszínűségének fenntartásában és népszerűsítésében.

Tudjuk, hogy nagyon fontos a tartalom átmentését szolgáló reproduktív digitalizálás, de még fontosabb az a hozzáadott érték, amely az eddig

külön síkon létező és élvezhető műfajok együttes alkalmazásában, illetve az eddig rejtve maradt szemantikai kapcsolatok mentén létrejövő, asszociatív elágazások kifejtésében ölt testet, új dimenzióba helyezve a korábban nyomtatásban napvilágot látott szövegeket.

Tószegi Zsuzsanna PhD

Irodalom

[1] Parragh Szabolcs: Ms 5386/9-10 – Hajnal újkora. Budapest, 2002. október. http://parszab.nir.hu/letoltes/ms5386.pdf

[2] Kondor Zsuzsanna: A kreativitás mintázata: Hajnal István.

http://zeus.phil-inst.hu/recepcio/htm/3/308_belso.htm

[3] Roger Chartier: Les méthamorphoses du livre. A könyv teljes szövege letölthető a következő címről: http://editionsdelabibliotheque.bpi.fr/livre