Representing Books Digitally

Mike Clarke
Adobe Systems Benelux BV
Amsterdam, The Netherlands

Abstract

As a result of the availability of digital printing devices and automated finishing equipment the digital production of books is already a feasible alternative to more traditional processes. And it seems that in some circumstances this new approach can offer economic advantages over conventional production systems. It is suggested that in most cases the most appropriate way to store the digital master of a book is as an Adobe® PDF file, since this format is able to provide an exact representation of any book page while also providing flexibility lacked by other formats. Features recently added to the format reinforce this argument, providing a basis for future developments in the digital book industry, covering both the print-on-demand and eBook business models.

Introduction

The conventional process for manufacturing and distributing books involves printing and binding a batch of copies, storing them in a warehouse, and distributing them through a supply chain that may involve both wholesaler and retailer. Given that the high set-up costs of offset printing imply large batch sizes, this process can be costly in terms of capital employed in stock holding, storage costs, transport costs and wastage (pulping of unsold books).

It is little wonder, then, that the book publishing industry is investigating the use of just-in-time manufacturing techniques. On the face of it, there is considerable potential for cost savings to be achieved both by manufacturing books in much smaller batch sizes (perhaps even a batch size of one), and locating the production facility physically closer to the location where the books are to be sold. Storing a book in digital form, transmitting it over networks and using digital printing coupled with automated finishing makes such an approach possible.

This paper provides a brief summary of the arguments for books-on-demand, and discusses in detail one specific technology component: the file format used to store the digital master of the book.

The Business Case

It is clear from any analysis of short-run book publishing that there are many alternative business models that might be implemented. The range extends:

- from the installation of a digital print device alongside traditional offset presses to handle small batch sizes when a book's sales do not warrant the use of offset printing,
- through the various alternatives in which printing and binding are moved further down the supply chain, to the wholesaler or even into the bookstore,
- to the extreme in which the traditional supply chain is dismantled, and the publisher takes orders from the customers over the web, manufactures the books himself and delivers them by post.

A recent study[1] by PIRA examined the alternative approaches taken by publishing, printing and retailing companies who are already actively involved in short-run book publishing, The study finds examples of the use all of the above models, and more. It analyses the strengths and weaknesses of each approach, and provides a financial model that can be used by companies to simulate the financial effect of implementing the various supply chains, using their own cost structures as data.

The general conclusion of the PIRA study is that although it is not always easy to quantify the savings exactly, there appears to be an economic argument for digital production for books with sales of less than about 1,000 copies per year. The savings may be correspondingly greater if the opportunity is taken to redesign the supply chain into one of the more radical models discussed above.

Note that a given title will pass through several stages during its lifecycle, with corresponding variations in sales volume. It may be that print-on-demand has a role to play in the early and late stages of the sales lifecycle of even a best-selling title that will benefit from the economies of scale offered by offset printing during its phase of peak sales. Print-on-demand's ability to address small batch sizes offers a way to lessen the risks associated with predicting future sales volumes at the launch of an new title, and also provides an economic way of keeping alive a publisher's back list.

So, it appears that there may be financial reasons to consider books-on-demand. But what of the technology? The latest advances in digital printing and finishing will be covered elsewhere in this conference. The remainder of this paper concentrates on the issue of file formats.

Background

Over the past decade, publishing workflows have been transitioning to a complete reliance on digital representations. This has been motivated by the desire to automate production in order to reduce costs, the availability of broadband networks for the transmission

of files, and by the evolution from film-setters to computer-to-plate systems. This last factor has been particularly important in driving the adoption of fully digital publishing systems and as a result, many of the technologies and procedures required for digital book printing have already been developed and tested in computer-to-plate workflows.

Computer-to-plate systems require digital representation of the documents to be imaged, and so from the point of view of data requirements it is but a small step from driving a CTP system to driving a digital press.

File Formats for Digital Books

There are three main methods available for the digital representation of book content:

1. as page-image bitmaps, usually coded as TIFF images
2. as an object-oriented page description such as Adobe Portable Document Format (PDF)
3. as tagged text expressed in a markup language such as SGML or XML

TIFF Images

These typically originate from scans of the pages of a printed book. They have the advantage of representing the original pages exactly, and can be printed easily. But they are completely inflexible since the page content cannot be accessed either for searching or for repurposing to other publications or to other media. File sizes tend to be large. Without special treatment, scans of half tones can be problematic.

PDF

PDF has become an established standard for the representation of digital documents, and it plays a similar role to that which PostScript played in the 1980s in providing one of the enabling technologies for the Desktop Publishing revolution. In fact PDF is closely related to PostScript, using the same imaging model, but is packaged in a much more robust file format.

PDF describes the appearance of a document in terms of the text, line art and images that go to make up each page, and uses vector representation where possible, leading to a resolution-independent description of the document.

Markup Languages

The principle of tagging texts to represent their logical structure has been in use since the early 1970s and gained wide acceptance in specific application areas with the development of the SGML standard. Today, XML is the preferred representation, although the principle remains the same—the representation of the content of the document is separated from its appearance. This is achieved by encoding the document's logical structure in the XML file, resulting in a very flexible description of content that can be formatted in different ways for different media.

File Formats for Print-On-Demand

When printing books, it is clearly essential to have complete control over the layout of the pages. This requirement mostly rules out the use of **XML** to represent the book content in a print-on-demand system. While the layout of an XML document *can* be specified using XSL or Cascading Style Sheets, the degree of control is currently not sufficient to represent anything more than simple typographic features. Graphics support is also limited. And since the conversion from document structure to page layout is performed on the fly at print time, the exact appearance of the printed output is not guaranteed to be predictable.

The appropriate use for markup languages is generally earlier in the workflow, as a way of storing the raw text of a book before page layout. Once the book has been paginated, PDF is the preferred way to store it in a print-ready state. But it turns out that there also is much to be gained from having access to the logical structure of the digital document at later stages of the workflow, and this will be addressed later in this paper.

PDF, on the other hand, is designed to reproduce the exact appearance of printed pages, and its rich imaging model allows any page to be printed exactly as was intended. Since its introduction in 1993, PDF's feature set has been developed to the stage where it is capable of being used in almost any commercial print workflow.

Amongst the reasons for PDF's widespread adoption are:

- it provides a convenient package for distributing digital documents over networks—file sizes are optimised through the use of compression, and everything that is required to print the document (including the fonts) is included in a single file.
- PDF is simple to generate from any application that can generate PostScript output, and it can be printed either directly or through Adobe Acrobat® on any PostScript output device.
- the document representation in a PDF file is not locked to any particular output device. So for example, today's PDF files will be able to take full advantage of future advances in digital print technology, such as higher resolutions.
- a PDF document can be viewed and printed using Adobe Acrobat on a range of platforms.
- Adobe Acrobat is extensible, and a range of plug-in tools for manipulating PDF files specifically for the professional publishing market is available from Adobe and from third parties.

Compared to PDF, **scanned pages** tend to produce larger file sizes, and are much less flexible. For example, while the text in a PDF document is searchable, that in a scanned image it is not. Nevertheless, printing on demand from page scans is a viable alternative, particularly when no digital representation of the book is available.

When starting from a printed copy the alternatives are:

- scan the pages and print from the page images
- scan the pages and use OCR to convert the page image into text
- rekey the whole text

…the second and third alternatives being most suited to text-only publications. It is interesting that in some cases rekeying can be less costly than OCR because of the manual work required to check the text in the latter case.

eBooks

eBooks represent a radical new business model for publishers that has generated a sometimes rather exaggerated wave of interest over the past year. Although eBooks apparently lie outside the scope of this conference, it may be a useful exercise to compare the eBook model with that of print-on-demand. Both involve expressing the book in digital format, and storing it in some kind of repository, which may or may not be distributed geographically. The difference comes only in the final stage of the distribution chain, where in one case the file is printed, bound and sold to the customer as hard copy, while in the other case the digital file itself is sold to the customer for on-screen reading.

The considerable overlap between the eBook and print-on-demand workflows suggest that there should be opportunities to share resources and developments between the two. So far as file formats are concerned, it is certainly feasible to consider using the same format for both eBook and print-on-demand services.

The two leading contenders for the standard for representing eBooks are Adobe PDF and Microsoft's *.lit* format. The latter is based on the OEB (Open Electronic Book) format, which is in turn based on XML. So OEB falls into the class of markup languages discussed above. It is made up of tagged text that describes the logical structure of the book, with some extra data to define aspects of layout. An OEB eBook is paginated on the fly in the eBook Reader application.

When comparing print-on-demand with eBooks it is possible to construct a matrix of the preferred formats for each kind of workflow:

	Print-on-demand	**eBooks**
Scanned images	X	
PDF	X	X
XML		X

So we can see that storing digital books as PDF files provides the opportunity to use them for either print-on-demand or eBooks.

Digital Rights Management

An important feature of eBook systems is the implementation of Digital Rights Management (DRM) systems that control and track the legal rights that are attached to the book. This involves encrypting the eBook file and providing keys that ensure that the book can be read only on specific hardware devices, thereby preventing unauthorised copying. In addition, detailed rights can be assigned, controlling, for example, the extent to which a particular customer is allowed to print from a particular book. Further, the DRM system maintains records of what payments are due to the various actors in the value chain as a result of each transaction.

The motivation for using DRM systems in an eBook system is obvious. Without them, it would be difficult to prevent the illegal distribution of content and to protect the interests of the copyright holders.

A similar situation exists in the case of print-on-demand. Once books are stored and distributed in digital form, there is little to prevent unauthorised copying and/or book production. And it is necessary to record all transactions and to keep track of what payments are due to which players. It may be that the DRM systems that are being developed for eBooks have application also for print-on-demand systems.

New Features of PDF

A number of new PDF features have recently become available, some of which are particularly relevant to digital books. Two are described here: the ability to code logical structure within a PDF file, and the support for metadata.

Logical Structure

As we have seen above, XML is basically a representation of the logical structure of a document, whereas PDF is principally a representation of its visual appearance. Both representations are valuable, and it would be very powerful to be able to package both in a single file. In fact PDF does support the coding of logical structure within the same file as the page descriptions, and Acrobat 5 includes some new tools that make use of this functionality.

Using this approach it is possible to create a PDF file that not only describes the appearance of the document, but also describes its structure. For example, a particular piece of text in the file will be described as being "Times Bold 14 point" in the appearance representation, and also as a "sub-heading within chapter 6" in the structure part.

With this dual representation, a digital book becomes much more flexible. When used as an eBook, it allows *reflow*—the feature whereby text can be rearranged to fit the current size of the display window. As the window size is changed, words move from line to line, within the current page.

The content of a structured PDF file can also be repurposed into other formats…by default in Acrobat 5 into RTF format.

Metadata

A digital book file needs to contain other information apart from the actual page content. Examples include the ISBN, copyright information, the publisher's and author's details and so on. Various standards bodies are working to establish international standards for the representation of this metadata. Acrobat 5 supports the embedding of metadata in a PDF file using an encoding based on XML schema. This means that the system is extensible and that it can be used with various standards.

Conclusion

Digital books offer intriguing opportunities for the publishing industry. At one extreme they can be used as input to digital print devices to provide a cost-effective way of printing short runs of books whose sales volumes do not warrant production in larger batches. At the other extreme they can be sold directly to the consumer as eBooks. And in between there is a myriad of possible business models that reshape existing printed book supply chains.

A key decision involves how to hold the master file that represents the digital book. Adobe PDF provides an established solution that suits all workflows and which is particularly valuable when complex typography and graphics are involved.

Going forward, the ability to embed document structure in a PDF file adds considerable flexibility to the digital book, offering the opportunity to repurpose and therefore resell content.

References

1. John Birkenshaw, Books on Demand—Digital Production Strategies and Cost Models, PIRA International (2000)

Biography

Mike Clarke holds a BA degree in Physics from Oxford University. He developed typesetting software on the first IBM PCs and worked on a clone PostScript interpreter before joining Adobe Systems in Amsterdam in 1990. Within Adobe he has been responsible for the technical support of PostScript and PDF developers in Europe, and is currently focusing on the developing eBook market.