Ebook software 2010

Jon Jermey*

Double-digit annual financial growth and the arrival of major players like Amazon and Apple have finally shaken the ebook market out of the financial doldrums. Old formats are being redesigned, and new ones developed, in an attempt to reach out to paying customers. But the "Tower of eBabel" resulting from multiple incompatible and constantly changing formats – compared, for example, with the music industry, now finally settling on MP3 files – indicates that ebooks still have a long way to go.

Soon someone will do a reliable poll and find out what proportion of our time looking at screens is spent reading different kinds of documents. My best guess right now is something like this:

- Emails – 40%
- HTML webpages – 40%
- PDF files – 8%
- Proprietary word-processing formats (eg MS Word) – 8%
- Plain text – about 2%
- Other formats – around 2%

In other words, despite two decades of ebook development, specialised ebook formats and software are still marginal players in the global computer environment. That is not to say that some people are not spending a lot of time with their Kindles, Iliads and Sony Readers, just that we should not expect to see ebook distribution formats spawning a software colossus like Microsoft or Google any time soon.

So with that in mind, let's take a look at the state of play regarding ebook formats and the associated software.

GENERIC FORMATS – EPUB

The closest thing we currently have to a standard ebook format is the EPUB specifications developed by the International Digital Publishing Forum (IDPF) and released in 2007, replacing the older Open ebook format. Designed as a free and open-source standard that would permit copy-protection, EPUB uses a system of "containers", with the inmost container holding the book in XML format, and the outer containers modifiable by the book distributor to allow for copy protection and other access controls. The Sony Reader, the Apple iPhone and several other readers support EPUB directly, while many other companies have incorporated modified versions into their own software. The EPUB format allows the inclusion of metadata, and supports the use of CSS style sheets for formatting control. EPUB text is reflowable around images and across pages of different sizes.

TEXT FILES (ASCII)

Text format files remain the lowest common denominator for transferring written material among all computer platforms. In fact, text is even more widespread than it first appears, because many formats, including RTF, HTML and XML, are simply text with formatting tags added. Text files as such are, of course, totally insecure, and unformatted text is less than optimal for reading anything but short blocks of material. However, it is extremely compact and – unlike some proprietary formats – pretty well guaranteed to still be around in 50 years' time.

Most PDAs and mobile phones can read text files, though they may require third party software (eg Pocket MS Word) to do so. The Kindle is reported to display text files oddly, but there are free utility programs available for the PC that can pre-process the text to prevent this. The Sony Reader will display text files out of the box, as will the Iliad.

* Jon Jermey is an indexer, computer trainer and writer. He is currently moving into the exciting world of Linux training. See http://www.learnubuntu.com.au.

PDFs

The PDF file format is nearly old enough to vote. It was developed by Adobe in 1993, primarily as a way of representing on a screen what would (or should) otherwise come out of a printer. Because it allows the contents of a document to be standardised and frozen, PDF has become popular as a document distribution format, but from an ebook perspective it has many drawbacks. Text does not normally flow when a PDF page is resized, making it difficult to read on small screens; and although this can be achieved through optional settings, the process is time-consuming and can introduce errors.

Ironically, as PDF files have become more popular because they could be frozen, there has been a growing demand for software that can unfreeze them, allowing users to make changes and add comments and marginalia. Adobe themselves have led the way with their Acrobat program, but there are many other commercial and freeware PDF editors available. Adobe have attempted to maintain author control by allowing file creators to lock the files with passwords that limit what the user can do – read-only, read-and-print, read-and-copy-but-not-print etc – but these are easily cracked and provide little security.

In recent years, Adobe has expanded the role of PDF files by allowing them to act as slide shows and picture galleries, as well as permitting the inclusion of other files as attachments. PDF became an open format in 2008, allowing developers outside Adobe to legally tinker with PDF file codes and develop their own formats and software.

Most PDAs and small PCs will display PDF files with no difficulty, using either their own programs or the Reader program, available as a free download from Adobe. The original Kindle software could be modified by users to display PDF files, and this modification has now been built into the Kindle 2. The Sony Reader and the Iliad support PDF formats out of the box, although there have been complaints on the web about displayed text and image quality and the speed of page turning. Currently, the Apple iPad supports PDF through third-party display software, similar to that which is currently available for the iPhone.

HTML

As mentioned above, HTML is simply text with embedded formatting codes, so any device that displays text will also display HTML; however, to process the codes and show formatting, links and images, a web-browser-type application is required. Most PDAs and mobile phones now come with one of these, usually a scaled-down version of Microsoft Internet Explorer, and some independent browser distributors such as Opera and Mozilla are now allowing genuine competition by making their products available for these devices.

The iRex Iliad will display files in HTML and its extended subset, XHTML. The Kindle will not display HTML as such, but it is possible to email an HTML file to Amazon and have it converted (for a fee) to a readable form. The Sony Reader provides users with a program for their PC that can convert HTML files into a readable format before they are moved onto the device.

XML

After a bright start a few years ago, enthusiasm for ebooks in XML (eXtensible Markup Language) appears to be lagging a little. Like HTML, XML is based in plain text, but unlike HTML, which is a genuine language, XML is more of a grammar, which developers can use to specify schema for particular types of data. The XML approach treats an ebook as a set of nested containers: books contain chapters, chapters contain paragraphs, paragraphs contain sentences, and sentences may contain formatted text in italics and bold, for instance. The properties of containers at each level can be specified, and any particular document tested (parsed) against these to check whether it matches the specifications. The EPUB format is based on XML, and HTML itself can be seen as a kind of free-form XML version. Most applications that display HTML will also display XML, although not necessarily in a readable form.

Mobipocket and Plucker formats

Mobipocket is one of the oldest ebook formats, and is based on an HTML file bundled into a container that can then be locked to a particular device, preventing file copying – although this locking has, of

course, been cracked. Mobipocket books are supported on PDAs and mobile phones by Mobipocket's own free software, and natively on the Kindle and the Iliad, but not on the Sony Reader. Mobipocket, a French company, was bought by Amazon in 2005, but continues to maintain a separate website and branded products. Mobipocket books and reading software are also available for PCs, and emulators are available to run it on Macintosh and Linux systems. Plucker is a similar, but less widely used, system developed for Palm PDAs and mobile phones.

Because of its versatility, the Mobipocket format – minus DRM – is often used for distributing public domain ebooks. For instance, both Mobipocket and Plucker formats are now available alongside plain text, HTML and EPUB for most of the free ebooks on the Project Gutenberg site.

PROPRIETARY READER FORMATS

The Amazon Kindle uses its own native AZW format, and Amazon offers a paid email-based service whereby other formats can be converted to AZW. It appears to be a simple modification of the Mobipocket format, and scripts are available to convert (though perhaps illegally) back and forth between the two. The Kindle also uses an internal format called Topaz, which appears to be an implementation of EPUB. The Sony Reader originally used a proprietary format called BBeB (BroadBand ebooks), but all new Sony ebook sales have been in the open EPUB format since late 2009.

DIGITAL RIGHTS MANAGEMENT

It is only in the last two years that ebook distribution has become a serious commercial proposition, in the United States at least, and both distributors and customers are rapidly raising their expectations as a result. Unfortunately, at a time when most music distributors are switching to unrestricted MP3 formats, some book distributors still cling to the hope that they can prevent unauthorised copying through proprietary formats involving passwords and encryption schemes, known generically as Digital Rights Management, or DRM. Customers are understandably resistant to these, especially as they are invariably cracked by hackers within a few days (or hours) of release, often necessitating further changes to the schemes and further annoyance for consumers.

One more enlightened approach to the issue is social watermarking, where an ebook is indelibly marked with the name of, or other identifying information from, the person who purchased it, making it possible in theory to name and shame those who make unauthorised copies. Another is the Digital Entertainment Content Ecosystem (DECE) – a grandiose title for a scheme whereby one purchaser can license the right to read/play/show a particular file on any device that they own. While these may avoid some of the consumer inconvenience normally associated with DRM, they too can be (and have been) quickly cracked.

Amazon, which currently dominates the market, has adopted a Big Brother approach: ebooks purchased and downloaded by Kindle users are checked against a master list when the user accesses the Amazon site, and any unauthorised versions can be removed without the reader's knowledge or permission. This made headlines in July 2009, when Amazon remotely deleted thousands of users' copies of books after it learnt that it did not hold the rights to their distribution (ironically the books concerned were Orwell's *Animal Farm* and *Nineteen Eighty-Four*). Kindle users quickly learnt to keep backup copies of their Amazon books on their own PCs, effectively neutering Amazon's powers of recall.

Amazon has now softened its terms, promising to delete books without permission only when instructed to do so by a court order; but the event serves as a reminder that under this kind of system, the purchaser never really *owns* an ebook – merely the licence to read it. Less well publicised, but equally damaging in the long run, was the collapse in January 2009 of Overdrive, the company that held decryption keys for ebooks sold by the popular Fictionwise site, making those books impossible to legally transfer between devices or replace if lost.

By contrast, technical publisher O'Reilly saw an increase of 105% in yearly ebook sales after dropping their DRM scheme in 2007. Sony also appears to have seen the writing on the wall, and dropped their encrypted BBeB format for the Sony Reader in August 2009, in favour of the EPUB

format. Sony ebooks can now legally be transferred to any other reader device that supports this system, although unlimited copying is still blocked. The iRex Iliad supports the Mobipocket format, which incorporates DRM.

EBOOKS AND BUNDLING

One of the most interesting phenomena in music distribution since the arrival of MP3 files has been the slow death of the album. Whatever cultural reasons there may be for putting together multiple tracks by the same performer, most listeners seem to regard it as a way of being forced to pay for tracks that they do not want alongside those that they do. Once online, and given the choice of downloading albums or individual tracks, they usually opt for individual items. At the same time genre CDs, which bundle together an assortment of tracks by different performers from a single period, or on a single theme, are becoming increasingly popular.

Are there implications here for ebooks? Something of the sort is already happening with technical books; for instance, CengageBrain, a division of Thomson, currently sells single textbook chapters for download (though not in Australia, alas) for US$1.99. Without any constraints on size and weight, there is no reason why the books we now regard as integral units could not be broken up and sold in separate parts; or conversely, why many individual physical books cannot be bundled into a single ebook if it makes sense to do so. For instance, many university courses that require course packs of multiple readings are now compiling these together in an electronic package.

AND THE FUTURE?

With steady double-digit growth in the ebook market there is profit potential for software developers. The next decade may see the same growth in the power and sophistication of ebook software that we saw in the 1990s for word processors and spreadsheet programs. The Tower of eBabel resulting from the proliferation of proprietary formats remains a problem, but if the music industry is anything to go by this will quickly resolve itself if and when the paying public shows its preference for open, DRM-free formats. At the same time as the number of ebooks on general purpose sites increases, we may see specialist ebook distributors appearing, focussing on particular kinds of books the way that some websites now focus on distributing particular genres of music. It is still the case that only a tiny fraction of all books ever published are available electronically, so there is plenty of room for new distributors.

Next issue's article will cover ebook distribution, and the multitudinous ways that ebook consumers can find that elusive volume.