Takáts Béla

A digitalizált mikrofilmek szolgáltatásának egyik alternatívája: a kétrétegű PDF

2010. február 4-én az Országos Széchényi Könyvtárban előadást tarthattam a helyi lapok digitalizálásának lehetőségeiről, a kétrétegű PDF-ek készítésének mikéntjéről, illetve ezek könyvjelzőzésének lehetőségeiről. Bár az eljárásban nincs semmi új és meglepő, talán érdemes írásban is összefoglalni az ott elhangzottakat, hogy 1. az itt leírtak kellően felboszszantsák az e kérdésekhez jobban értőket és hozzászólásaik nyomán kollégáim újabb ismereteket szerezhessenek, és/vagy 2. választ adjanak az előadáson el nem hangzott, vagy az azóta hozzám érkezett kérdésekre.

Miről is van szó?

Néhány városi könyvtárnak – egy állományvédelmi digitalizálási program keretében – módja nyílt az *Országos Széchényi Könyvtár*tól (*OSZK*) megrendelni a helyi lapjaikról készített mikrofilmek digitális másolatait, illetve megvásárolhatta az ABBYY FineReader (Professional Edition) optikai szövegfelismerő rendszer 9.0-ás, valamint az Adobe Acrobat Pro szintén 9.0-ás verzióját. A támogatás elnyerésének feltétele volt, hogy a számítógépes programok segítségével a hírlapokról, folyóiratokról készült képeket használatra, kutatásra tegyék alkalmassá, tegyék lehetővé munkájuk eredményéhez a minél szélesebb körű hozzáférést.

Mi tehát a teendő a programok megvásárlása és a mikrofilmek digitális másolatainak megrendelése, megérkezése után?

Nézzük meg, mink van (és hogy veheti hasznát szinte azonnal az olvasó)

Ahhoz, hogy lássuk, mit kaptunk az OSZK-tól, először is szükségünk lesz egy képnézegető programra. Erre a célra az ingyenesen, magyar nyelven is használható IrfanView programot ajánlom[1].

Ha e programmal megnyitjuk és megnézzük a képeket[2], először is azt tapasztaljuk, hogy egy kép *két* újságoldalt tartalmaz és ennek következtében szükségszerűen találunk olyan képeket, amelyek két lapszám egy-egy oldalát tartalmazzák. A képek tulajdonságairól a képnézegető program az *i* betű lenyomása után tájékoztat (*1. ábra*).

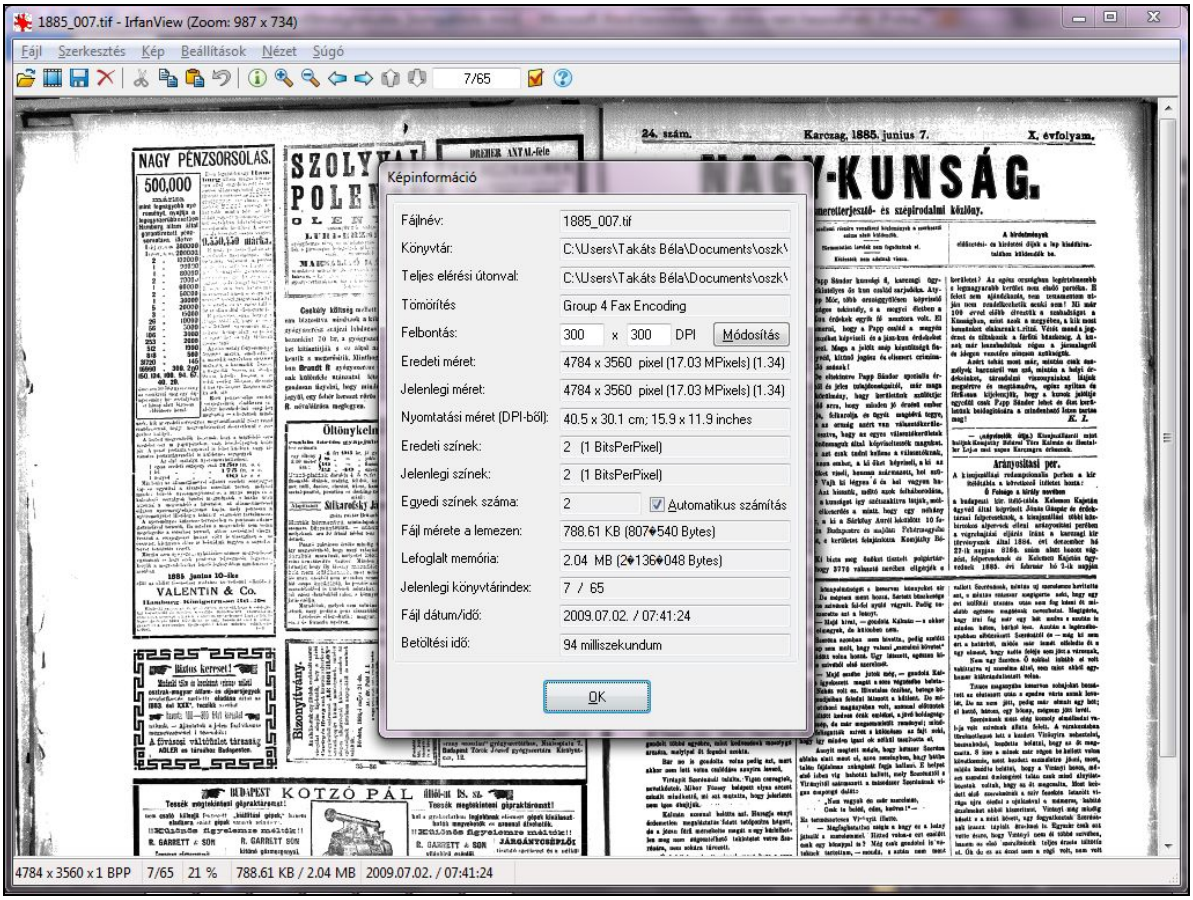
Az így nyert információkból megállapíthatjuk, hogy 300 DPI felbontású, 4784x3560 pixel méretű, kétszínű, TIF-képek állnak rendelkezésünkre.

Mit tegyünk, hogy *azonnal* szolgáltatni tudjuk szerzeményünket? Először is a megvásárolt képekről készítsünk *mentést* és gondoskodjunk az elmentett állományok biztonságos tárolásáról. Ezután rendezzük el a fájlokat egy, az olvasók számára is áttekinthető struktúrába, például így: Létrehozunk egy alkönyvtárat a lap nevével, és ebbe évenként, folyamatosan számozva helyezzük el a képeket:

```
nagy_kunsag[3]
        1885
                1885_001.tif
                1885_002.tif
                1885_003.tif
                ...
        1886
                1886_001.tif
                1886_002.tif
                1886_003.tif
                ...
        ...
```

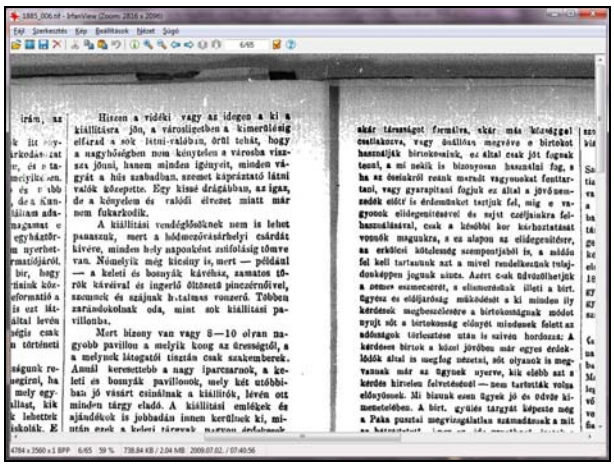
(Ehhez célszerű egy „commander” programot telepíteni és megtanulni a fájlok csoportos átnevezésének módját.) A sorszámozásnál mindig figyeljünk a kellő számú „vezető nulla” használatára (001, 002, 099, 999), hogy fájljaink sorba rendezése megfelelő legyen!

Olvasóink így – a mikrofilmmel szemben – egy kényelmesebben használható, nyomtatott és elektronikus másolatot *is* nyújtani tudó szolgáltatáshoz jutnak, ami – lássuk be – nagy előrelépés!



1. ábra **Képinformációk megtekintése az IrfanView programban**

E szolgáltatásunk persze még sok kényelmetlenséget okoz. Az olvashatóság érdelében a képernyőn jelentősen fel kell nagyítani az oldalak képét, ez zavarja az áttekinthetőséget és ebből a szempontból (sem) szerencsés, hogy egy kép két újságoldalt is tartalmaz (*2. ábra*).



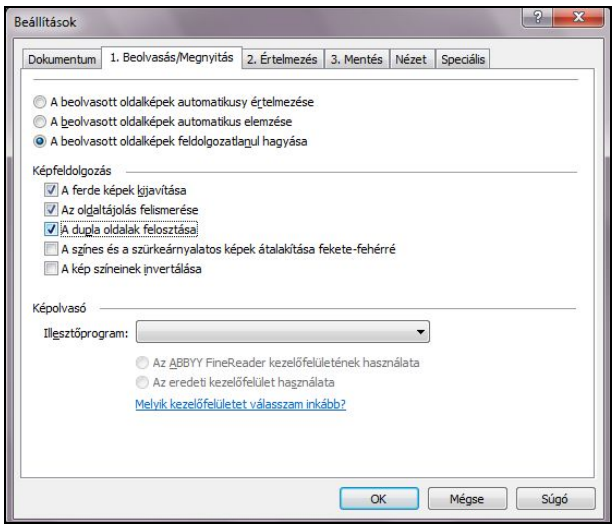
2. ábra **Az eredeti képek áttekintésének nehézségei**

Az információkhoz való hozzáférést az segíti igazán, ha a képeinken látható szöveget kereshetővé tesszük. Az első problémát (oldalak szétvágása) ugyan megoldhatnánk képnézegető programunkkal, egyéb célszoftverrel is, de ne tegyük. A megvásárolt karakterfelismerő program 9.0-ás verziója ezt a problémát is meg fogja oldani.[4]

Indítsuk el és győződjünk meg néhány beállításáról!

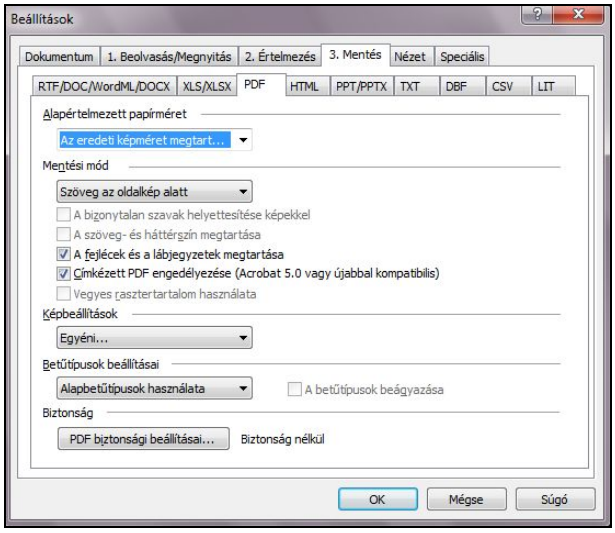
Kétrétegű PDF készítés

Az ABBYY FineReader 9.0 használata során rengeteg beállítási lehetőségünk van. Ezek közül számunkra most kettő kiemelten fontos. Ahhoz, hogy a két oldalt tartalmazó képek szétvágásával ne kelljen foglalkoznunk, az „Eszközök” legördülő menü „Beállítások” pontjára kattintás után megnyíló ablak „1. Beolvasás/Megnyitás” fülén (az egérrel) tegyünk egy pipát „A dupla oldalak felosztása” felirat előtti négyzetbe (*3. ábra*).



3. ábra **Az ABBYY FineReader 9.0 beállítása (1)**

Ahhoz, hogy kétrétegű PDF-ünk (a két réteg: a kép és a szöveg) első, látható rétege a dokumentum *képe* legyen, a felismert szöveg pedig a háttérben maradjon, a „3. Mentés” fül „PDF” alfülén szintén győződjünk meg néhány beállításról (*4. ábra*).

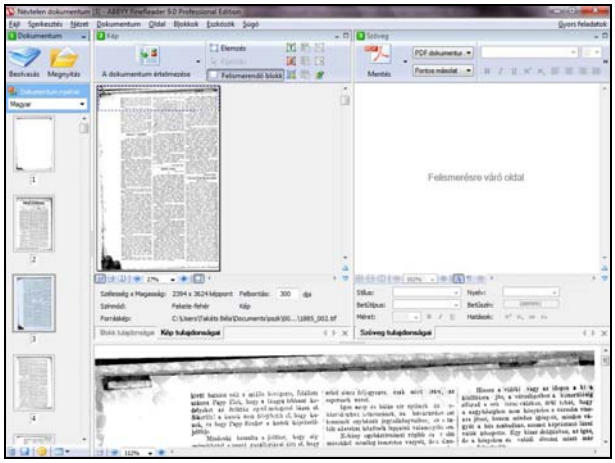


4. ábra **Az ABBYY FineReader 9.0 beállítása (2)**

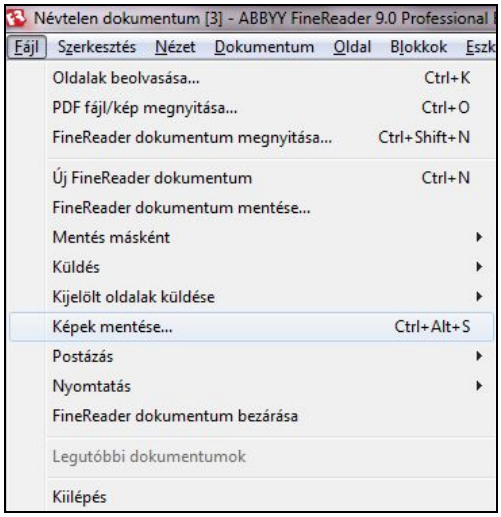
Itt figyeljünk arra, hogy az „Alapértelmezett papírméret” pontnál „Az eredeti képméret megtartása”, a „Mentési mód”-nál pedig a „Szöveg az oldalkép alatt” opció legyen kiválasztva.

Ha e két beállításról gondoskodunk, akkor képeink betöltése után[5] a következőket látjuk (*5. ábra*).

A program tehát – beállításunk nyomán – automatikusan elvégzi a képek szétvágását. Ezeket a képeket a „Fájl” – „Képek mentése” menüpont segítségével feltétlenül mentsük el és nevezzük át őket, hogy olvasóink – addig is, míg munkánk végére nem érünk – a már említettnél kényelmesebben használható szolgáltatáshoz jussanak. (*6. ábra*).



5. ábra **Az ABBYY FineReader használata**



6. ábra **Képek mentése az ABBYY FineReader 9.0 program segítségével**

A képek jobb használhatósága érdekében a következő fájlszerkezet kialakítását javaslom:

nagy_kunsag
1885
1885-05-31_01.tif
1885-05-31_02.tif
1885-05-31_03.tif
1885-05-31_04.tif
1885-06-07_01.tif
...

Azaz: megjelenés éve-hónapja-napja_sorszám. kiterjesztés. Itt se feledkezzünk meg a „vezető nullákról”. Fontos: a képeket úgy nevezzük el, hogy a majdan használandó, keresésre alkalmas PDF állományokban beazonosíthatók és igény esetén visszakereshetők legyenek, mert e képekből jobb minőségű nyomtatásokat lehet készíteni, mint munkánk leendő végeredményéből.

A dokumentumok szövegének felismertetéséhez kattintsunk a „Dokumentum értelmezése” gombra és várjunk (ill. foglalkozzunk egyéb feladatainkkal), míg a program végez a betöltött képek felismerésével...

Ez messze nem lesz hibamentes. A képernyő bal oldalán található kis képek bal alsó sarkában megjelenő, a felismerés befejezését jelző ikon mellett, a jobb oldalon feltehetően sűrűn látni fogjuk a felismerés hibáira utaló felkiáltójelet, a képernyő közepén pedig feltűnik a hibalista is (*7. ábra*).

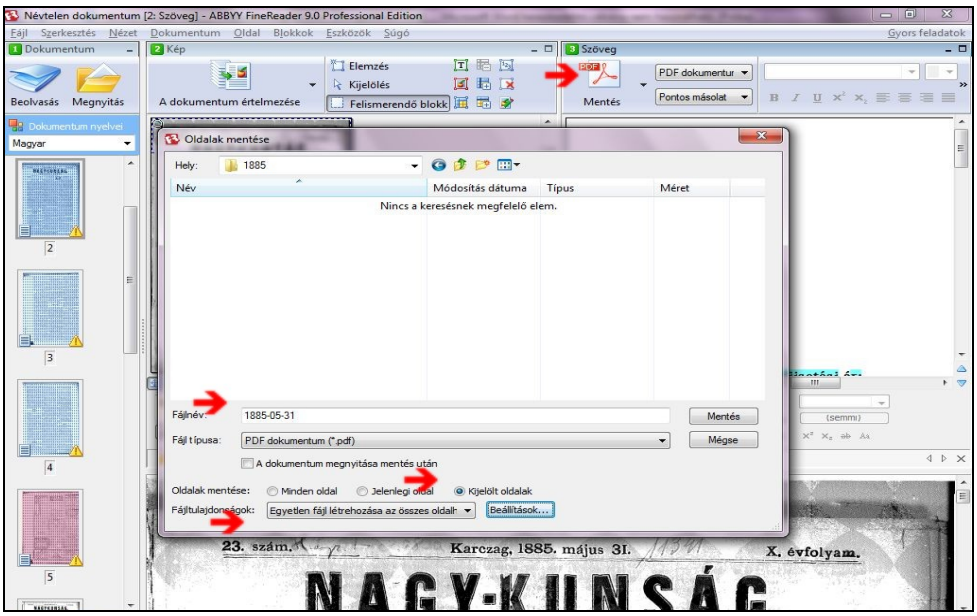
Mivel a képek minőségének javítására nincs módunk, ebbe törődjünk bele.[6]

7. ábra **Hibaüzenetek a karakterfelismerő programban**

A felismertetés után a FineReader programban lehetőségünk nyílik a szöveg korrektúrázására. Ennek felvállalását azonban alaposan meg kell fontolni, mert irdatlanul nagy munkát vállalnánk vele. Végiggondolhatjuk, kipróbálhatjuk a szöveg meghatározott részeinek (címek, alcímek, szerzők neve, szövegben szereplő nevek stb.) korrektúrázásának lehetőségeit is, mert ez később megkönnyíti a könyvjelzőzés folyamatát.[7] A szövegfelismerő programok feltételezhetően még sokat fognak fejlődni, remélhetőleg előbb-utóbb alkalmasak lesznek 19. századi szövegek helyesebb felismerésére is, illetve talán előbb-utóbb lehetőség nyílik arra is, hogy hírlapjaink mikrofilm-másolatai helyett az *eredeti* kiadványokról szerezzünk be – a másolat másolatáról készítettnél – sokkal jobb minőségű képeket.

Abban az esetben, ha egy lapnak csak egy-két évfolyama jelent meg egy településen, fel lehet vállalni ezt a sziszifuszi, de kétségtelenül nagyon érdekes feladatot is.

A felismert oldalak kétrétegű PDF fájlokba mentése nagy odafigyelést igényel. Érdemes lapszámonként készíteni a mentést oly módon, hogy a Ctrl gomb lenyomása mellett egérkattintással kijelöljük az egy fájlba menteni kívánt oldalakat. (Ha helyesen végeztük el a fájlok elnevezését, ezek egymás alatt lesznek!) A „Mentés” gomb megfelelő beállításával a következő ábra szerint lehet elvégezni ezt a feladatsort (*8. ábra*).



8. ábra **Az ABBYY FineReader 9.0 beállításai a PDF dokumentum mentésekor**

Amire figyelni kell: helyes legyen a fájlnév (ez lehet a megjelenés éve-hónapja-napja), a „Kijelölt oldalak” előtti rádiógomb legyen bekapcsolva, és a „Fájltulajdonságok”-nál az „Egyetlen fájl létrehozása az összes oldalhoz” opció legyen beállítva.

A már elmentett oldalaknál a felismerést jelző ikon mellett megjelenik a mentést jelző ikon is (*9. ábra*).

9. ábra **Az értelmezést és mentést jelző ikonok**

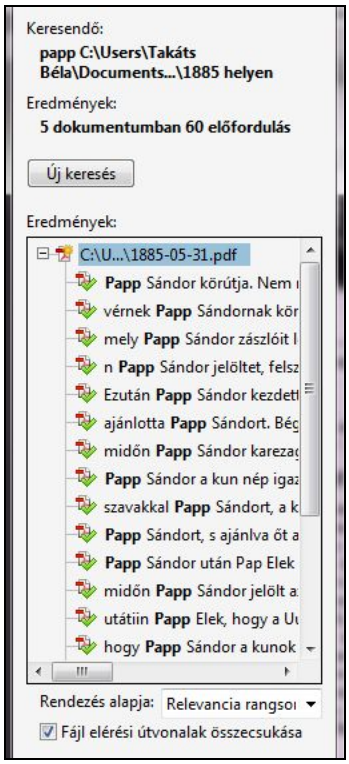
Nincs más dolgunk, mint az egyenként elmentett lapszámokat egy alkönyvtárba összegyűjteni és nem megfeledkezni arról, hogy a PDF fájlok használatára ingyenesen elérhető Adobe Reader program „Bővített keresés” funkciójával (Shift+Ctrl+F) az egy alkönyvtárban összegyűjtött összes, meg nem nyitott fájlban is lehet keresést végezni. A bővített keresés a későbbiekben létrehozandó könyvjelzőkre is kiterjed. Ne felejtsük el erre felhívni olvasóink figyelmét! (*10. ábra*).

A könyvjelzőzés

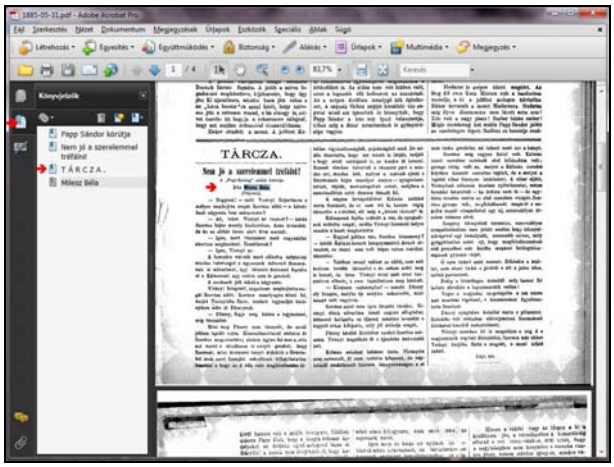
Már jeleztem, hogy a szöveg felismerése nem 100%-os. A könyvjelzők készítésének célja épp ezért az, hogy a szövegben az általunk kijelölt elemek szó szerint és betűhíven kereshetők legyenek a dokumentumban, illetve hogy ezekhez a pontokhoz gyorsan „oda tudjon ugrani” a majdani felhasználó.

Nagy körültekintést kíván a könyvjelzőzendő elemek kiválasztása. Az alábbi példában a címek, a szerzők neve és a rovatcímek kerülnek be a könyvjelzők közé, de az elemek köre természetesen igény szerint bővíthető.

Maga a munka végtelenül egyszerű. Kétrétegű PDF dokumentumainkat megnyitjuk az Adobe Acrobat programmal, a megfelelő ikonra kattintva láthatóvá tesszük a könyvjelző eszköztárat, a képernyőn kijelöljük azt a szövegrészt, ahová könyvjelzőt szeretnénk elhelyezni, majd megnyomjuk a Ctrl+B billentyűkombinációt... (*11. ábra*).



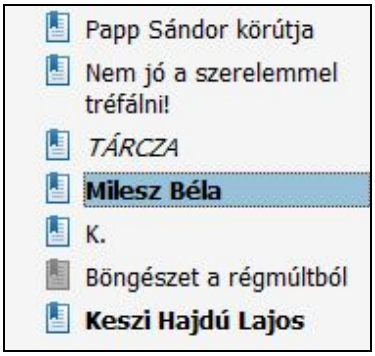
10. ábra **Bővített keresés az Adobe Reader programmal**



11. ábra **Könyvjelzők készítése Adobe Acrobat programmal**

A képen látszik, hogy például a „T Á R C Z A .” könyvjelzőt kell átszerkesztenünk TÁRCZA formára, mert a keresés csak így fogja megtalálni.

Ha az egeret az elkészült könyvjelző fölé visszük és a jobb oldali gomb megnyomása után a lokális menüből a „Tulajdonságok” pontot választjuk, a különböző kinézetű könyvjelzők megkülönböztetésére is lehetőségünk lesz (*12. ábra*).

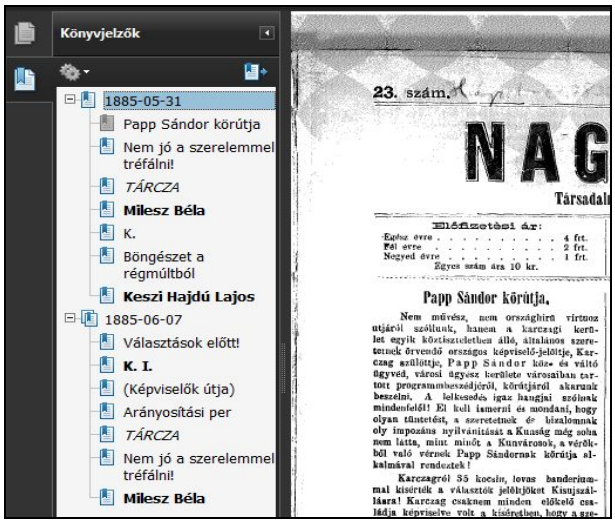


12. ábra **A könyvjelzők megjelenítési lehetőségei**

Ez esetben a szerző nevét félkövér, a rovatcímet dőlt, a címet normál betűvel különböztettem meg.

Ha végeztünk az egyes lapszámok könyvjelzőzésével, a PDF fájlok összevonásával a könyvjelzők listája is összevonható.

Ezt a feladatot szintén az Adobe Acrobat programmal lehet elvégezni. A „Fájl” – „Egyesítés” – „Fájlok egyesítése egyetlen PDF fájlban” lehetőség kiválasztása után a megjelenő ablakban (bal felső sarok) a „Fájlok hozzáadása...” pontra kattintás után kijelölhetjük azokat a lapszámokat, amelyeket egyetlen állománnyá szeretnénk összevonni. Az összevont fájl az Adobe Reader programban így jelenik meg (*13. ábra*).



13. ábra **Az összevont könyvjelzők megjelenése**

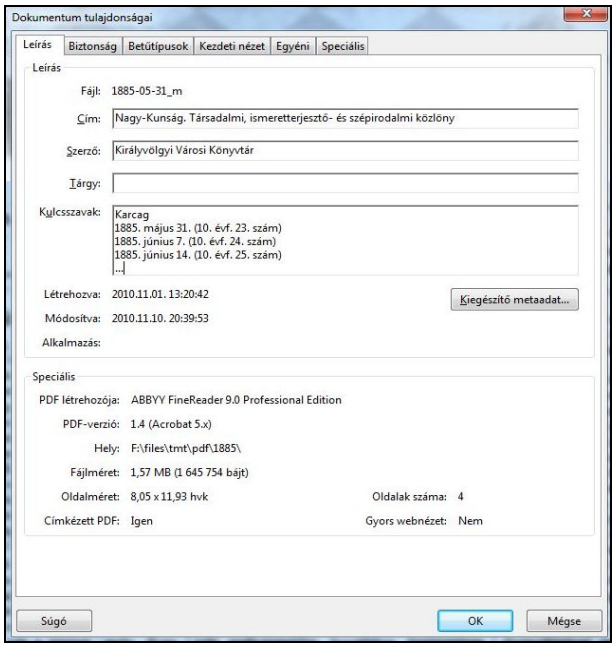
Látszik, hogy az egyes lapszámokhoz tartozó könyvjelzőket (a példadokumentumban nem minden könyvjelző készült el) a fájlok neve fogja egybe, a névadásra tehát tényleg érdemes odafigyelni és a választott formát következetesen alkalmazni! Az így létrehozott fájl mérete természetesen már jóval nagyobb lesz, összevonáskor tehát (az átlagos terjedelem és számítógépeink minőségének függvényében) laptípusonként érdemes meghatározni az összevonandó lapszámok mennyiségét (havi, negyedéves, féléves, éves összesítés).

Az így összevont fájlokat – méretük miatt – interneten keresztül már nem érdemes szolgáltatni.

Még néhány szó...

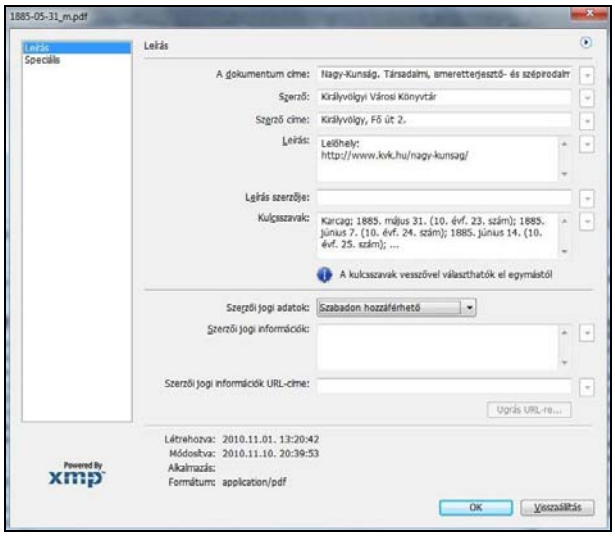
A kétrétegű PDF dokumentumok lokális kereséséről (Adobe Reader, bővített keresés) már esett szó. Ha a szolgáltatás az internetre kerül – ingyenes megoldásként – a kereséshez szeretném mindenki figyelmébe ajánlani a Google „Egyéni kereső” szolgáltatását[8], amivel – ha nem is 100%-os, de – korrekt szolgáltatás készíthető.

A hálózaton közreadott PDF fájlokban célszerű rögzíteni a dokumentum tulajdonságait. Ezt az Adobe Acrobat program „Fájl” menüjének „Tulajdonságok...” pontjával (vagy a Ctrl+D billentyűkombináció lenyomása után) az ábrán látható ablakban tehetjük meg (*14. ábra*).



14. ábra **A dokumentum tulajdonságainak rögzítése**

Az itt rögzíthető adatok köre kiegészülhet a „Kiegészítő metaadat...” gombra kattintás után (*15. ábra*).



15. ábra **Kiegészítő metaadatok**

A metaadatok rögzítése lehetővé teszi, hogy az internetes keresők a találati listában megfelelően jelenítsék meg dokumentumaink adatait, illetve, hogy a felhasználók a letöltött fájlok forrását, lelőhelyét később is azonosítani tudják. Fontos tudni, hogy ha az egyes lapszámok PDF állományait egyenként látjuk el ezekkel az adatokkal, a fájlok összevonása után a sorrendben legelső szám adatai fognak az új, összevont dokumentum egészére vonatkozni, az adatokat tehát ilyen esetben újra kell szerkeszteni.

A metaadatok rögzítésére a fentiekben természetesen csak egy példát mutattam be. Célszerű erre helyi szabályzatot alkotni és a szerint – de mindenképpen következetesen – eljárni addig, amíg ezek kezelésére könyvtárainkban ki nem alakul egy együttes gyakorlat, vagy nem születik erre egy megfelelő ajánlás.

Biztos vagyok abban, hogy a régen megjelent helyi lapok internetes közzététele nagy érdeklődést fog kiváltani a városi könyvtárak honlapjain. Tudom azt is, hogy egy ilyen digitalizálási műveletsort sokkal könnyebb megmutatni, mint leírni, de kedvcsinálónak talán megfelelnek a leírtak... Internetes fórumainkon, személyes kapcsolatainkon keresztül, a jó példák tanulmányozásának segítségével sokat tanulhatunk egymástól. Használjuk ki ezeket a lehetőségeket![9]

Jegyzetek

1 Letölthető pl. innen: *http://www.irfanview.com/*. A cikk írásakor e program 4.27-es verzióját használtam.

2 Előadásomhoz a karcagi Városi Csokonai Könyvtártól kaptam közel száz képet a Nagy-Kunság c. lap 1885-ben megjelent számaiból. Írásom elkészítésekor is ezeket használtam, itt is megköszönve a karcagi kollégák szíves segítségét.

3 A könyvtáraknak, fájloknak természetesen adhatunk ékezetes neveket is. A későbbi – interneten is elérhető – szolgáltatás zavartalansága érdekében azonban jobb, ha hozzászokunk az ékezet nélküli megnevezésekhez.

4 A képek szétválasztásának természetesen egyéb módjai is vannak. A feladat megoldható pl. a *http://scantailor.sourceforge.net* helyről letölthető, ingyenesen használható Scan Tailor programmal is. Valamilyen alternatív megoldásra valószínűleg szükségünk lesz, mert nem biztos, hogy a FineReader minden képünket megfelelően fogja kettévágni. Ha csak egy-két esetről van szó, a szétválasztás a képnézegető programmal is megoldható, de ez lassabb, körülményesebb megoldás.

5 A betölthető képek mennyisége számítógépünk kapacitásától függ. Egy alkönyvtár összes képét a Ctrl/A billentyűkombinációval tudjuk kijelölni a megnyitáshoz.

6 Elvileg lehetne próbálkozni a kontraszt és a fényerő változtatásával az IrfanView csoportos konvertálás funkciójával, de ezeknél a gyenge mikrofilmes képeknél nemigen segít. Viszont a felismerési nyelvre vonatkozó figyelmeztetésekre érdemes odafigyelni, mert lehet, hogy idegen nyelvű (pl. német) rész van valamelyik újságoldalon, és ha mi csak a magyar nyelvet állítottuk be, akkor a FR nyilván hibásan ismeri fel ezt a részt.

7 E cikk szerzője sosem a FineReader program segítségével korrektúrázott, másoknak talán ez jobban „kézre áll”.

8 L. *http://www.google.hu/cse/*

9 E cikk szerzője is sokat tanult Drótos László lektori megjegyzéseiből, javaslataiból, melyek legtöbbjét be is építette írásába. Köszönet érte!

Beérkezett: 2010. XI. 21-én.

Takáts Béla
a Jász-Nagykun-Szolnok Megyei Verseghy Ferenc Könyvtár és Művelődési Intézet szakmai igazgatóhelyettese,
az MKE Helyismereti Könyvtárosok Szervezetének elnöke.
E-mail: *tbela@vfmk.hu*