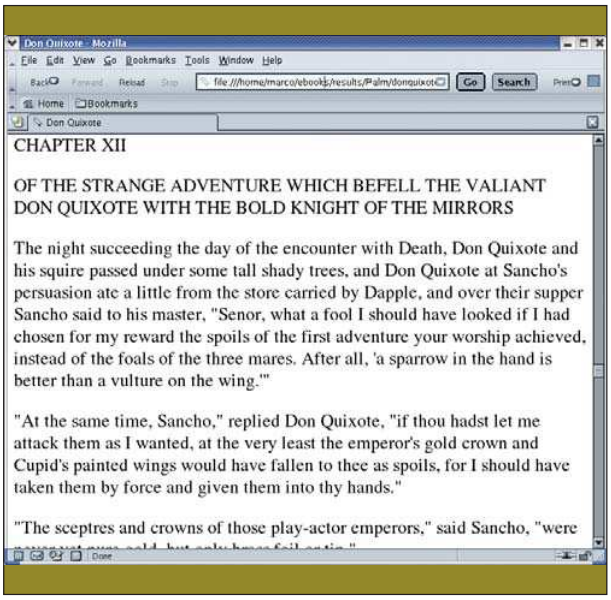
e-könyvek átalakítása nyílt formátumokra

Az e-könyvek a gyártófüggő formátumok elképesztő kavalkádját hozták magukkal. Alakítsuk hát HTML formátumúra őket, így bármilyen eszközön meg tudjuk jeleníteni tartalmukat!

A digitális formátumú könyvek, e-könyve (elektronikus könyvek, e-book) olyan készülékeken jeleníthetők meg, amelyek általában túl kevés erőforrással rendelkeznek egy normál webböngésző futtatásához. Sok kiadó, a ***Project Gutenberghez*** hasonló tervezeteket már nem is említve, több ezer új és klasszikus kiadványt jelentetett meg digitális formában. Baj egyrészt a hardverrel van – legyen szó akár általános célú zsebtitkárról, akár célkészülékről –, másrészt magával az e-könyvek kiadásával foglalkozó ágazattal, amely jóval széttöredezettebb, mint a személyi számítógépek és a webböngészők piaca. Lehet tehát, hogy megveszünk ma egy e-könyvet, és tíz év múlva már nem tudjuk elolvasni – sőt, lehet, hogy már holnap sem, ha például új hordozható gépet vagy zsebtitkárt vásárlunk. Mivel a széttagoltság elleni fellépés sokak érdeke, írásomban néhány olyan parancssori eszközt szeretnék ismertetni, amelyek segítségével a népszerűbb formátumú e-könyveket ***ASCII*** vagy ***HTML*** formátumú anyagokká tudjuk alakítani. Jelenleg gyakorlatilag nem léteznek eszközök az e-könyv formátumok ***PDF*** vagy ***OpenDocument*** formátumba (az ***OpenOffice.org*** által használt új ***OASIS*** szabvány) történő kimentésére; szerencsére ez nem jelenti azt, hogy az átalakítás ne volna megoldható. Ha egyszer a szöveg már ***ASCII*** vagy ***HTML*** formátumot kapott, szövegböngésző – például w3m – vagy egyéb program – mint a ***html2ps*** – segítségével már könnyedén átalakítható egyszerű szöveggé vagy ***PDF*** fájllá. Ha ezt az utat választjuk az átalakításra, akkor azt akár már ma is elvégezhetjük, hiszen utóbbi nyílt formátum, ahogy 20 év múlva is az lesz.

PalmDoc

PalmOS alatt az eredeti és legelterjedtebb e-könyv formátum a ***PalmDoc***, más néven ***AportisDoc*** vagy egyszerűen ***Doc***; nem keverendő össze a ***Microsoft Word*** hasonló nevű, ***.doc*** formátumával. A ***Doc*** formátumú fájlok a ***.pdb (Palm Database, Palm adatbázis)*** vagy a ***.prc (Palm Resource Code, Palm erőforráskód)*** kiterjesztésről ismerhetők fel, mindkettő lényegében összefűzött rekordokat tartalmazó ***PalmPilot*** adatbázist rejt.



1. ábra PalmDoc fájl böngészőben jobban megjeleníthető HTML formátumúra alakítva

A szabványból később több változat is kialakult, többek közt az alapszintű formátumot ***HTML*** címkékkel bővítő ***MobiPocket***.
Minden ***Palm*** e-könyv három részből áll: a fejrészből, a szövegrekordok sorozatából és a könyvjelzők sorozatából. Alapesetben a fejrész 16 bájtos, bizonyos ***Doc***-olvasók ezt futási időben, egyedi adatok tárolása céljából kibővítik. Alapesetben a fejrész adja meg többek közt a tömörítetlen szöveg teljes hosszát, a pillanatnyilag megjelenített rész pozícióját, illetve tartalmaz egy két bájtos, előjel nélküli egész számokból álló tömböt, mely az egyes szövegrekordok tömörítetlen méretét jelzi. A rekordok maximális mérete jellemzően 4096 bájt, és a rekordok tömörítése egyenként történik.
A könyvjelző rekordok egy 16 bájtos névből és egy 4 bájtos, a szöveg kezdetétől számított eltolásból állnak. Mivel a könyvjelzők elhagyhatók, sok ***Doc*** e-könyv nem is tartalmazza őket, a ***Doc***-olvasók pedig sokszor

1. kódrészlet Egyszerű Perl parancsfájl a Pyrite által kimentett szöveg HTML-lé alakítására

```
#! /usr/bin/perl
undef $/;
$TEXT  = <>;
$TEXT =~ s/\n\n/<p>/gm;
print <<END_HTML;
<html><body>
$TEXT
</body></html>
END_HTML
```

másféle, vagyis nem hordozható módszereket is támogatnak megadásukra. Az egyes olvasóprogramok saját kiterjesztései kategória, változatszám, illetve e-könyvek közötti hivatkozások megadására is alkalmasak lehetnek. Utóbbi adatok szinte mindig a *.pdb* vagy *.rc* fájlon kívül tárolódnak, vagyis ne várjuk megőrzésüket, amikor e-könyveinket átalakítjuk.

A ***Pyrite Publisher***, korábbi nevén ***Doc Toolkit*** egy tartalomátalakító eszközöket magába foglaló készlet ***Palm*** rendszerekhez. Jelenleg csak néhány szövegformátum átalakítására van lehetőség, ám szolgáltatásait ***Python*** beépülő modulok segítségével bővíteni is lehet. A ***Pyrite Publisher*** alkalmas az átalakítandó dokumentumok közvetlenül, webről történő letöltésére és a letöltött könyvjelzők közvetlenül a kimeneti adatbázisba való beillesztésére is. A csomag használatához 2.1-es vagy újabb ***Python*** szükséges, futtatása pedig parancssorból vagy a ***wxWindows*** alapú grafikus felülettel történik. A program ***Linux*** és ***Windows*** alá érhető el, forrás és bináris fájl formájában egyaránt. Ha az utóbbit választjuk, akkor ne feledjük el, hogy a legtöbb előre lefordított program a */usr* könyvtárban keresi a ***Pythont***. A linuxos változat az átalakított fájlokat a ***JPilot*** vagy a ***pilot-link*** segítségével azonnal át tudja másolni zsebtitkárra is.

A ***Pyrite*** telepítése és futtatása ***Fedora Core 2*** alatt problémátlan volt. Az egyéb, később említendő parancssori átalakítókkal ellentétben a ***Pyrite*** csak ***ASCII*** formátumba tud menteni, ***HTML***-be nem. A futtatható fájl neve ***pyrpub***. A ***.pdb*** fájlok átalakításához a következő parancsformátumot kell használni:

```
pyrpub -P TextOutput -o don_quixote.txt \
Don_Quixote.pdb
```

Ha csak gyorsan tárgymutatót akarunk készíteni egy digitális könyvtárról, akkor a ***Pyrite***-on kívül másra nem is lesz szükségünk. Ugyanakkor a kimenetet nem nehéz tovább formázni, amivel böngészőben is könnyebben olvashatóvá válik. Az alábbi, Perlben készült kódrészlet ugyan csúnya, ám tökéletesen megfelel a ***Don Quixote*** HTML változatának előállítására. (*1. kódrészlet*)

A parancsfájl betölti a teljes korábban a ***Publisher*** segítségével létrehozott ***ASCII*** szöveget, és minden alkalommal, amikor új sor karaktert talál, a helyére a bekezdés ***HTML*** címkéjét illeszti. Az alapszintű ***HTML*** formázású eredményt a szabványos kimenetre írja ki. Ha meg akarjuk változtatni az igazítást, a betűtípust vagy a színeket, akkor csak annyit kell tennünk, hogy a megfelelő stíluslapot beillesztjük a `<html><body>` sor mögé.

A várhatóan 2005 tavaszán megjelenő ***OpenOffice.org 2.0*** ***.pdb*** formátumban is képes lesz menteni a szövegeket. Ha olvasni is képes lesz az ilyen fájlokat, akkor tömeges átalakítási szolgáltatásával (***File›AutoPilot›Document Converter***) az egész átalakítási kérdés egy csapásra elegáns megoldást nyer. Kipróbáltam a szolgáltatást az 1.9.m65-ös előzetes kiadásban, de egyelőre csak egy általános ki/beviteli hibára utaló üzenetet sikerült kapnom. Remélhetőleg a későbbi változatokban már tökéletesen fog működni ez a szolgáltatás is.

A P5 Perl csomag

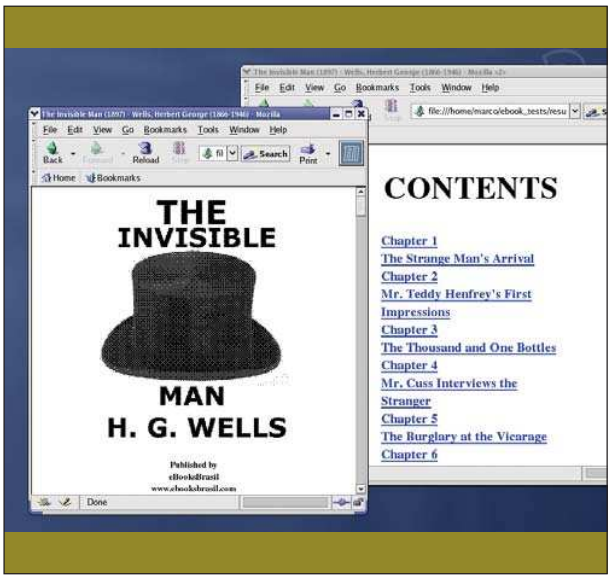
A ***Pyrite Publishert*** elsősorban normál ***HTML*** vagy szöveges fájlok ***Palm*** gépeken olvasható formátumra alakítására készítették, a több lehetőség inkább mellékhatásként fogható fel. A fent ismertetett eljárás csak nehézkesen alkalmazható, ha például nagy mennyiségű, egyedi ***HTML*** címkéket, hiperhivatkozásokat és metaadatokat tartalmazó palmos e-könyvet kell átalakítani. Ilyenkor a legjobb megoldás egy ***Perl*** parancsfájl alkalmazása, igénybe véve a nyelv szabványos ***XML*** vagy ***HTML*** és ***P5-Palm*** moduljait. A modulokat a ***Comprehensive Perl Archive Network*** (lásd az internetes forrásokat) webhelyről lehet elérni. A ***P5-Palm*** modulkészletben található osztályok segítségével a ***PalmOs*** alapú készülékek által használt ***.pdb*** és ***.prc*** adatbázisfájlok olvasása, feldolgozása és írása egyaránt megoldható.

A Rocket Ebook és a MobiPocket

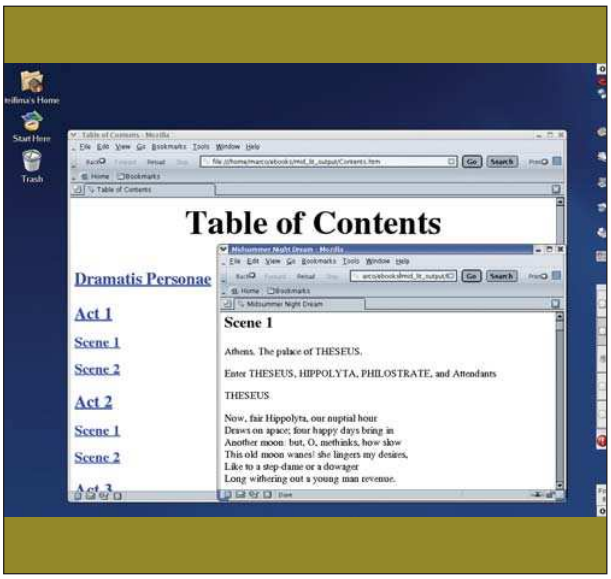
A ***RocketBook*** e-könyvek több érdekes jellegzetességgel is rendelkeznek, például támogatják a tömörített ***HTML*** fájlokat, valamint a bekezdések formázásait és a hivatkozási nevek pozícióit összegző indexeket. Ezekről és az ***.rb*** fájlok pontos felépítéséről az internetes források között található, az RB formátumot ismertető oldalon lehet bővebben olvasni. A ***Rocket Ebook*** és a ***Mobipocket*** fájlokat szétbontani az ***Rbmake*** nevű parancssori eszközkészlettel lehet. A készlet honlapján forráskódot, bináris csomagokat, levelezési listát és hibajelentések elküldésére alkalmas címet egyaránt találunk. Az `rbmake` használatához szükség van a ***libxml2 2.3.1***-es vagy újabb változatára, a ***pcre*** (***Perl-Compatible Regular Expressions***) könyvtárra, illetve a tömörítések kezelése miatt a ***zlib***-re. A forrásból történő fordítás elvégzéséhez – legalábbis ***Fedora Core 2*** alatt – külön telepíteni kell a ***pcre-devel*** csomagot is.

Az Rbmake könyvtár

Az ***Rbmake*** egyik pozitívuma modulárisan összeállított forráskódja. Ha úgy akarjuk, egy teljes objektumorientált



2. ábra Az Rbmake a RocketBook fájlok összes összetevőjét kibontja, ide értve a szöveget és a képeket is



2. ábra A Convert Lit jól olvasható HTML fájlt állít elő, hiperhivatkozásokból álló tartalomjegyzékkel

2. kódrészlet Az OPF egy XML alapú formátum könyvek jellemzőinek leírására

```
<dc:Title>A Midsummer-Night's Dream</dc:Title>
<dc:Creator role="aut"
    file-as="Shakespeare, William, 1564-1616">
        William Shakespeare, 1564-1616
</dc:Creator>
<dc:Description>fiction, poetry</dc:Description>
```

C könyvtárat le tudunk fordítani, függetlenül a csomag egyéb részeitől, így bármilyen más programból is kezelni tudjuk az .rb fájlokat. Ezzel a módszerrel saját, akár minden részletében testreszabott ***Rocket Ebook*** átalakítót is tudunk írni, esetleg az összes e-könyvünkről jegyzéket készíthetünk egy adatbázisba; minden esetben csak arra a kódrészre van szükségünk, amely az ***.rb*** formátum tényleges írásáért és olvasásáért felelős, vagyis az ***RbFile*** osztályra. Ez a kódrész megnyitja a fájlt, visszaadja a könyvet alkotó szakaszok listáját, majd futás közben csak a főprogram által ténylegesen igényelt részeket bontja ki. Ha éppen arra van szükségünk, a könyvtár keresési és cserélési eljárásokkal is rendelkezik, ezeket a ***Perl***-ben megszokott reguláris kifejezésekkel tudjuk használni.

Az Rbmake eszközöknek minden korszerű GNU/Linux terjesztésen gyorsan és gondok nélkül le kell fordulniuk. A forrás .tar állományban részletes HTML leírást is találunk. A HTML fájlok előállítására alkalmas bináris fájl az rbburst. A program az eredeti .rb konténerben lévő összetevők – szöveg, képek és leíró adatok – mindegyikét kibontja. A 2. ábrán két Mozilla ablakban látható, hogy ***H. G. Wells The Invisible Man*** című könyvén az `rbburst`-öt futtatva milyen fedlapot és tartalomjegyzéket kapunk.

Microsoft Reader

A ***Microsoft Reader*** fájlokat a ***.lit*** kiterjesztésről ismerhetjük fel. Számos a hagyományos könyvekre jellemző tulajdonságuk van, mint például oldalszámok, kiemelések és jegyzetek. A kulcsszó alapú keresést és a hiperhivatkozásokat is támogatják, ám egyetlen olvasóprogramhoz kötődnek.

Az ilyen fájlok átalakítására szolgáló eszköz neve spártai egyszerűséggel ***Convert Lit***. A programot a `-help` kapcsolóval futtatva – hűen a unixos hagyományokhoz – az összes parancssori kapcsolót megtekinthetjük.

A program háromféle üzemmódot ismer: ***robbantás (explosion)***, ***lefelé átalakítás (downconversion)*** és ***dedikálás/beleírás (inscribing)***. Meglévő ***.lit*** fájlt ***OEBPS***-megfelelő csomagba a robbantás móddal tudunk átalakítani. Az ***OEBPS***-ről (***Open eBook Publication Structure***) később még lesz szó.

A *3. ábra* ***Shakespeare A Midsummer's Night Dream*** című művének a ***Convert Lit*** programmal robbantott változatát szemlélteti. A lefelé átalakítás ezzel ellentétes eljárás, általa ***Microsoft Reader***-megfelelő eszközökön használható ***.lit*** fájlt kapunk. A beírásnál a lefelé átalakítás közben felhasználó által megadott feliratot csatolunk a ***.lit*** fájlhoz. A pontos szintaxist illetően a program honlapján találunk részletes tájékoztatást (lásd a forrásokat).

Már volt róla szó, hogy a ***Convert Lit*** egy különböző fájlokból álló ***OEBPS*** csomagot állít elő. A fenti példánál a teljes fájllista a következő: ***Contents.htm, copyright.html, ~cov0024.htm, cover.jpg, MidSummerNightDream.opf, MobMids.html, PCcover.jpg,***

PCthumb.jpg, *stylesheet.css* és *thumb.jpg*. A *HTML*, a *CSS* és a *JPG* fájlok különösebb meglepetést nem jelentenek, de vajon mi célt szolgál az *.opf* fájl? Ez egy *XML* konténer, ez írja le az eredeti könyvben található metaadatok egyes részeinek szerkezetét. Az *OPF* kiterjesztés az *electronic book package formatra*, az elektronikus könyv csomag formátumra utal. Az *OPF* az e-könyv egyéb részeire mutató hivatkozásokat is tartalmaz, illetve magába foglalja jellemzőik leírását is. Szerepét könnyebben átláthatjuk egy példán keresztül, ezért a 2. *kódrészletbe* bemásoltam a *MidSummerNightDream.opf* egy kisebb részét.

A gyakorlatban mindebből az fakad, hogy a *Convert Lit* akkor is jól használható lehet, ha teljes gyűjteményünket zárt, gyártófüggő formátumban akarjuk hagyni. Ekkor elég futtatnunk a programot az összes *.lit* formátumú e-könyvünkön, majd az *.opf* fájlok kivételével törölnünk mindent. Ezt követően egy rövidke parancsfájllal vagy akár egy nagyobb tudású *XML* feldolgozó programmal végig tudunk menni ezeken, és az általunk kívánt adatbázisba tudjuk másolni a tárgymutatót.

A *Convert Lit* a *digitális jogkezelő (digital rights management, DRM)* részeket is eltávolítja az e-könyvekből, legalábbis ami a régebbi, *DRM1* változatot illeti. Ha *Microsoft Reader* e-könyveket gyűjtünk, valószínűleg *Microsoft Windows* operációs rendszerrel és a *Microsoft Reader* egy jogtiszta példányával is rendelkezünk. A *Convert Lit* weboldala szerint a *Convert Litet Windows* alatt lefordítva, a *Windows DRM* segítségével az újabb *DRM5*-öt alkalmazó e-könyveket is át tudjuk alakítani *DRM1*-essé.

Tömeges átalakítás

Eddig inkább parancssori átalakításokról volt szó. Aki viszont nagyobb, többféle formátumból összeálló e-könyvgyűjteménnyel rendelkezik, az egyetlen héjparancsfájllal egyszerre is elvégezheti ezek átalakítását. Mint már láttuk, ha a szöveg átkerül *ASCII* vagy *HTML* formátumba, a határ a csillagos ég. A ciklust egy-két sorral bővítve a tárgymutató elkészítése a `glimpse` vagy a `ht::dig` használatával is megoldható; akár mindent beírhatunk egyetlen *PostScript* könyvbe és így tovább.

OEBPS

Jelenleg még fejlesztés alatt álló megoldás e-könyvek – legalábbis a közeljövőben beszerezhető példányok – nyílt formátumba alakítására. Pontos neve *Open eBook Publication Structure (nyílt e-könyv kiadási szerkezet, OEBPS)*. Célja egy *XML*-alapú, a meglévő nyílt szabványokra épülő szabálygyűjtemény összeállítása többféle e-könyv rendszeren történő tartalomszolgáltatáshoz. Az *OEBPS*-t, mely immár 1.2-es változatban érhető el, az *Open eBook Forum* tartja karban. A csoportban több mint 85 szervezet található meg, hardver- és szoftvergyártók, kiadók, szerzők és az elektronikus kiadványok piacán érdekelt felhasználók egyaránt. Az *OEBPS* önmagában, közvetlenül semmilyen digitális jogkezelést nem szolgál. Az *OeBF Rights and Rules Working Group (jogok és szabályok munkacsoport)* azonban szorgalmasan tanulmányozza a vonatkozó kérdéseket, munkájának célja „egységes, kölcsönösen támogatott előírásgyűjtemény összeállítása az elektronikus kiadói közösség számára". Még meglátjuk, mi sül ki belőle.

Bármi is történjék, az *OEBPS* hátteréként szolgáló nyílt szabványok szilárd alapokra épülnek. Az *XML*, az *Unicode* és az *XHTML* mellett a *CSS1* és a *CSS2* bizonyos részei is szerepelnek benne. Az *Unicode* kódolások családja, alkalmazásával több tízezer karaktert is félreértések és keveredések nélkül lehet kezelni. Az *XHTML* a *HTML 4* új, *XML* alapú megtestesülése. Az *OEBPS* röviden talán úgy jellemezhető, mint az *XHTML* e-könyvek kezelésére testreszabott változata – olyasvalami, ami akkor is fennmarad, ha a támogatását adó vállalatok némelyike esetleg el is tűnik az üzleti élet porondjáról. A képek *PNG* vagy *JPEG* formátumúak lehetnek. A metaadatok (szerző, cím, *ISBN* stb.) kezelése a *Dublin Core* szótáron keresztül fog történni.

Az *OEBPS* tehát alkalmas arra, hogy az összes e-könyvünket megőrizzük, és biztosítsuk időtállóságukat, még akkor is, ha némelyik hardver- vagy szoftvergyártó neve néhány év múlva homályos emlékké válik. Sajnos az ilyen „nyílt" e-könyvekre alkalmazott *DRM* megoldások egy-egy gyártóhoz köthetnek bennünket. Amíg az *OEBPS* e-könyveket *DRM* nélkül tudjuk beszerezni, addig az *OEBPS* a legjobb megoldás annak biztosítására, hogy gyűjteményünk még akkor is használható maradjon, ha az ezen a piacon érdekelt cégek akár mindegyike a süllyesztőbe kerül.

Linux Journal 2005. június, 134. szám

Marco Fioretti hardver-rendszermérnök, a szabad szoftverekkel mint EDA alaprendszerekkel és a hatékony asztali környezet létrehozására irányuló RULE Project vezetőjeként áll kapcsolatban. Marco Olaszországban, Rómában él a családjával.

KAPCSOLÓDÓ CÍMEK

- www.cpan.org
- www.linuxmafia.com/pub/palmos/development/doc-format and www.timwentford.uklinux.net
- www.blorf.net/~wayne/rb_format.html
- www.pyrite.org/doc_format.html
- www.pyrite.org/publisher
- freshmeat.net/projects/rbmake
- www.kyz.uklinux.net/convlit.php
- www.openebook.org/oebps/oebps1.2/index.htm
- userpage.fu-berlin.de/~mbayer/tools/html2text.html
- user.it.uu.se/~jan/html2ps.html