Converting a Legacy Print Book to an EPUB with Pinpoint Index Linking

By David K. Ream

David K. Ream has worked over 35 years with publishers in the areas of typesetting design and production, databases, editorial systems, indexing, and electronic publication design and production. Currently, Dave is one of the three Co-Chairs of ASI's Digital Trends Task Force.

In the electronic realm, page numbers lose their meaning. What is a page? It can vary depending on the way you choose to display the material on your eReader based on the font and type size. The index locators for page 10 have no unique meaning for two viewers of the same eBook.

A really simple solution is that eBooks should be linked not to pages but to more specific points within the text. New tools are likely to be developed to assist publishers with accomplishing this for new publications. But how can this be accomplished for already published books? This article will talk about a real-life project involving such a scenario and describe the workflow and software involved, and the pitfalls and unexpected issues that arose.

Getting Started

The starting point of the process was easy enough. The publisher sent the PDF file used to print the book for conversion. Off it went to a conversion service, which returned an eBook. (There are two major file formats for eBooks: MOBI, which is used on Amazon Kindles and EPUB[1], which is used on most other eReaders. This article describes work that was done on an EPUB file.)

Both the PDF and EPUB indexes have page numbers. During the creation of the EPUB, the conversion service inserted anchors for each page. The page break anchors are something you cannot see on the screen, but they are in the coding allowing the locators to link to page breaks from the Table of Contents or the Index.

Before going further, it is worth mentioning that an EPUB is meant to be a dynamically reflowable document because it can be viewed on an eReader (e.g., Nook), on a PC or Mac (using software, e.g., Calibre[2]), on a tablet (e.g., iPad), or even on a mobile phone. It is not meant to be an exact replica of the book's layout[3]. In addition, it can have totally new content including rich media (audio, video). This discussion, however, is confined to conver sion of legacy publications, not new books or new features.

The reflowable quality does affect some conversion results though. In this converted book, the dot leadering was not kept in the Table of Contents. The page numbers were just one space away from the text (chapter heading, for example) to which they applied. These small imperfections in conversion can lead to user confusion, especially if the chapter heading ends with a number as well. A Table of Contents example of this would be:

The Top 10 24

As an Index entry, it is much clearer when a comma is used as a separator:

The Top 10, 24

There are many other quirks like this that introduce problems in conversions, but we will only focus on the index processing.

Changing the Links in the Index

First, the index had to be available in indexing software to allow the addition of the pinpoint[4] IDs. Since the print book indexer was not the same indexer adding the IDs, the index from the PDF file was converted so it could be imported into CINDEX software to accomplish this step[5].

Next, the indexer adding the IDs needs to know what they are! Since conversions to EPUB do not provide pinpoint IDs of any nature, a new tool is needed. Leverage Technologies has created such a tool, let's call it EAI[6], to do this. It also was used for a couple of other processing steps discussed later on. This first step read the EPUB and created an intermediate or indexer version of the EPUB. The page break anchors were made visible when reading the EPUB to assist the indexer. For every paragraph, heading, list item, etc., EAI added a pinpoint ID and made it visible. In other words, a unique identifier was added to every chunk of text, whether a chapter title, subchapter title, box, paragraph, figure, or table.

This intermediate EPUB version was delivered to the indexer so that it could be used to walk through the book while also walking through the index in page order. The indexer identified the point on the page, say a paragraph, that was a more specific point than the page break to link to and added that ID to the page number. Let's say that the term iPads was mentioned in chapter 2 on page 37 and that page had 5 paragraphs with pinpoint IDs of 2.23, 2.24, 2.25, 2.26, and 2.27. The indexer determined that iPads were specifically discussed in the paragraph with ID 2.26. So the ID 2.26 was added to the locator field using the vertical bar as a separator to page number 37. This produced the new locator 37|2.26. [A different character other than vertical bar could have been used.] This process was repeated for each page number. When the index was completed, the indexer ensured that every entry (that was not a cross- reference entry), contained a locator of the form #|#. At this point the index was returned to Leverage Technologies.

The index was exported from CINDEX into the tagged format ready for HTML/Prep[7] to process. Before running HTML/Prep though, the exported file was run through step two of the EAI tool. This procedure transformed the locators, such as 37|2.26, into a form that contained the page number and a URL that would work within the EPUB. In our example, ID 2.26 was mapped to the EPUB filename for that chapter and the pinpoint ID in a format too arcane to show for our purposes here.

Now HTML/Prep was run on the transformed file with the mapped IDs, which created the new XHTML index file for the EPUB. HTML/Prep also linked the cross-references. (These were not linked during the conversion of the PDF file.) Of course any cross-references that didn't link were reported (for example, if they were not an exact match or if they were generic), and decisions were made about changing the index to allow them to link.

One last time the EAI tool was run (step 3) which accomplished several things:

- Removal of the visible page break IDs and the visible pinpoint IDs;
- Replacement of the page-break-linked index file with pinpoint-linked index; and
- Addition of index style definitions to the CSS style sheet to match those HTML/Prep generated.

Finally we had the final EPUB, and it was time to review the result and test it. This was done using Calibre since viewing this on the computer was quicker and easier than sending it to a physical eReader device. Not all eReaders or reading software act the same! Bet you are not surprised to read that. Similar to how a web page may display differently in different browsers that work on HTML or XHTML, so too there are display discrepancies that can occur on different reading systems. For instance, the resulting index looked fine using Calibre but didn't format properly in another reading software Cool Reader. Another difference occurred when the reviewer jumped to the target by clicking on a link. Calibre showed the target paragraph at the top of a window that was sized similar to that of an eBook screen. When maximized to the full size of the computer screen, after clicking the link, the target paragraph appeared in frame but was no longer the top paragraph.

What Went Awry

While many issues arose during the initial run through, i.e., errors reported by the software, and were resolved as we went along, other issues only became apparent at the review stage. Once these were addressed, the processing cycle steps were repeated until we had the final "good" EPUB.

- Roman numerals: The original book indexer had worked from first-pass pages in a PDF file, which had numbered front matter pages with small roman numerals. After turning the index in, the publisher added two pages to the front matter and, of course, renumbered the front matter pages, but neglected to adjust the corresponding locators in the index which ended up being off by two (or as the Romans would have said *ii*). This was discovered during the page number mapping process. These page numbers had to be adjusted.
- PDF index conversion: While the text data was converted with full fidelity, style settings were lost so italics needed to be restored for headings that were book titles and such.
- eBook indexer changes: The indexer adding the pinpoint IDs (not the original book indexer) came across situations where the concept embodied by the entry could not be found on the referenced page. There seemed to be nothing else to do but to delete these entries. Maybe these were caused by a typo in the page number, but without fully re-indexing the book from scratch they could not be easily found. There were also passing mentions indexed on one page where the meat of the discussion was on the next page. Pinpoint IDs were used that reflected the entrée not the appetizer.

At last, we had the EPUB ready to deliver back to the publisher who would provide it for download when sales were made. Throughout, this discussion has focused on an eBook in EPUB format that can be read by the Nook and other devices. MOBI is the format used by Kindles. At this point converting the EPUB to MOBI would be the best way to create an analog MOBI file. EPUB is an open standard and well defined. MOBI is not and currently the EAI tool doesn't work on MOBI files.

Summary

A lot was learned during this first EPUB index relinking project. Most important was that this is not a push the button workflow scenario yet. A lot of review had to take place at each step during the process and fixing errors that were reported along the way. The second indexer was faced with the philosophical dilemma: either maintain the fidelity of the book index in the face of errors and typos, or improve the index for the eBook. And even then, the question arises whether making such changes constitutes a copyright violation? Fortunately, the publisher owned the copyright in our case.

The book in question had very simple page numbers without additional notations like '*t*' and '*f*' for tables and figures. So how this process would work when a box, figure, table, or footnote is indicated needs to be analyzed for future projects. Similarly an index citing section numbers or legal citations may also present issues not encountered in this project.

Nonetheless, the issues I have raised here on relinking the index and what to watch out for should provide some guidance to both programmers and indexers for the future.

Notes

1. This is basically a zip file with the text of the book represented in XHTML. There are other files contained in the zip file but not pertinent to the discussion in this article. See *http://idpf.org*, and my article in this issue.
2. One of the many eReader software available for PCs or Macs. *http://calibre-ebook.com*
3. There are books that will remain in their print format because the layout is critical or the design is part of the allure: art books, comic books/manga, cookbooks, etc.

4 The term pinpoint will generally be used to mean any specific target. It most often will be a paragraph but could also be a list item, table row or cell, footnote, figure/illustration, etc.

5 Conversion services are available from Indexing Research, Leverage Technologies, and there are some third-party tools that may work as well.

6 The tool doesn't yet have an official name. For now it is still in development and undergoing alpha testing.

7 *http://www.levtechinc.com/publishing-indexing-products/utilities/html-prep.asp* can create HTML and XHTML forms of indexes from CINDEX or Sky Index.