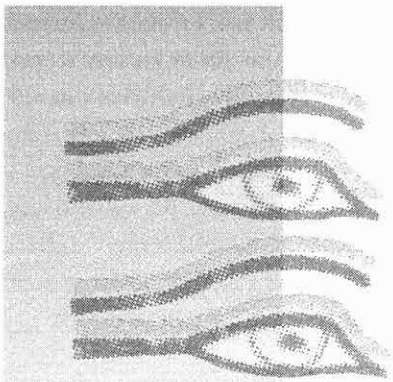

TANULMÁNYOK

A szövegdigitalizálás döntési folyamata

TÓSZEGI Zsuzsanna

Az utóbbi évtizedben gyakran találkozhattunk a hazai könyvtári szaksajtóban a digitalizálás módszereivel, eszközeivel, eredményeivel, de kevés olyan publikáció jelent meg, amely a napi gyakorlat, és azon belül is az eldöntendő kérdések felől közelíti meg ezt a sok összetevőből álló folyamatot. Az alábbiakban a nyomtatott művek digitalizálása során számba veendő és mérlegelendő szempontokat vesszük sorra – azt remélve, hogy a döntési folyamat bemutatásával a könyvtáros kollégák segítségére lehetünk a digitalizálási feladatok ellátásában.

A fogalmi keretek

Bár 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 eljárás-

sal készül, tehát *digitális formában jön létre*. A digitalizálás során viszont a *korábban más hordozón megjelent műveket* valamilyen *digitalizáló eszközzel kódoljuk át* a számítógép nyelvére, illetve *rögzítjük* számítógéppel olvasható adattároló eszközre. Az eredeti mű hordozója lehet papír, bakelit lemez, celluloid szalag stb., a rögzített információ lehet szöveg, hang, állóvagy mozgókép, illetve ezek együttese.

A cikkünk tárgyát képező *szövegdigitalizálás* azoknak az eszközöknek, módszereknek, eljárásoknak az összességét jelenti, amelyek segítségével az analóg eljárással nyomtatott dokumentumról a számítógép által kezelhető, digitális jelek sorozata jön létre.[1] A digitalizálás során az analóg jeleket valamilyen digitalizáló eszközzel alakítják át a számítógép által olvasható jelekké (kódokká). Más szavakkal azt is mondhatjuk, hogy *a digitalizálás eredményeként az analóg nyomat számítógépes reprezentációja jön létre.*

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

A szövegek digitalizálására használatos eszköztár igen szűkös – gyakorlatilag a számítógépbillentyűzetre és a szkennerre korlátozódik. A digitalizáló eszközök azonban 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.

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 egyenértékű, azzal egyező hatást vált ki. Ebbe a csoportba elsősorban a fakszimile állományok (képfájlok) tartoznak. A bibliográfiai feldolgozás során a digitalizált változat azonosítói között mindenképpen meg kell adni a forrásmű adatait.
- 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 eredeti és a digitalizált dokumentum közötti bibliográfiai kapcsolatra hivatkozni kell.
- 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. Katalogizáláskor az azonosító adatok között utalni kell az eredeti műre.

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 mű 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 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[2] publikált művekre is. A hálózaton keresztül elérhető művek viszont többé-kevésbé dinamikusak, hiszen a szolgáltató szervereken tárolt állományok képernyőn való megjelenése a kliens oldali számítógép beállításától, illetve a felhasználó által futtatott programoktól is függ.

A digitalizálási folyamat célrendszere

Maga a digitalizálás nem túlságosan bonyolult folyamat, előkészítése azonban igen nagy körültekintést igényel. Mielőtt a megvalósításhoz hozzákezdenénk, végig kell gondolnunk azokat a legfőbb szempontokat, amelyek segítségével pontosan meg tudjuk határozni a digitalizálás célrendszerét.

A digitalizálás célja és a digitalizálhatóság

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 dokumentumok elérhetővé tétele a nagyközönség számára;
- *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 digitalizáljuk;
- *on-demand szolgáltatás*, amelynek keretében konkrét megrendelésre digitalizálunk.

Az alábbiakban a döntéshozatal során figyelembe veendő legfontosabb szempontokat vesszük sorra, a „kinek, miért, mit, hogyan digitalizáljunk?" leegyszerűsített kérdések köré csoportosítva.

A „kinek és miért digitalizálunk?" kérdéseket csak együtt érdemes föltennünk, ugyanis ha meghatározzuk a *felhasználói célcsoportot*, az is eldől, hogy a *szerzői jogi törvény* alapján milyen feltételekkel digitalizálhatjuk a szóban forgó művet. E tekintetben három fő célcsoport, illetve cél közül választhatunk:

1) magánszemély – saját maga, saját céljaira digitalizál;
2) intézményi, belső célokra digitalizálunk;
3) a nagyközönség számára, tartalomszolgáltatási céllal digitalizálunk.

A szerzői jogi szabályokra a következő fejezetben térünk vissza, itt csak annyit jegyzünk meg, hogy a harmadik pontban jelzett szolgáltatásnak igen szigorú előfeltételei vannak: addig nem szabad, illetve nem érdemes a digitalizáláshoz hozzákezdenünk, amíg a szerzői jogi feltételek nem tisztázódnak.

A „miért digitalizálunk?" kérdésre adott válaszok között annak is helyet kell kapnia, milyen *időtartamra* tervezzük a digitalizált mű közzétételét. Más technológiát kell választanunk, ha *hosszútávú* megőrzésre szánjuk a digitalizált szöveget, mint ha csak *rövidtávra* tervezzük az adott szolgáltatást.

A *prioritási sorrend* is csak a digitalizálás céljainak ismeretében fogalmazható meg. A döntés során meg kell határoznunk, hogy a *legértékesebb*, a *legnagyobb érdeklődésre számot tartó*, a *legkutatottabb*, vagy a *legveszélyeztetettebb* dokumentumokat részesítjük-e előnyben. A „mit digitalizáljunk?" kérdésre adott válaszokban *tudományos, gyakorlati, üzleti* stb. szempontok egyaránt érvényesülhetnek.

Ha a *szelekció* nehéz kérdésén túljutottunk, következik a digitalizálandó mű elemzése az *állományvédelmi szempontok* alapján, majd meg kell vizsgálni a digitalizálandó szöveg *adathordozójának fizikai adottságait*, illetve a *forrásmű szövegének jellemzőit.*

Végezetül a tartalomszolgáltatás minőségét alapvetően meghatározó, a „hogyan digitalizáljunk?" kérdéskörbe tartozó szempontokat kell sorra vennünk, hogy válaszolni tudjunk az alábbi kérdésekre:

- Milyen feldolgozási minőséget akarunk garantálni?
- Kívánunk-e letöltési, nyomtatási, másolási stb. lehetőséget adni a felhasználóknak?
- Kereshetővé akarjuk-e tenni a szöveg egyes elemeit?
- Milyen adathordozón tesszük közzé a szöveget?
- Hogyan akarjuk megtalálhatóvá tenni a digitalizált művet?

A felsorolt szempontokra a későbbiekben visszatérünk, ugyanis a „kinek, miért, mit, hogyan digitalizáljunk?" kérdésekre adott válaszok jelölik ki a feladat minőségi és mennyiségi mutatóit és határozzák meg a digitalizálás erőforrásigényét.

A szerzői jogi szempontok

Nem lehet eléggé hangsúlyozni, hogy a digitalizálás szempontjából a szerzői jogi kérdések a legfontosabbak közé tartoznak. Egy szerzői joggal védett mű esetében, ha nem kapjuk meg a jogtulajdonos hozzájárulását, akkor sem az interneten, sem CD-ROM-on, sem más hordozón nem adhatjuk közre a digitalizált művet – márpedig a korábban nyomtatásban megjelent műveket főként azért digitalizáljuk, hogy az interneten vagy CD-ROM-on hozzáférhetővé tegyük őket a nagyközönség számára.

A szerzői jogi szabályok szerint a *digitalizálás a mű többszörözésének minősül*, amelynek engedélyezése a szerző kizárólagos joga – ezért minden esetben először azt kell megvizsgálnunk: a szerzői jog szempontjából védett műről van-e szó? Ha nem, akkor nincs akadálya a digitalizálásnak. Ha igen, akkor fel kell kutatnunk a szerző(ke)t (illetve a jogtulajdonosokat), meg kell velük kötni a felhasználási szerződést, és csak ezután kezdődhet a munka.

A szerzői jogi szabályok

Minden olyan egyéni, eredeti alkotás[3] szerzői jogi védelemben részesül, amely egy vagy több szerző szellemi, művészi teljesítményének eredményeként jött létre. A szerzői jog – minden külön regisztrációs kötelezettség nélkül – a művet annak létrejöttétől védi.

A szerzőket a műveik után megillető jogok két részre: a *személyhez fűződő*, illetve a *vagyoni jogokra* oszlanak. A személyhez fűződő jogok – amelyek *nem ruházhatók át* – a következőkből tevődnek össze:

- a név feltüntetésének, a szerzői minőség elismerésének a joga,

- a nyilvánosságra hozatal és a visszavonás joga,
- a mű sérthetetlenségének a joga.

A *vagyoni jogok* – amelyek *bizonyos feltételekkel átruházhatók* – legfontosabb szabályai:

- a szerző a műve felhasználásáért anyagi ellenszolgáltatásra jogosult;
- a szerzői jogot a szerző kedvezményezettje örökölheti.

A személyhez fűződő jogok közül *a név feltüntetésének és a szerzői minőség elismerésének a joga soha nem évül el.* A személyhez fűződő további jogok, illetve a vagyoni jogok csak bizonyos ideig, az ún. *védelmi időn* belül érvényesíthetők; irodalmi művek esetében a védelmi idő a legtöbb európai országban a *szerző halálától számított 70 év.*

Nemcsak a szerzők, hanem a művek nyilvános előadásában, a nyilvánossághoz való közvetítésben szerepet játszó *közreműködők*, sőt, a jelentős ráfordítással létrejövő *adatbázisokat előállítók* érdekeit is védik az ún. *kapcsolódó jogok*. A közreműködők a mű nyilvánosságra kerülésétől számított 50 éven át élhetnek a *szomszédos jogok* által biztosított jogaikkal; az adatbázisok, adattárak létrehozóira pedig 15 évig érvényesek a *sui generis* jogok.

Aki a védelmi időn belül vagy többszöröz ni[4], vagy nyilvánossághoz közvetíteni akar egy szerzői művet, jogviszonyba kerül a jogtulajdonosokkal[5], így a mű *felhasználójává* válik. A felhasználás körülményeit és feltételeit írásbeli szerződésbe kell foglalni. A felhasználási szerződés tartalmára vonatkozóan nincsenek kötelező előírások, de az alábbiakra mindenképpen érdemes kitérni:

- a szerződő felek azonosítására szolgáló adatok;
- a szerződés tárgya (a szerzői mű megnevezése);
- a szerző nyilatkozata a mű eredeti voltáról[6];
- a felhasználás köre (internet, CD-ROM, nyomtatás, intranet stb.);
- a felhasználás időtartama (amelyen belül a felhasználó jogosult a mű felhasználására);
- a vagyoni jogokra vonatkozó megállapodás (a szerzői jogdíj mértéke, a fizetés módja stb.).

Amennyiben a digitalizálásra szánt műhöz előadóművészi és/vagy közvetítői teljesítmény is kapcsolódik, a szomszédos jogok jogosultjaival is kell felhasználási szerződést kötni.

A szerzői jogi rendszer nemcsak a jogtulajdonosok, hanem a felhasználók érdekeit is szolgálja. A tudományos és művészeti alkotások megismerését azok a szerzői jogi rendszerbe épített korlátozások is elősegítik, amelyek bizonyos értelemben határt szabnak a jogtulajdonosi monopol-jogok érvényesítésének. Az össztársadalmi érdekeknek a szerzők jogaival szembeni – bizonyos szűk határokon belüli – érvényesülését a jog *szabad felhasználásnak* nevezi. A szabad felhasználás körébe tartozó esetekben a felhasználók legálisan és jogdíjfizetés nélkül juthatnak hozzá a szellemi javakhoz.

A szabad felhasználás viszont csak akkor érvényesíthető, ha a felhasználás egyszerre tesz eleget az alábbi követelményeknek:

- nem jelenthet sérelmet a szerzőre nézve,
- meg kell felelnie a tisztesség követelményeinek és
- közvetve sem szolgálhat kereskedelmi érdekeket.

A jelen írás adta keretek nem teszik lehetővé a szerzői jogi rendszer – és ezen belül a szabad felhasználás – kimerítő ismertetését, így most csak a tárgyunk szempontjából legfontosabb eseteket emeljük ki. Archiválási célra lehet másolatot készíteni (tehát digitalizálni is), ha az

- belső intézményi célokat,

- tudományos kutatást,
- könyvtárközi kölcsönzést szolgál, vagy
- korábban megjelent mű kisebb részéről, vagy folyóirat-, illetve újságcikkről készül.

Fontos tudni azonban, hogy archiválási céllal sem szabad másolatot készíteni építészeti műről, műszaki létesítményről, szoftverről és számítástechnikai eszközzel működtetett adatbázisról.

A szabad felhasználás körébe eső, *magáncélra* történő digitalizálásra is hasonló szabályok érvényesek. Fontos megjegyeznünk azonban, hogy a magáncélra történő digitalizálásra csak akkor érvényesíthetők a szabad felhasználás kedvezményes szabályai, ha a magánszemély saját maga készíti, és nem mással készíttetni a digitális másolatot.

A forrásmű kiválasztása

Könnyű lenne azt mondani, hogy digitalizáljunk minden korábban nyomtatásban megjelent művet, tegyük közzé a hálózaton, és bízzuk a felhasználókra a válogatást. Ez az út azonban – főként anyagi okokból – még a leggazdagabb országok számára sem járható. Sokba kerül maga a digitalizálás, sok élőmunka árán lesznek a művek visszakereshetővé, és hatalmas számítógépes tárkapacitásra van szükség, ha nem csak tárolni, de szolgáltatni is akarjuk a digitalizált műveket.

A döntési folyamat egyik legnehezebb kérdése a szelekció. A „mit digitalizáljunk" kérdésre adott válasz 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. Éppen ezért az a jó megoldás, ha a szelekció kérdését szakavatott döntéshozókra, a *tudományos műhelyekre* bízzuk, mert – a köz szolgálatára és a tudomány érdekeire együttes figyelmet fordítva – ők tudnak eleget tenni a kiválasztás nehéz feladatának.

Szelektálni egyáltalán nem könnyű, ugyanis sehol a világon nincs olyan elfogadott kánon, amely meghatározná, mi készüljön el legelőször, mi kerüljön a másod- vagy harmadvonalba. A kockázatot és a döntés felelősségét azonban vállalni kell: meg kell állapítani a szerzők sorrendjét, és a különböző kiadások közül ki kell választani a legjobb minőségű szövegeket. Ha létezik olyan forrásmű, amelyet többéves kutatómunkával rendeztek sajtó alá, akkor ezt kell választani.

A digitalizálandó mű elemzése

Alapelvként leszögezhetjük, hogy a digitalizálásnak nem szabad kárt okoznia: rongálni, roncsolni csak olyan dokumentumokat szabad, amelyekből másik példány is rendelkezésünkre áll. A döntési folyamatban elsőként az állományvédelmi szempontok alapján kell kiválasztanunk a megfelelő eljárást. Ezt követően sorra kell vennünk a rendelkezésre álló analóg példány jellemzőit, illetve a tervezett szolgáltatás céljait; ezek úgyszintén nagyban befolyásolják a digitalizálási technológiát.

Az állományvédelmi kérdések tisztázása

A forrásművet először állományvédelmi szempontból kell megvizsgálni, hogy ki tudjuk választani a megfelelő digitalizálási eljárást.[7] Többféle szkennelési technológia létezik, amelyek közül azonban feltehetően csak egy lesz alkalmazható a konkrét környezetben az adott műre. Az alábbiakban számba vesszük a két szélső értéket és a leggyakrabban alkalmazott megoldást.

a) A kötetet lapokra vágják szét – a kötet roncsolásával járó megoldás akkor alkalmazható, ha a szóban forgó dokumentum könnyen hozzáférhető, és akár több példányban is rendelkezésre áll. Ezzel a technológiával a szkenneléssel kapcsolatos élőmunka költségei jóval alacsonyabbak lesznek, a szabadlapos szkenner beruházási költségei azonban magasak.

b) A forrásmű kontakt-szkennerrel digitalizálható, ha úgy ítéljük meg, hogy a dokumentum nem károsodik a laponkénti széthajtogatástól és a szkenner által kibocsátott, igen erős megvilágítástól.

c) Ha az eredeti dokumentumok fokozott védelemre szorulnak, a pótolhatatlan értéket jelentő példányok digitalizáló eszközeként szóba sem jöhet az erős megvilágítást alkalmazó kontakt-szkenner. A nagy értékű műveket kizárólag rendkívül alacsony megvilágítási értékkel dolgozó, speciálisan e célra kifejlesztett, igen drága eszközökkel szabad és lehet digitalizálni.

Végezetül azt is az állományvédelem szemszögéből kell eldönteni, a forrásműveket szállítsák-e a digitalizáló műhelybe, vagy az eszközöket vigyék a dokumentumok őrző helyére.

A forrásmű fizikai adottságainak és szövegének elemzése

Mielőtt a digitalizálás technológiáját, illetve az *output* iránt támasztott követelményeket meghatároznánk, elemeznünk kell az *input*, vagyis a digitalizálás alapjául szolgáló forrásmű tulajdonságait:

1) *a szöveg struktúrája*

- csak alfabetikus írásjelekből áll, vagy képletek, szimbólumok, az ábécétől eltérő jelölések is vannak a szövegben;
- a főszöveg nyelvében előfordulnak-e idegen nyelvű elemek – ha igen, ezek a főszöveg, vagy más írásrendszer karakterkészletéből állnak-e, és milyen arányt képviselnek a főszöveghez képest;
- szerkezetileg mennyire tagolt a szöveg, vannak-e hierarchikus szintek (főcím, fejezet-, szakasz-, egyéb részcím, vagy kötet-, ciklus-, verscím stb.)?

2) *a szemantikai sajátosságok*

- hányféle, a szöveg értelmezése szempontjából lényeges elem definiálható (utalók, indexek, mutatók, idézetek, mottók, ajánlások, nevek stb.)?

3) *tipográfiai adottságok*

- folyamatos, vagy hasábokba, szakaszokra tördelt-e a szöveg, vannak-e a szöveg formázásának sajátosságai (például versforma, táblázat, kiemelés, lábjegyzet, végjegyzet stb.)

A következő oldalon látható táblázatban összefoglaljuk azokat a legfontosabb döntési szempontokat, amelyek befolyásolják, melyik esetben milyen digitalizálási módszert válasszunk.

Tekintetbe kell még venni a dokumentum fizikai jellemzőit is: milyen a papír és a nyomdai előállítás minősége, vannak-e a lapokon esetleges elszíneződések, foltosodás stb.?

Mielőtt a szkennelés és karakterfelismertetés közös alkalmazása mellett döntenénk, ne hagyjuk figyelmen kívül, hogy az OCR technológia a lézernyomtatóval, famentes papírra, folyó szövegként kinyomtatott, mai helyesírású szövegekre van optimalizálva. Ha a forrásmű bármely vonatkozásban eltér ettől a négy kritériumtól, a karakterfelismertetés során jelentős minőségromlást fogunk tapasztalni, és számíthatunk rá, hogy az eredmény nem lesz megfelelő.

Gyakorlati tanácsként azt mondhatjuk, a döntés előtt érdemes próbát tenni: szkenneltessünk be három jellemző oldalt, majd ismertessük föl a szoftverrel, és elemezzük a kapott eredményt. Abból, hogy mennyi időt kell a hibajavításra szánni, ki tudjuk kalkulálni, egy óra alatt milyen mennyiséget lehet kijavítani. Elképzelhető, hogy a kapott eredmény alapján úgy döntünk, inkább begépeltetjük a szöveget. Ez utóbbi professzionális megoldásai között jól ismert a kétszeres bevitel: ekkor két különböző leíró gépeli le a szöveget, majd számítógéppel összehasonlítják a két változatot, és csak azt a szöveget vetik össze az eredetivel, ahol a két változat eltér egymástól. Annak a valószínűsége ugyanis,

Elemzési terület	Elemzési szempont	Igen	Digitalizálási módszer	Nem/ Nincs	Digitalizálási módszer
1. a szöveg struktúrája	1/a) csak alfabetikus írásjelekből áll?	X ⇒	OCR	X ⇒	képfájl
	1/b) képletek, szimbólumok, az ábécétől eltérő jelölések vannak a szövegben?	X ⇒	képfájl	X ⇒	OCR
	1/c) vannak idegen nyelvű elemek?	X ⇒	leírás/ képfájl	X ⇒	OCR
	1/d) az idegen nyelvű szövegek a főszöveggel egyező karakterkészletből állnak?	X ⇒	leírás	X ⇒	képfájl
	1/e) az idegen nyelvű szövegek a főszövegtől eltérő írásrendszer karakterkészletéből állnak? (például latin betűk között görög karakterek)	X ⇒	képfájl	X ⇒	leírás
	1/f) vannak hierarchikus szintek?	X ⇒	képfájl	X ⇒	OCR
2. szemantikai sajátosságok	vannak a szöveg értelmezése szempontjából lényeges elemek? utalók, indexek, idézetek, mottók, ajánlások, nevek stb.	jelölő rendszer (SGML, XML) alkalmazása ajánlott		nem érdemes jelölő rendszert alkalmazni	

3. tipográfiai adottságok	3/a) folyamatos a szöveg?	X ⇒ OCR	X ⇒ leírás/ képfájl
	3/b) hasábokba, szakaszokra tördelt a szöveg?	X ⇒ leírás/ képfájl	X ⇒ OCR
	3/c) vannak a szöveg formázásának sajátosságai?	X ⇒ leírás/ képfájl	X ⇒ OCR
	3/d) versforma	X ⇒ leírás/ képfájl	X ⇒ OCR
	3/e) táblázat	X ⇒ leírás/ képfájl	X ⇒ OCR
	3/f) kiemelés	X ⇒ leírás/ képfájl	X ⇒ OCR
	3/g) lábjegyzet	X ⇒ leírás/ képfájl	X ⇒ OCR
	3/h) végjegyzet	X ⇒ leírás/ képfájl	X ⇒ OCR

hogy két leíró ugyanazon a helyen ugyanazt a hibát „állítja elő", rendkívül csekély.

Ha túl bonyolult szöveggel állunk szemben, és/vagy a rendelkezésre álló anyagi eszközeink szűkösek, illetve az outputot illetően nem elsődleges fontosságú a teljes szöveg visszakereshetővé tétele, az eredeti mű fakszimile formátumának megjelentetése kínálkozik megoldásként. Ebben az esetben elmarad az összes, egyenként is jelentős élőmunka-ráfordítást igénylő művelet: a karakterfelismertetés utáni többszöri korrektúra, a strukturális, szemantikai és tipográfiai sajátosságok kódolása – viszont elveszítjük a számítógépes feldolgozás legnagyobb előnyét: a sok szempontú visszakereshetőséget.

A döntési folyamat során további megfontolás tárgya a költségek elemzése: ki kell kalkulálni, mekkora beruházás, mennyi élőmunka-ráfordítás szükséges a kívánt eredmény eléréséhez? Erről bővebben „A digitalizálás költségigénye" c. fejezetben olvashatunk.

A tartalomszolgáltatás szempontjai

Akár online, akár offline módon elérhető szolgáltatást tervezünk, az eddig felsoroltakon túlmenően további számos kérdésre kell előre megadnunk a választ annak érdekében, hogy helyesen tudjuk kialakítani a digitalizálás célrendszerét, majd kiválasztani a megfelelő eljárást. A tartalomszolgáltatás olyan sokrétű tevékenység, amelynek alapos ismertetése szétfeszítené e cikk kereteit, ezért kénytelenek vagyunk a legjellemzőbb kérdések föltevésére és a rájuk adott legtipikusabb válaszokra szorítkozni.

- Milyen feldolgozási minőséget akarunk garantálni? – Hasonmás műveknél eldöntendő kérdés, elegendő-e fekete-fehérben reprodukálni az oldalakat, vagy inkább színhelyesen akarjuk még az elszíneződéseket, foltokat stb. is tükröztetni. A karakterenként kódolt szövegnél 100%-os pontosságra kell törekedni, egyúttal meg kell határozni a megengedhető maximális hibaarányt[8]. Vannak olyan szövegek, amelyekben – legalábbis elvileg – egyáltalán nem fordulhat elő hiba.
- Hogyan akarjuk megtalálhatóvá tenni a digitalizált művet? – A dokumentumot el kell látni a keresőprogramok számára „érthető", kezelhető metaadatokkal (a dokumentum azonosító adataival, illetve kulcsszavakkal).[9]
- Milyen adathordozón, milyen formátumban[10], milyen feltételekkel bocsátjuk közre a szöveget? – Sem szerzői jogi, sem menedzselési szempontból nem mindegy, hogy offline hordozón (pl. CD-ROM-on), vagy online módon (az interneten vagy intraneten), továbbá ingyenesen, vagy térítés ellenében tesszük hozzáférhetővé a dokumentumot. Ha térítéses szolgáltatásban gondolkodunk, akkor rendelkeznünk kell az elektronikus kereskedelem folytatására szóló engedéllyel. Ha kötelező regisztrációhoz kívánjuk kötni a szolgáltatás igénybe vételét, akkor eleget kell tennünk az adatvédelmi törvény szigorú adatkezelési előírásainak stb.
- Kereshetővé akarjuk-e tenni a szöveg egyes elemeit? – Itt nemcsak a szöveg karakterekké alakítása a kérdés, hanem az is, vannak-e olyan szövegelemek, amelyek minősített keresését[11] fontosnak tartjuk. További kérdések is adódnak:
  - A szövegben eltérő háttérszínnel ki akarjuk-e emelni a keresett elemet?
  - Több találat esetén hogyan oldjuk meg a találatok közötti léptetést?
- A megjelenítés milyen funkciókat támogasson? E kérdéskörben elsősorban a következőkre kell válaszolnunk:
  - Engedélyezzük-e a letöltést, a másolást, a nyomtatást? – Ha szerzői jogi, üzleti, vagy más megfontolásból nem kívánjuk megengedni a digitalizált mű többszörözését, akkor PDF formátumban tegyük közzé a szöveget, és tiltsuk le a felsorolt funkciókat. A legújabb PDF verziók az előbb felsoroltakon túl lehetőséget adnak a dokumentumok hitelesítésére is[12].
  - A nyomtatott művekhez hasonlóan akarjuk-e lapozni a szöveget? – Ezt a funkciót kétféleképpen teljesíthetjük: vagy megtartjuk az eredeti dokumentum oldalankénti elrendezését, vagy egy program segítségével a felhasználó gépének adottságaihoz igazítjuk a virtuális oldalakat, így könnyítve meg a mű olvasását.

A szövegek reprezentációjára szolgáló formátumok, jelölőrendszerek

A szöveg értelmezésének három szintje ismeretes: a *formai* (layout), a *logikai* (szintaktikai) és a *tartalmi* (szemantikai). Vannak olyan szövegformátumok, amelyek csak a formai adottságokat, mások pedig a szintaktikai és szemantikai elemeket is tudják kezelni.

Ha a hosszú idejű megőrzés mellett a szöveg bizonyos elemeinek minősített keresésére[13] és a számítógépes hardver- és szoftvereszközök adottságaitól független, széles körű használhatóságra egyidejűleg törekszünk, akkor nem elégedhetünk meg a ma elterjedt HTML formátummal[14]. Bár sokkal több előkészületet igényel, és nagy az élőmunka-ráfordítás igénye, hosszabb távon megéri az SGML szabványt, vagy annak legújabb „leszármazottját", az XML-t alkalmazni.

Az SGML szabványt[15] 1986-ban fogadták el. Az

elmúlt két évtized során számos tudományterületre és annak jellemző dokumentumtípusaira kidolgozták a speciális SGML alkalmazásokat, a világot az 1998-ban napvilágot látott XML[16] változat hódította meg, amely érvényesíti az SGML előnyeit, de igyekszik kiküszöbölni annak hátrányait.

Az SGML szabvány alkalmazása jelentős előkészítő munkát igényel, melynek során ki kell dolgozni a tartalmi elemek jelölését, rögzíteni kell a különböző információtípusok közötti kapcsolatokat, a dokumentum struktúrájára vonatkozó szabályszerűségeket. Azt is 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 alkalmazások „lelke" a DTD, amely nem más, mint az egyes szövegtípusok (ez lehet például vers, dráma, szabadalmi leírás stb.) szövegmodellje.

Az SGML szabvány szerint feldolgozott dokumentumban a szövegbe ágyazva, de attól speciális határoló jelekkel elválasztva jelöljük a metaadatokat. A metaadatok három típusát különböztetjük meg:

- *elemek* (a dokumentum alkotóelemei, amelyek lehetnek kötelezőek vagy opcionálisak – például: a ‚Cím' lehet kötelező, az ‚Alcím' pedig opcionális elem);
- *attribútumok* (hivatkozások a szövegen belüli vagy kívüli, az egyes elemek tulajdonságait jelölő objektumokra – attribútum lehet például a szöveg nyelve, a nyelvkódok szabványos jelölésével megadva);
- *entitások* (a dokumentumban használható speciális karakterek, beágyazott képek, táblázatok stb.) Olyan entitásokat, amelyeket nem határozunk meg előre, nem lehet a dokumentumban alkalmazni.

Másként fogalmazva az *elemek* a dokumentum logikai szerkezetét határozzák meg, az *entitások* pedig a logikai szerkezet mögött lévő *fizikai szerkezetet* írják le.

A felsorolásból látható, hogy a dokumentumok formai jegyeit az SGML állományok nem tartalmazzák. Mindazt az információt, amelyet az egyes dokumentumtípusok megjelenítésével kapcsolatban fontosnak tartunk, részben a DTD fájlokban, részben a külön definiálandó stíluslapokban határozhatjuk meg.[17] A dokumentumok megjelenítésére külön szabvány, a DSSSL[18] szolgál.

Az SGML egyik legnagyobb előnye, hogy független a hardver- és szoftvereszközök fajtájától és típusától, illetve a számítógép operációs rendszerétől[19]. A szabvány további kedvező adottsága, hogy különválasztja a tartalmat a formától; hátránya viszont, hogy bonyolult és drága, mert alkalmazása speciális tudást igényel. Ha valóban hosszú távra akarunk digitalizálni, akkor viszont megéri az SGML technológiát alkalmazni, mert a szabványos eszközökkel kódolt szöveghez mindig lehet olyan konverziós programot írni, amelyik az output oldal mindenkori kívánalmainak megfelel.

A digitalizált állomány megőrzésének kérdései

Az informatikai hardver- és szoftvereszközök rendkívül gyorsan elavulnak, ezért a ma rendelkezésünkre álló digitalizálási eljárások eredményeként létrejövő számítógépes állományok várható élettartama igen rövid. A gyors technológiai avulás következtében – nem véletlenül – a jelentős ráfordítást igénylő digitalizálás egyik kulcskérdése a megőrzés, illetve a tartalomszolgáltatás tervezett időtartama. Egy adott számítógépes környezetben alig néhány évig tudunk úgy dolgozni, hogy a gépünkből kikerülő szövegek és egyéb állományok a mindenko-

ri átlagos színvonalon lévő számítógépekkel értelmezhetőek legyenek. Sajnos, ez igaz a merevlemezen tárolt fájlokra, de még inkább az offline hordozókra (például CD-ROM-okra) kiírt állományokra. Ahogy ma már – lejátszó egységek híján – nem tudjuk a 80-as évekbeli mágnesszalagokat, vagy a 90-es évek elején használt 5,25 inches hajlékony lemezeket leolvasni, tíz év múlva ugyanígy nem lesznek eszközeink a ma általánosságban elterjedt háttértárolókon lévő állományok olvasására.

A digitalizált állomány megőrzése részben a fizikai, részben a technikai környezettel szemben támaszt követelményeket. Fizikailag biztosítanunk 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 szövegdigitalizálás módszerei

Szkennelés, fakszimile kép előállítása

A szkennelés eredményeként a digitalizált oldal *képe* jön létre, amely az eredeti oldal hű leképezése. A szkenner által létrehozott képfájl a hagyományos nyomdatechnikában ismert „fakszimile", „hasonmás" oldalra hasonlít. A digitalizálás célkitűzései között meg kell határozni, kielégíti-e céljainkat a képfájl, vagy kereshetővé akarjuk tenni a szöveget – ez utóbbi esetben a képen látható szöveget át kell kódoltatni számítógéppel olvasható formátumra.

A szkennerekről elég annyit tudni, hogy – a digitalizálandó mű hordozójától, fizikai paramétereitől és az állományvédelmi szempontoktól függően – különböző típusok közül választhatunk. A legismertebbek az ún. síkágyas szkennerek, amelyek leginkább a jól ismert másológépekre hasonlítanak, de vannak olyanok is, amelyek szabadlapok, vagy mikrofilmek, vagy diaképek digitalizálására alkalmasak. Bár nem a szkennerek között tartjuk számon őket, a digitális fényképezőgépek is használhatók dokumentumok digitalizálására.

Szövegfelismerés (OCR)

Amennyiben a digitalizálás célja *számítógéppel olvasható szöveg* előállítása, akkor szükség van a szkennelt képek konvertálására, vagyis a képi elemekként tárolt információk szöveggé való visszafejtésére. E célra speciális karakterfelismerő szoftvereket[20] fejlesztettek ki, amelyek működési elve a következő. A képfájl egészen apró elemekből, pontokból (ún. pixelekből) épül föl. A digitalizált kép adottságaitól függően minden egyes képpont hordoz valamilyen információt: a legegyszerűbb esetben, amikor a fehér lapon csak fekete betűk szerepelnek, ez az információ az **igen/nem** (vagyis az **1** és a **0**) váltakozására szorítkozik. Esetünkben az „igen" a feketével fedett, a „nem" pedig a nem fedett képpontot jelenti.[21]

A szkenner a lap minden egyes apró pontjáról tárolja azt az információt, van-e ott festék („igen"), vagy nincs („nem"). A szövegfelismerő szoftver ezen a képfájlon halad végig, és a képpontok eloszlását hasonlítja 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 szövegfelismerő szoftverek nemcsak az egyes írásjelek képét tárolják, de terjedelmes szótárakat is tartalmaznak. Még a szövegfelismertetés ele-

jén ki kell választani a munkanyelvet, amelynek szókészletével a szoftver a feldolgozás utolsó fázisában összeveti az általa felismerni vélt szavakat, és a képernyőn (általában valamilyen színes kiemeléssel) jelzi, ha két szóköz-jel között olyan karaktersort állapított meg, amelynek megfelelője nem található meg a szókészletben. A számítógép által hibásnak jelzett szavakat mindenképpen ellenőrizni – és szükség esetén javítani – kell. Gyakran előfordulnak azonban alakilag hibátlan, az adott szövegkörnyezetben mégis hibásnak számító szavak, amelyeket szintén javítani kell. A korrektúrázást csak intellektuális munkával lehet elvégezni – éppen emiatt igényel a szövegfelismerés jelentős élőmunkaráfordítást.

A kijavított szöveget – a digitalizálás célfüggvényében meghatározott elvek alapján – vagy újból ellenőrizzük és javítjuk, vagy nem. Ha jó minőségű szöveget kívánunk reprodukálni, minimum kettő, de sok esetben három korrektúra-fázisra is szükség van. Minél bonyolultabb, összetettebb szöveget digitalizálunk, annál több emberi beavatkozásra van szükség a karakterfelismerés során előállt hibák kijavítása érdekében.

Begépelés

Nehéz egzakt módon meghatározni, hol van az a határ, amelynél a begépelést érdemesebb választani. Általánosságban azt mondhatjuk, ha régies helyesírású, vagy sok idegen szót tartalmazó, vagy különleges tipográfiai elemeket (például sok dőlt betűt, vagy hasábokra tördelést) tartalmaz a szöveg, akkor kifizetődőbb leíratni a szöveget, mint a karakterfelismerés után korrektúráztatni. Korábban már említettük: ha a cél igazán jó minőségű szöveg előállítása, akkor érdemes két leíróval legépeltetni ugyanazt a szöveget, amelyet egy számítógépes programmal összehasonlítva csak ott korrektúrázzák, ahol eltérés mutatkozik a két változat között.[22]

A különböző szövegdigitalizálási eljárások előnyeit és hátrányait az alábbi táblázatban foglaltuk össze:

eljárás	eredmény	előny	hátrány
begépelés	szövegfájl	pontos	költséges
szkennelés	képfájl (átalakítható szöveggé)	olcsó	tárolása nagy tárkapacitást igényel
szkennelés + szövegfelismerés	képfájl + számítógéppel olvasható szöveg	rongálja az eredeti példányt	csak a mai helyesírású szövegeknél ad jó eredményt
digitális fényképezés	képfájl + számítógéppel olvasható szöveg	olcsó, az eredeti példányt védi	tárolása nagy tárkapacitást igényel

Mindig a cél határozza meg, melyiket választjuk!

A digitalizálás sikere számos, egymással ellentétes hatást gyakorló tényező helyes megválasztásán múlik, így a munka megkezdése előtt még az alábbiakat is mérlegelnünk kell:

igény	következmény
nagy mennyiségű szöveg részletekbe menő feldolgozása	hosszú időtartam, jelentős költségráfordítás
bonyolult felépítésű digitalizált állomány	nehezebb használhatóság és érthetőség
a pillanatnyilag hozzáférhető, olcsóbb technológia alkalmazása	a hosszú távú megőrzés ellen hat
sok szempontú felhasználhatóság	nem felel meg egyes speciális igényeknek

A digitalizálás költségigénye

A megfelelő kapacitású digitalizáló rendszer kialakításának költségei több összetevőből állnak, melyek közül vannak kötelezően, illetve opcionálisan előfordulók. Az alábbi táblázatban azokat a költségtípusokat foglaljuk össze, amelyek egy tipikus digitalizálási feladat végrehajtása során jelentkeznek.

megnevezés	feladat	költségtípus	eseti / folyamatos	megjegyzés
szakértői díjak	a rendszerterv elkészítése, a forrásművek kiválasztása, előkészítése stb.	személyi / dologi kiadás	eseti folyamatos	
szerzői jogdíj	felhasználási szerződés megkötése	jogdíj	eseti	a védelmi időn belül lévő művek jogtulajdonosai számára
hardver és szoftver beruházás	a digitalizáló eszközpark kialakítása	felhalmozás	eseti	ha saját digitalizáló műhelyt akarunk berendezni
feldolgozás	digitalizálás, korrektúrázás, metaadatok készítése és karbantartása	személyi / dologi kiadás	folyamatos	
állományvédelem	restaurálás	dologi kiadás	eseti	
archiválás	tárolás	dologi kiadás	folyamatos	
szolgáltatás	a publikus hozzáférés hardver-, szoftver- és telekommunikációs költségei	dologi kiadás	folyamatos	

Az előzetes költségszámítások alapján lehet eldönteni, mely fázisokat tudjuk „házon belül" elvégezni, és melyeket kell kiadni külső vállalkozónak. Saját digitalizáló műhelyt akkor érdemes kialakítani, ha annak kapacitását folyamatosan ki tudjuk használni. Azt is érdemes tekintetbe vennünk, hogy az informatikai rendszerek és eszközök amortizációja 2–3 év, ezért már a tervezés során kell az infrastruktúra megújítására gondolnunk.

A szerzői jogok védelme digitalizált művek esetében

A digitalizálás során a szerzői jog védelmével nemcsak a korábban nyomtatásban megjelent művek felhasználójaként, hanem tartalomszolgáltatóként is tisztában kell lennünk. A jelentős ráfordítással digitalizált és ugyancsak komoly költséggel működtetett rendszerben szolgáltatott művek illegális felhasználása ellen nemcsak saját érdekünkben, de azoknak a jogtulajdonosoknak az érdekében is fel kell lépnünk, akiknek mi annak rendje-módja szerint kifizettük a jogdíjat. A jogtulajdonosok elvárják a tartalomszolgáltatóktól, hogy akadályozzák meg műveik engedély nélküli letöltését és esetleges illegális forgalmát.

A digitálisan hozzáférhető állományok szerzői jogvédelmére a hagyományos eszközök nem alkalmasak, ezért informatikai megoldásokat fejlesztettek ki erre a célra. A digitális tartalmakhoz való hozzáférést lehetővé tevő, illetve 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 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 tartalomátvitelt a jogosulttól a felhasználóig, és
- a felhasználási díj elszámolását.

A DRM rendszer két alapvető funkciója:

- egyrészt a szellemi alkotásokat, az alkotásokon fennálló jogokat, illetve az alkotókat és a szerzői jogi jogosultakat *azonosítja,*
- másrészt a dinamikusan változó felhasználási környezetben a *joggyakorlás érvényesítését szolgálja.*

A dokumentumok azonosítására szolgál a *Digital Object Identifier* (DOI) és a *digitális vízjel.* A szerzők, illetve a jogosult felhasználók a *digitális aláírás* segítségével igazolhatják személyazonosságukat. A digitális tartalomátvitel során a jogosulatlan hozzáférést megakadályozó jogkezelő eljárás a hitelesítés és a titkosítás.

Összegzés

A digitalizálási feladatoknak akkor tudunk a legjobb színvonalon eleget tenni, ha mindenkor a nemzetközi szinten elterjedt szabványos megoldásokat alkalmazzuk.[23] Ebben az esetben még abban is bízhatunk, hogy időről időre elkészül majd az a konverziós és/vagy migrációs eljárás, amely úgy váltja föl a jelenleg ismert szabványos eljárásokat, hogy adatainkat és a digitalizált állományokat veszteség nélkül tölthetjük át az új rendszerbe.

Jegyzetek

1. 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.

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

3. Fontos megemlíteni, hogy az ötlet, az elgondolás, az eljárás, az elv, a működési módszer, a matematikai művelet stb. nem részesül szerzői jogi védelemben.

4. A szerzői jog szempontjából többszörözésnek minősül a mű másolása bármilyen hordozóra (papírra, filmre, elektronikus adatrögzítő eszközre stb.).

5. Jogtulajdonos: a szerző vagy jogutódja.

6. A mű eredetiségére vonatkozó nyilatkozatban a jogtulajdonos kijelenti: a mű eredeti szellemi termék, amelynek ő rendelkezik a felhasználási jogaival.

7. Annak eldöntése, hogy a szkennelést vagy a begépelést választjuk, elsősorban nem az állományvédelmi szempontok, hanem a szöveg adottságainak, illetve a megcélzott szolgáltatás minőségének a függvénye. Erről bővebben „A forrásmű fizikai adottságainak és szövegének elemzése" c. alfejezetben olvashatunk.

8. A Digitális Irodalmi Akadémia (www.irodalmiakademia.hu) gyűjteménye esetén a maximálisan megengedhető hibaszám 3 ezrelék.

9. A leíró adatokhoz a Dublin Core metaadatszabványt, a dokumentumot kísérő azonosító jelzetre a DOI-t (Digital Object Identifier) ajánljuk. A Dublin Core, illetve a DOI honlapja: dublincore.org és www.doi.org

10. A formátumok közül a leggyakoribb a HTML, a PDF, az XML, a képfájlok közül a TIFF, a JPG és a BMP.

11. A minősített keresésről a következő fejezetben esik szó.

12. Egy változat hiteles volta különleges jelentőséggel bír például a hivatalos közlönyöknek az internetes közzétételénél.

13. A számítógépes szövegszerkesztők és a HTML-t értelmező Web-böngésző programok alapértelmezésben a szöveges keresést „tudják"; ekkor a számítógép karakterről karakterre hasonlítja össze a keresőkérdést a keresett szöveggel, és csak az egyező karakterláncot értelmezi találatként. Ezzel szemben a „minősített keresés" alatt azt értjük, hogy egy előre kidolgozott séma alapján megjelöljük azokat a szövegelemeket, amelyeket kereshetővé akarunk tenni. A számítógép a jelölők alapján találja meg a meghatározott elemeket. Ha például egy szövegben az összes név kereshetősége fontos, minden név elé beillesztjük a <name> jelölőt, így megtaláljuk az összes Kiss, Nagy, István, Pista, Júlia, Julcsi stb. nevet. Ha azonban külön-külön akarjuk látni a vezeték-, a kereszt- és a beceneveket, három jelölőt alkalmazunk: <familyname> <forename> <nickname>.

14. A HTML is az SGML szabványból származik, de gyakorlatilag a dokumentumoknak csak a formai jegyeit tudja kezelni.

15. Standard Generalized Markup Language, ISO 8879:1986

16. Extensible Markup Language

17. Egy példa: 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.

18. Document Style and Semantics Specification Language, ISO/IEC 10179:1996

19. A számítástechnikában ezt „platformfüggetlenségnek" nevezik.

20. A karakter-, vagy szövegfelismerő szoftvereket az angol nevük – Optical Caracter Recognition – rövidítése alapján gyakran OCR programoknak nevezik.

21. A színes képek esetében is hasonló a folyamat, azzal a különbséggel, hogy a „nemfehér" képpontok esetében a szkenner érzékelősora felváltva tapogatja le a három különböző színkép-tartományban érzékeny pixeleket. (A legismertebb az RGB [vörös-zöld-kék] színrendszer.)

22. Annak a valószínűsége, hogy két ember ugyanazon a helyen ugyanazt a hibát véti, gyakorlatilag nulla.

23. A cikkben említett néhány, a digitalizálási tevékenységhez kapcsolódó nemzetközi szabvány ebbe a kategóriába tartozik, így ezeket bízvást alkalmazhatjuk.