INDEXES IN TOMORROW'S WORLD

Glenda Browne

Freelance Indexer

www.webindexing.biz

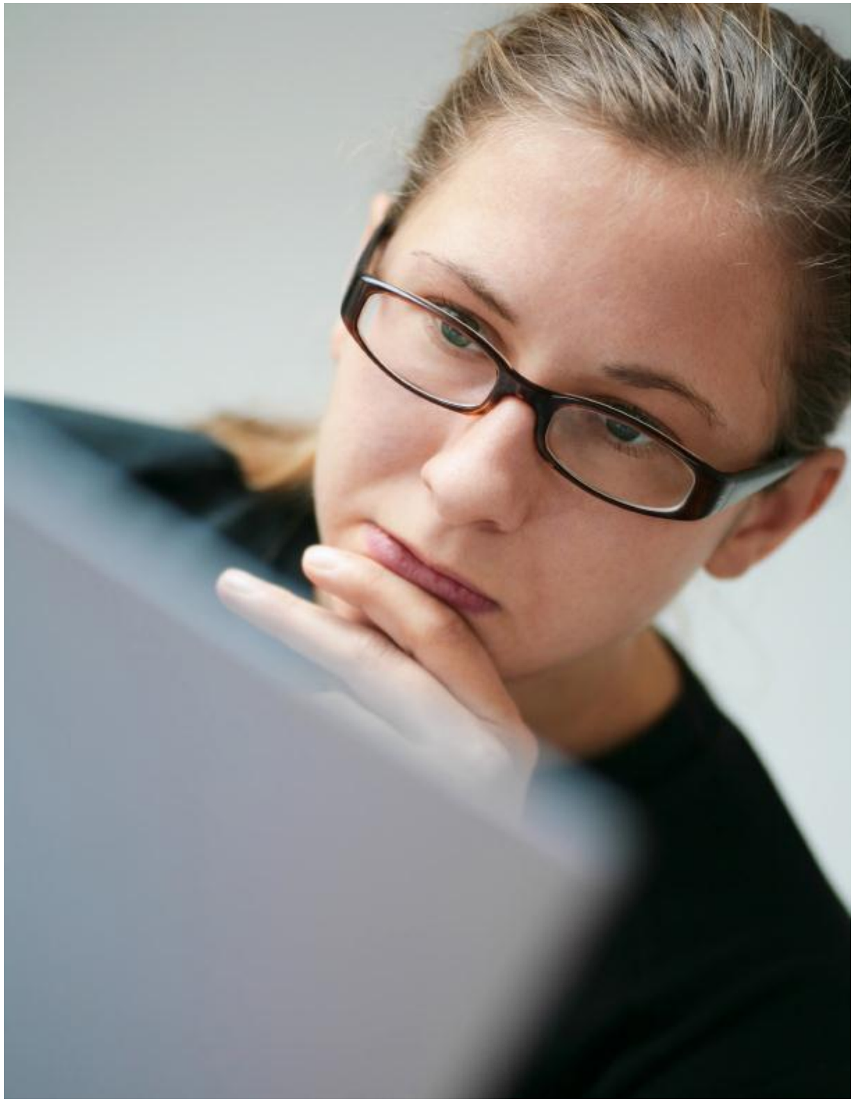
Enduring indexes

- *ASTC 2003: Engage the online reader*
  Why websites need indexes/indexers (and metadata), and the great future ahead.
- *ASTC 2012: Emerging and evolving*
  Why ebooks need indexes/indexers (and metadata), and the great future ahead.

Overview

- Why index when you can search?
- Ebook indexes to date
- Ebook index developments
- Adapting indexing for ebooks
- Software
- Metadata

Why index when you can search?



Indexes…

- …are explorable
- …concentrate
- …aggregate
- …disambiguate
- …are selective
- …provide categories
- …contribute to the tone of a book
- …can be fun.

Indexes are explorable – they provide easy paths to relevant content

Old water canal, Tenerife, Canary Islands, Spain

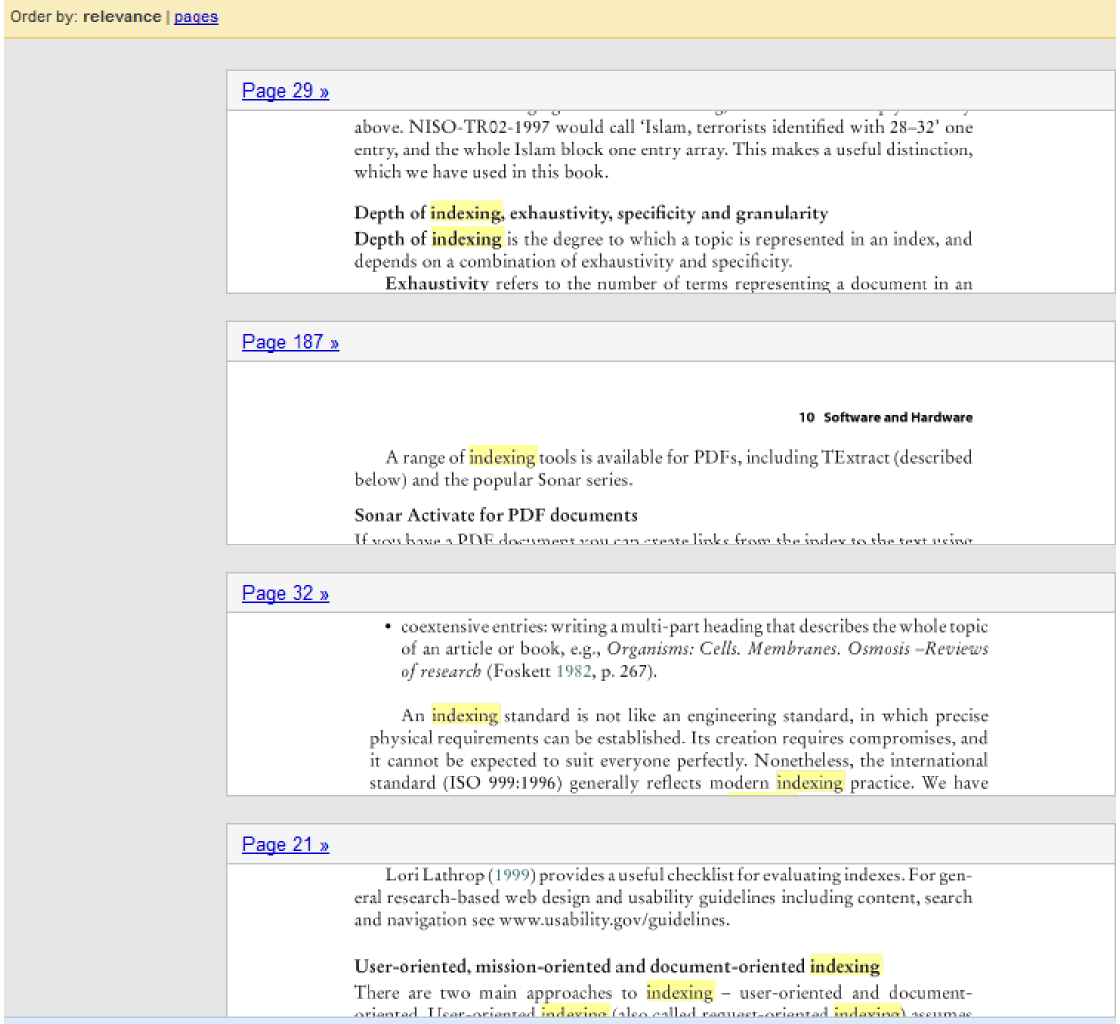
Indexes are explorable

- You can:
  - Use hierarchical relationships to zoom in and out (looking at broader and narrower terms)
  - Use subheadings to explore aspects of a topic
  - Use related term relationships to get ideas of other places to look
  - Use locator information to get an idea of the relative importance and abundance of discussions by noting section length, special formatting (eg, bold for key discussions) and the number of locators. (Locators refers to page numbers or their equivalents, including hyperlinked text.)

Indexes concentrate…

indexing *see also* user-oriented indexing
  depth of 29
  software for 187
  standards for 32
user-oriented indexing 21

…so you don’t get long lists to scan



Order by: **relevance** | pages

Page 29 »

above. NISO-TR02-1997 would call ‘Islam, terrorists identified with 28–32’ one entry, and the whole Islam block one entry array. This makes a useful distinction, which we have used in this book.

Depth of indexing, exhaustivity, specificity and granularity

Depth of indexing is the degree to which a topic is represented in an index, and depends on a combination of exhaustivity and specificity.

Exhaustivity refers to the number of terms representing a document in an

Page 187 »

10 **Software and Hardware**

A range of indexing tools is available for PDFs, including TExtract (described below) and the popular Sonar series.

Sonar Activate for PDF documents

If you have a PDF document you can create links from the index to the text using

Page 32 »

- coextensive entries: writing a multi-part heading that describes the whole topic of an article or book, e.g., *Organisms: Cells. Membranes. Osmosis –Reviews of research* (Foskett 1982, p. 267).

An indexing standard is not like an engineering standard, in which precise physical requirements can be established. Its creation requires compromises, and it cannot be expected to suit everyone perfectly. Nonetheless, the international standard (ISO 999:1996) generally reflects modern indexing practice. We have

Page 21 »

Lori Lathrop (1999) provides a useful checklist for evaluating indexes. For general research-based web design and usability guidelines including content, search and navigation see www.usability.gov/guidelines.

User-oriented, mission-oriented and document-oriented indexing

There are two main approaches to indexing – user-oriented and document-oriented. User-oriented indexing (also called request-oriented indexing) assumes

Indexes concentrate 2

- Indexes often provide a level of specificity intermediate between the table of contents and search results.
- ‘…an index is a better guide than a table of contents to the sweep and depth of knowledge available in the text, revealing far more opportunities for discovery and learning.’ (Geoffrey Marnell)

Indexes aggregate

- They group the *Matterhorn* with *Monte Cervino* and group *Bonnie Prince Charlie* with the *Young Pretender* and *Charles Edward Stuart.* (SI PTG)
- These are easily identified groups, but even so, Kindle X-Ray (only in Kindle Fire) doesn't always manage this yet, failing to group *Buffalo Bill* with *William Cody,* while grouping, incorrectly, *Daniel Burnham* senior and junior. This is despite the fact that this information is in Shelfari, which they say they use as a resource. (Wright 2012)
- Indexes also provide more subtle groupings, eg, of *farm waste pollution* in Chapter 2 grouped with *agricultural runoff to watercourses* in Chapter 12. (SI PTG)

Indexes disambiguate 1

ASTC (Australian Society for Technical Communication)

ASTC (Alice Springs Town Council)

ASTC-design

Indexes disambiguate 2

- ‘Jack’ (‘King Burrigan’) 39
- ‘Jack’ (native servant) 115
- ‘Jackey’ (c.1780-1840) 7
- ‘Jackey’ (c.1800-1834)
  - arrested and charged 44–6
  - as ‘domesticated Aborigine’ 41–2
  - transportation of 36–7, 56
  - trial of 48–51
  - wounded by leg irons 251
- ‘Jackey’ (c.1825-1840) 7
- ‘Jackey’ (sentenced 1865) 148
- ‘Jackey’ (‘Tallboy’) 129, 133–6
- ‘Jackey Jackey’ 85–7, 89, 93–4, 98–9, 141
- ‘Jackson’ 14, 255
- Jackson, Constable 199
- Jackson, William 136–7
- ‘Jacky Jacky’ (Koenghegulluc) 104–7, 123
- ‘Jacob’ 164
- ‘Jaggy Jaggy,’ ***see*** Lo.gir.ma.koon

Indexes are selective

- A cookbook index entry for *eggs* may show you the *custards*, *soufflées* and *meringues*, but not every recipe that has an egg in it.
- An index won’t lead you to ‘as we saw in the chapter on Leonardo’ when you look up ‘Leonardo’, but a search engine will. (SI PTG)

Indexes provide categories

- chilled desserts
- gluten-free cakes
- Sicilian main meals

- controversies about indexing
  - filing order
  - function words in subheadings
  - indexing ‘indexing’
  - passing mentions
- controversies, indexing of (Browne and Jermey)

Indexes contribute to the tone of a book

- Academics expect indexes in quality books

  ‘An index and a bibliography have been integral components of scholarly monographs,…This is where they tend to start reading a new book – from the back...’ (Agata Mrva-Montoya)

Indexes can be fun

- When Sarah Palin published her memoir without an index, Darby created one for her:

Iraq war

Palin's informed perspective on 214

Media's skewing of Palin's informed perspective on 238

Lies told about Sarah Palin 74-75, 77, 79, 95, 102, 148, 202-204, 215, 232, 236-239, 246-247, 272-275, 289, 314, 318-320, 338, 343, 346-348, 350-352, 365-366, 378, 380

Lies told by Sarah Palin N/A

BNA Usability Study

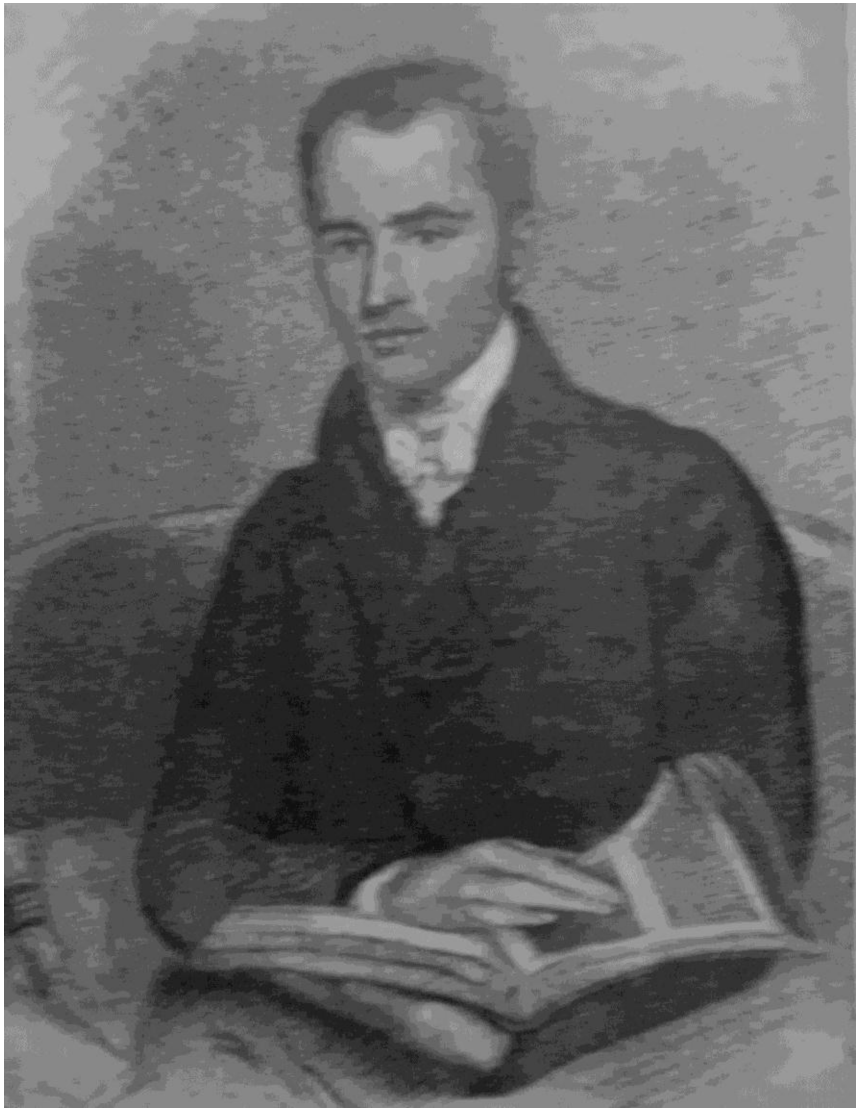
- ‘In the BNA Usability Study [into use of legal books], index users had an 86 percent success rate while text searchers had only a 23 percent success rate. The study included both single answer and more complex research tasks.’
- They also found that use of indexes saved time.

Too big to browse, too small to search

- Search algorithms that work for web-wide search don’t usually work for smaller collections, including intranets and ebooks.
- Ebooks don’t have the content nor the number of users to provide adequate feedback for relevance ranking. (Mark Baker)
- A compromise solution was suggested: Structuring the materials into online help portal pages, streamlining the index to refer to those pages, and relying on search for simple concepts can make a feasible help system.’ (Jan Wright)

Ebook indexes to date

Book indexes of the past worked well for print books



Thomas Bulman Browne, 17??

But unfortunately most early ebook indexes provided just the bare bones

Indexes were…

- Omitted
- Were provided but with…
  - No page numbers or unlinked page numbers
  - Linked badly-displayed page numbers
- Just a few were great indexes
  - They were embedded or linked, to provide one-click access to the relevant text
  - They were created with care, and checked to see that they worked
- Unfortunately the existence of so many bad indexes leads people to say that ebook indexes are worthless.

Provided with no page numbers or unlinked page numbers

- An informal study of Kindle ebooks found that of 21 titles that had indexes in print form, only two had fully functional, linked indexes in the ebook.
- Also, when Amazon says that a book has an index, there is almost no way of finding out in advance whether it is an active, linked index or just a copy of the print-format index. (Pierke Bosschieter)
- Mike Cane reported on unlinked indexes that had notes suggesting ‘entries in this index, and other terms, may be easily located by using the search feature of your e-book reader.’

Provided with linked page numbers, but badly-displayed

- Problems with Amazon indents

Ebook index developments

ASI Digital Trends Task Force

- American Society for Indexing Digital Trends Task Force was set up to address the lack of quality indexes in ebooks:
  - It has a LinkedIn group
  - It wrote the charter for, and chairs, the IDPF EPUB Indexes Working Group
  - ASI DTTF members have had discussions with Adobe about InDesign indexing

International Digital Publishing Forum

- IDPF is the trade and standards association for the digital publishing industry
- Develops and maintains the EPUB ebook standard
- International and open
- Uses existing standards where possible, e.g.
  - XHTML5
  - CSS3 (Cascading Style Sheets)
  - SVG (Scalable Vector Graphics)
  - SSML/PLS/CSS 3 Speech (for text-to-speech rendering)
  - SMIL (for synchronising text and audio playback)
  - DAISY (talking book standard)
  - Dublin Core

EPUB

- EPUB is a free and open e-book standard (originally designed for reflowable content, although it now has a fixed format option as well).
- Derived from Open Ebooks standard
- Ebooks in EPUB format are zipped files containing content documents and navigation documents. They are similar to collections of webpages, but have a set reading order – http://www.asindexing.org/files/DTTF/Anatomy_of_ebook.pdf
- Presentation is determined by publishers' style sheet and Reading System choices (and potentially user preferences).
- ANZSI is a member of the IDPF and the EPUB IWG.

EPUB IWG use cases and implementation suggestions

- Chapter-like index – the electronic equivalent of a standard book index.
- Index term search (pop-up index) – access to the index from any place in the text, by selecting a word or phrase in the text, or by typing an entry into a search box (possibly with predictive search).
- Index locator search (reverse index) – retrieval of all index entries that are attached to a highlighted range of text
- Standalone indexes – ebooks that contain nothing but indexes to other ebooks (eg, journal collections; books in series).

EPUB3 CFIs

- EPUB3 CFIs (Canonical Fragment Identifiers) are automatically generated pointers to every part of the text. They describe every location in a book through its relationship to the start of the book.
  - In non-technical terms, something like ‘CFIstartshere, Chapter 1, Section 3, Paragraph 5, 87th character in’
  - A real example is ‘epubcfi(/6/4[chap01ref]!/4[body01]/10[para05]/3:10)’.
- CFIs are potential anchors for links from indexes but they are not easily human readable, and there is as yet no software to easily insert them into indexes.

Future of page numbers

- Page numbers lose meaning in reflowable ebooks
- They retain value for legacy books for citation and comparison with printed versions
- There will be a transition period, after which it is likely that paragraph or section numbers will take over when numbers are needed.
- Indexing decisions about page numbers will largely depend on book-wide decisions

Accessibility of ebooks

- EPUB has a commitment to accessibility, and has incorporated DAISY requirements into EPUB3.
- Amazon’s KF8 format does not support DAISY, although the US Department of Education expects everyone to. (McIlroy)
- Acessibility increases with:
  - Semantic markup – saying what things are, not how they should look. For example, an index will say ‘subentry’ rather than ‘indent 5 spaces’.
  - Dynamic lookups (e.g. auto-completing search boxes) that don’t involve manual navigation
  - Text-to-speech synthesis
  - More structure, e.g. sectioning the index by letter of the alphabet

EPUB sample HTML5 markup

- Sample high-level HTML5 markup can be seen on the web.
- A final version is expected next year.
- The inclusion of indexing in the EPUB standard will make it easier for publishers to include indexes and for reading systems to display them effectively.

Non-EPUB index developments

Ebook index developments – mashups

- Index mashup of 5 books by Seth Godin
  - http://indexmasher.com/online/Linchpin-AllMarketers-PurpleCow-Tribes-Dip.html
- Eat your books
  - A web-based recipe search engine (with lessons for ebooks) that indexes a large collection of cookbooks – you can limit the search to the cookbooks you own, or get ideas of other cookbooks to buy – http://www.eatyourbooks.com.

Ebook indexes as a marketing tool

- Include indexes in ebook samples (Joe Wikert)
- Use indexes as a cross-selling tool by including live links to content in different ebooks [with a way for users to filter them out if they wish].’ (Kevin Broccoli)

Ebook index developments – X-ray

- Kindle X-Ray in Kindle Fire (not available in Australia)
  - Provides a sparkline graph showing occurrences of names (and other topics) throughout the book
  - Needs curation or checking to ensure appropriate connections are made

Ebook index developments – visual index

- *The Fry chronicles: an autobiography* (iPhone app edition)
- http://www.itsbeenreal.co.uk/files/gimgs/46_all5-iphone-flatindexhibit.jpg
- Key theme tags in the book are allocated to 4 major groups: People, Subjects, Emotions, and ‘Fryisms’.
- The whole book is represented by a circular wheel of ‘spines’, each of which represents a section of text. The arcs around the outside of the wheel connect sections that are tagged with the same theme.
- The book is written in a way that allows easy access from many entry points.

Adapting indexing for ebooks

Indexing for ebooks – small screens

- Indexing for a small screen (or a variable-sized screen) – indexer decisions
  - Shorter headings and subheadings to avoid turnover lines
  - Shorter subheading lists so main headings don’t scroll off the page. This is especially important when coming to the middle of an index from a search
- Indexing for a small screen (or a variable-sized screen) – system decisions
  - Keeping the main heading accessible from all subheadings, eg, displaying it when you hover over a term
  - Providing more control over the amount of content that is displayed, eg, allowing users to display only main headings.

Indexing for ebooks – pages and points

- Pages, paragraphs or points
  - Index to exact locations or to sections
    - Sections give context, and are easier to update and translate
    - Exact locations might work better for specific terms such as names
  - Index starting points or ranges
    - Ranges could be highlighted
    - Ranges are essential for Locator search

- Locators
  - Text of entry, section number, page number, arbitrary marker
  - Unique locators – should we limit to one per concept?

- Cross references
  - More double entry and less *see* references

Indexing for ebooks – alternative search terms in EPUBs

- When searching the index from the text, the exact characters matter (unless fuzzy searching is used).
  - Include ‘fetus’ as an alternative metadata term (hidden) at the entry ‘foetus’
  - Include ‘Samuel Langhorne Clemens’ and ‘Mark Twain’ as alternative metadata terms (possibly displayed) at the entry ‘Twain, Mark’.

Indexing for ebooks – metadata about terms and locators

- EPUB will define coding for metadata about terms (e.g. saying that they are flowers, or mountains, or names of authors). This could be used to:
  - expand generic cross references, eg, ‘prime ministers, see also names of specific prime ministers’)
  - filter the index, eg, ‘show only names of authors’.
- EPUB will also define coding for provision of metadata about locators, eg, to say that the locator links to a table or a figure.

Indexing for ebooks – workflow and planning considerations

- Indexing as you write (if you are an author and indexer)
- Optimising indexing for different outputs (single-sourcing)
- Re-using indexes effectively (eg, for new editions and translations)

Software

Linking indexes

- Embedded indexes
  - Embedding in Adobe InDesign, MS-Word, LibreOffice, etc. Not all convert readily to EPUB format
  - Embedding in XML
- Standalone linked indexes
  - Creation of a standalone index linked to anchors in the text
  - Creation of active links in PDF documents using Sonar Activate

Adobe InDesign and FrameMaker

- InDesign
  - InDesign CS6 strips the index entries out of EPUB files when exported. There are workarounds (Wright, Castro)
  - ASI DTTF members have talked with Adobe about the indexing feature of InDesign
  - There is an InDesign Indexing YahooGroups mailing list
- FrameMaker
  - The latest version of Adobe Technical Communication can convert a document created in FrameMaker to EPUB via RoboHelp. Apparently the index ‘doesn’t look very good’ in the EPUB output. (Cheryl Landes)

Metadata

Metadata

- Metadata is crucial for findability of ebooks
- EPUBs must include title, identifier and language along with the modified property using the Dublin Core standard.
- ONIX3 bibliographic metadata can be included in EPUB3 files. ONIX for Books is an XML-based standard for book metadata (http://www.editeur.org/83/Overview/).
- ONIX is used with BIC subject codes (http://publishers.asn.au/index.cfm?doc_id=271), which are from the UK. There are plans to merge BIC with BISAC (the North American book trade subject headings)

The Future

- The IDPF EPUB Working Group shows that the industry is committed to quality indexing.
- The inclusion of indexes in the EPUB standard will make it easier for publishers to provide linked indexes, and when more ebooks have effective indexes, the expectation that they can and should be included will be greater.

References – 1

ASI DTTF LinkedIn group, http://www.linkedin.com/groups/ASI-Digital-Trends-Task-Force-4005509?gid=4005509&trk=hb_side_g

BNA Law School Education Series. ‘Using online indexes’, http://dl.dropbox.com/u/2248375/Using%20BNA%20Indexes%20study.pdf

Bosschieter, Pierke. ‘The Kindle and the indexer’ *The Indexer* 28 (3):116-118

Broccoli, Kevin. Interview with Joe Wikert, http://toc.oreilly.com/2012/03/ebook-indexes-mashups-toc-podcast.html

Browne, Glenda and Jermey, Jonathan. *The indexing companion*. Cambridge University Press, 2007.

Cane, Mike, 2009. ‘ePub: the death of the index?’, http://ebooktest.blogspot.com.au/2009/08/epub-death-of-index.html

Castro, Liz, 2010. ‘Creating an index for EPUB with InDesign and GREP’, http://www.pigsgourdsandwikis.com/2010/07/creating-index-for-epub-with-indesign.html

Darby, Seyward. ‘The “Going Rogue” Index: Sarah Palin wouldn't make one, so I did it for her’, http://www.tnr.com/article/politics/the-going-rogue-index#

EPUB Indexes Working Group, ‘high-level HTML5 markup’ [sample]

EPUB Indexes Working Group. ‘Main wiki page for EPUB Indexes Working Group’, https://code.google.com/p/epub-revision/wiki/IndexesMainPage?ts=1322858948&updated=IndexesMainPage

References – 2

Garrish, Matt, 2011. *What Is EPUB 3?: An Introduction to the EPUB Specification for Multimedia Publishing*, O'Reilly Media. This book is 24 pages long and can be downloaded free from http://shop.oreilly.com/product/0636920022442.do

The Indexer March 2012, single issue focussing on ebook indexing available at www.lulu.com

Kasdorf, Bill, 2011. ‘EPUB 3 (not your father’s EPUB): opening Pandora’s box in the world of e-books’ *Information Standards Quarterly*, v.23 i.2, http://www.niso.org/publications/isq/2011/v23no2/kasdorf

Landes, Cheryl, 2012. Feedback from an Adobe Webinar, Posted by to the LinkedIn ASI DTTF group, 13 Sept 2012

Marnell, Geoffrey. 2012. ‘A lament for the vanishing index’ *ASTC(NSW) Newsletter* July 2012

McIlroy, Thad. 2012. ‘Ebook formats are a mess - here’s why’ *Learned Publishing* 25(4): 247-250

Mrva-Montoya, Agata, ‘Indexing monographs in the digital age’, http://blogs.usyd.edu.au/sydneypublishing/2012/08/indexing_monographs_in_the_dig_1.html

References – 3

Mrva-Montoya, Agata, ‘Indexing monographs in the digital age’, http://blogs.usyd.edu.au/sydneypublishing/2012/08/indexing_monographs_in_the_dig_1.html

Posavec, Stefanie. ‘MyFry iPhone app’, http://www.itsbeenreal.co.uk/index.php?/new/myfry-iphone-app

Publishing Technology Group, Society of Indexers (UK). ‘Characteristics of analytical indexes’, http://dl.dropbox.com/u/2248375/SI_Analytical_Index.pdf

Wikert, Joe. ‘Rethinking samples’, http://jwikert.typepad.com/the_average_joe/2012/03/rethinking-samples.html

Wright, Jan and Ream, David (ASI DTTF). IDPF Index functionality in ePub, 12 October 20122, http://dl.dropbox.com/u/2248375/IDPF%20Index%20functionality%20in%20ePub.pdf

Wright, Jan. ‘InDesign ePub Scripts’, http://www.wrightinformation.com/Indesign%20scripts/Indesignscripts.html

Wright, Jan. ‘The devil is in the details: indexes versus Amazon’s X-Ray’,*The Indexer* 30(1):11–16