IDPF, EPUB and ebook indexes

Glenda Browne

In the rapid move to ebook publishing, indexes were often ignored, or presented poorly. Indexing societies are working to remedy this. One key development has been the International Digital Publishing Forum EPUB Indexes Working Group, which was developed following a proposal by the ASI.[1]

Ebook publishing has taken off in the last year or so, but indexes are often not included, or are not effectively presented, in ebooks. As Jan Wright reported in her editorial in the January issue of *The Indexer* (2012), following a proposal from the ASI Digital Trends Task Force, the International Digital Publishing Forum (IDPF), which develops and maintains the EPUB standard, has started an Indexes Working Group. IDPF has members from around the world, and includes the 'heavyweights' Google, Apple and Sony, as well as traditional publishers, ebook conversion houses and other interested organizations.[2]

The Charter for the Indexes Working Group is at http://idpf.org/charters/2012/indexes/. ASI members David Ream and Michele Combs are leading the group, and Mary Russell and Glenda Browne are representing ANZSI Inc. Minutes and working documents for the Indexes Working Group are all publicly available at https://code.google.com/p/epub-revision/wiki/IndexesMainPage.

EPUB

EPUB3 is a free and open ebook standard designed for reflowable content (that is, content without fixed page breaks). Reflowable content means the text display can be optimised for the device used or according to the user's preferences. EPUB is an international standard, and supports scripts and reading directions other than those required for English and other Western languages. It also has a strong focus on accessibility.

EPUB publications are basically zipped collections of resources that can be interpreted by reading systems and rendered (presented, visually or aurally or both) for users. Some of the resources in the collection provide metadata and navigation information while others contain the actual content (http://idpf.org/epub/30/spec/epub30-overview.html).

EPUB3 supports audio, video and interactivity. It will be possible to link from an index to a specific time within an audio or video file.

The EPUB standard uses existing open standards wherever possible. These include XHTML (eXtensible HyperText Markup Language), CSS (Cascading Style Sheets), SVG (Scalable Vector Graphics), SSML/PLS/CSS 3 Speech for text-to-speech rendering (Speech Synthesis Markup Language/Pronunciation Lexicon Specification) and SMIL for synchronizing text and audio playback (Synchronized Multimedia Integration Language) (Garrish, 2011).[3]

A new feature of EPUB3 is CFIs (canonical fragment identifiers) which allow links to every part of a document without the target markers having to be added individually. These will be important for indexes in the long term, although there are as yet no software products that provide the functionality necessary for their use in indexing.

The minimal bibliographic metadata requirement for EPUB publications is three elements from the Dublin Core Metadata Element Set (DCMES) – title, identifier and language – along with the modified property (date on which the resource was changed). Additional optional metadata is expressed using the DCMES optional elements and the meta element (http://idpf.org/epub/30/spec/epub30-publications.html).

For a short introduction to EPUB, see the article by Bill Kasdorf (also a member of the Indexes Working Group) in *Information Standards Quarterly* (2011).

IDPF EPUB Indexes Working Group

The IDPF EPUB Indexes Working Group meets by telephone conference every fortnight, and discusses issues via a mailing list. There is also a wiki in which documents are developed.

There is currently a wiki page for the collection of definitions that will be needed when the specifications are written, and another which is gathering a list of atomic elements (https://code.google.com/p/epub-revision/wiki/IndexesAtomicElements). Atomic elements are the essential components of an index which cannot be broken down further into smaller elements.

There has been some discussion in the group about the types of cross-references needed, of complex locators, and about making a 'container' to capture a whole entry, but no firm decisions have been made yet.

Other discussions have focused on the links from the index to the text. There have been two parts to this discussion – one is the question of whether you link to print-page equivalents, book sections (such paragraphs) or individual words, and the other is about the nature of the targets within the document.

For legacy content (pre-existing content also published in print format) the simplest option is to insert page break markers in the ebook to show where the print pages break, and to link the existing print index entries to the top of these breaks. This gives almost the same user experience as a print index does (a bit better because it links you directly to the

'page', but a bit worse because with a small screen you might have to scroll to find the content that fitted on one print page).

Digital conversion companies convert print books to ebooks on behalf of publishers. They often charge a per page rate on a sliding scale depending on the complexity of the work (see for instance http://ebookarchitects.com/conversions/services.php).

For new content, publishers have two options. They can use embedded indexing, in which the index entries are inserted into the text to which they refer, or they can create a standalone index that is linked to targets within the text. These targets can be (X)HTML anchor tags or CFIs. CFIs are new in EPUB3. They are automatically generated pointers to every part of the text that describe locations through their relationship to the start of the book (in non-technical terms, something like 'CFIstartshere, Chapter 1, Section 3, Paragraph 5, 87th character in'[4]). The advantage of CFIs is that they automatically define every location in the document. The disadvantages are that there are currently no mechanisms for easily inserting them into an index, and they are machine-readable rather than easily human-readable.

EPUB focuses on semantic description of content rather than presentation. That is, it will identify key features of indexes (such as main headings, subheadings, cross references and locators) and work out ways of describing them and their relationships. It cannot, however, tell reading devices how these should be presented. So if an EPUB ebook says that something is a main heading, the reading device can choose to display it using bold font; it can present the index in one or two columns – this is all up to the reading device (and the publisher's style sheet), and not up to EPUB.

One question the Working Group has considered is how ranges will be dealt with. This is easy with print pages, but needs to be considered with ebooks, unless all locators are to take users simply to the start of the discussion of interest to them, without giving them any idea where the discussion will stop. There is some technical complexity in the tagging of ranges; whether and how these are displayed to the reader will depend on the reading device, but it could be by highlighting the text within the range with a coloured background.

The existence of standards is crucial for efficiency in the ebook business. The inclusion of indexes in the EPUB3 standard to cater for the incorporation of linked indexes to text locations or to print page equivalents, using links or CFIs, will be of great benefit to publishers, indexers and readers of nonfiction books throughout the world.

Notes

1 An earlier version of this article was published as 'Indexes Working Group of IDPF', *ANZSI Newsletter* **8**(4) (May 2012), 6–7, www.anzsi.org/UserFiles/file/May%202012.pdf. A longer article on this topic is Browne (2012).

2 In one presentation the relationship of the traditional publishers to the 'heavyweights' was described as being like 'the chickens watching the foxes'.

3 This book is 24 pages long and can be downloaded free (you still have to go through the checkout process as if you were purchasing the book).

4 A real example is 'epubcfi(/6/4[chap01ref]!/4[body01]/10[para05]/3:10)'.

References

Browne, G. (2012) 'Ebook indexes, EPUB and the International Digital Publishing Forum.' *Online Currents* **26**: 127–30 (June). Available at: http://webindexing.biz/ebook-indexes-epub-and-the-international-digital-publishing-forum/

Garrish, M. (2011). *What is* EPUB*3? An Introduction to the* EPUB *specification for multimedia publishing*. O'Reilly Media. Available at: http://shop.oreilly.com/product/0636920022442.do

Kasdorf, B. (2011) 'EPUB3 (not your father's EPUB): opening Pandora's box in the world of e-books' *Information Standards Quarterly* **23**(1): 2. Available at: www.niso.org/publications/isq/2011/v23no2/kasdorf

Wright, J. (2012) 'Editorial.' *The Indexer* **30**(1): 1.

Glenda Browne is a freelance indexer, co-author of The indexing companion *and ANZSI Inc representative on the IDPF EPUB Indexes Working Group. Email:* glendabrowne@gmail.com

A full life . . .

see also admirers, literary; admirers, political; artistic connections; brotherhoods; characteristics; clients; collections; craft skills; designs; disciples; education; employees; family relations; finances; friends, female; friends, male; health; homes; influences on; literary connections; Morris, Marshall, Faulkner & Co.; Morris & Co.; Morris, William, biographies and studies; Morris, William, literary depictions of; Morris, William, portraits; political associates; political connections; products; servants; sexual orientation; shops/showrooms; sporting activities; travels; views, writings

From the index entry on the main character in Fiona MacCarthy, *William Morris*, London: Faber & Faber, 1994.

LIFE
DEBTS
HEALTH
PERSON
POLITICS
RELATIONSHIPS
WORKS

Subentries to the entry for the eponymous character in Amanda Foreman, *Georgiana, duchess of Devonshire*, London: HarperCollins, 1998.