not what it used to be.'
» Paul Valery (1871-1945)
'Time, occasion, chance and change
To these all things are subject.'
» Percy Bysshe Shelley
'Never let the future disturb you
u will meet
if yo
ve to, with the same we
s of reaso
whi
day arm you against the
nt.'
» Meditations. Marcus Aureliu

The Matrix Reloaded

EPUB ebook indexing

Presented by Glenda Browne. Based on the Matrix presentations by Jan Wright and Glenda Browne and the EPUB presentation by David K. Ream at the ANZSI conference in Wellington, NZ, 2013.

May 2015

The Matrix – indexing options

Trinity: We can never see past the choices we don't understand.

History of The Matrix

The Matrix documents tabulate the choices that indexers and publishers have to make about ebook indexing tools and approaches.

http://www.asindexing.org/about-indexing/digital-trends-task-force/

Created and maintained by Michele Combs, David Ream, Jan Wright, Pilar Wyman, and Glenda Browne. Updated November 2013

Overview

- Ebook indexes, IDPF, EPUB and XHTML
- EPUB 3 Indexes Specification
- Software tools for ebook indexing
- Approaches to indexing ebooks
- Publishers and ebook indexing

Ebook indexes, IDPF, EPUB and XHTML

Ebooks

- For this talk, I am using ebooks to mean electronic books with reflowable text, that don’t necessarily have page numbers, and are read using specialised software on mobile devices or computers.
- I have not included PDF books or o-books (books on the web).

Ebook indexes

- For this talk, I am using ebook indexes to refer to active, linked indexes where users can click on entries or locators and be taken to the place the content is discussed in text.
- Locators is the word we use to refer to page numbers and their equivalents.

Ebook reading

- Ebooks can be read on:
  - specialised ereaders, eg, Kobos and Nooks
  - software installed on a PC
  - smart phone apps.
- Mobi (Kindle) and KF8 (Kindle Fire) are proprietary formats.
- EPUB 2 and EPUB 3 are open source, community-developed standards.
- This presentation mainly discusses EPUB 3.

EPUB 3

- EPUB 3 is an open source ebook standard.
- It is developed and maintained by the International Digital Publishing Forum (IDPF).
- EPUB 3 uses existing standards where possible (eg, XHTML, CSS3, SVG, DAISY, Dublin Core).
- It is international, and has a strong focus on accessibility.

EPUB 3

- EPUB 3 uses XHTML. An EPUB book has a lot in common with a website. Some of the differences relate to the packaging of the ebook as a unit.
  - Compression (a zipped file with a .epub extension)
  - Navigational elements (a set reading order)
    - manifest lists the content elements
    - spine enumerates the order of files
  - Structural elements (eg, <section>, <aside>, and <figure>)
  - Book-wide metadata (based on Dublin Core)

EPUB 3 indexes

- Ebook indexes can be created using plain XHTML coding. Going beyond that, the EPUB 3 Indexes Specification (http://www.idpf.org/epub/idx/) provides a standard that allows for ebook indexes that do everything print indexes do and more.
  - Charter put forward by ASI, approved December 2011
  - Specification approved March 2014
  - Waiting for:
    - Adoption Readiness Roadmap
    - Implementation by reading systems & publishers

EPUB 3 Indexes Adoption Readiness Roadmap

- Supporting documentation – working groups to consider the ‘big picture’ for implementation of the specification.
  - Samples
  - Epubcheck support (for validation)
  - EPUB Conformance Test Suite (epubtest.org) – for testing an ebook reading system's support for EPUB 3 features
  - Authoring tool support
  - Reading system support

EPUB 3 implementation by reading system developers & publishers

> ‘…upgrading to EPUB 3 is not a trivial undertaking, nor is it one that can be reasonably taken unilaterally. ‘

Sanders Kleinfeld
http://toc.oreilly.com/2013/02/oreillys-journey-to-epub-3.html

EPUB 3 Indexes Specification

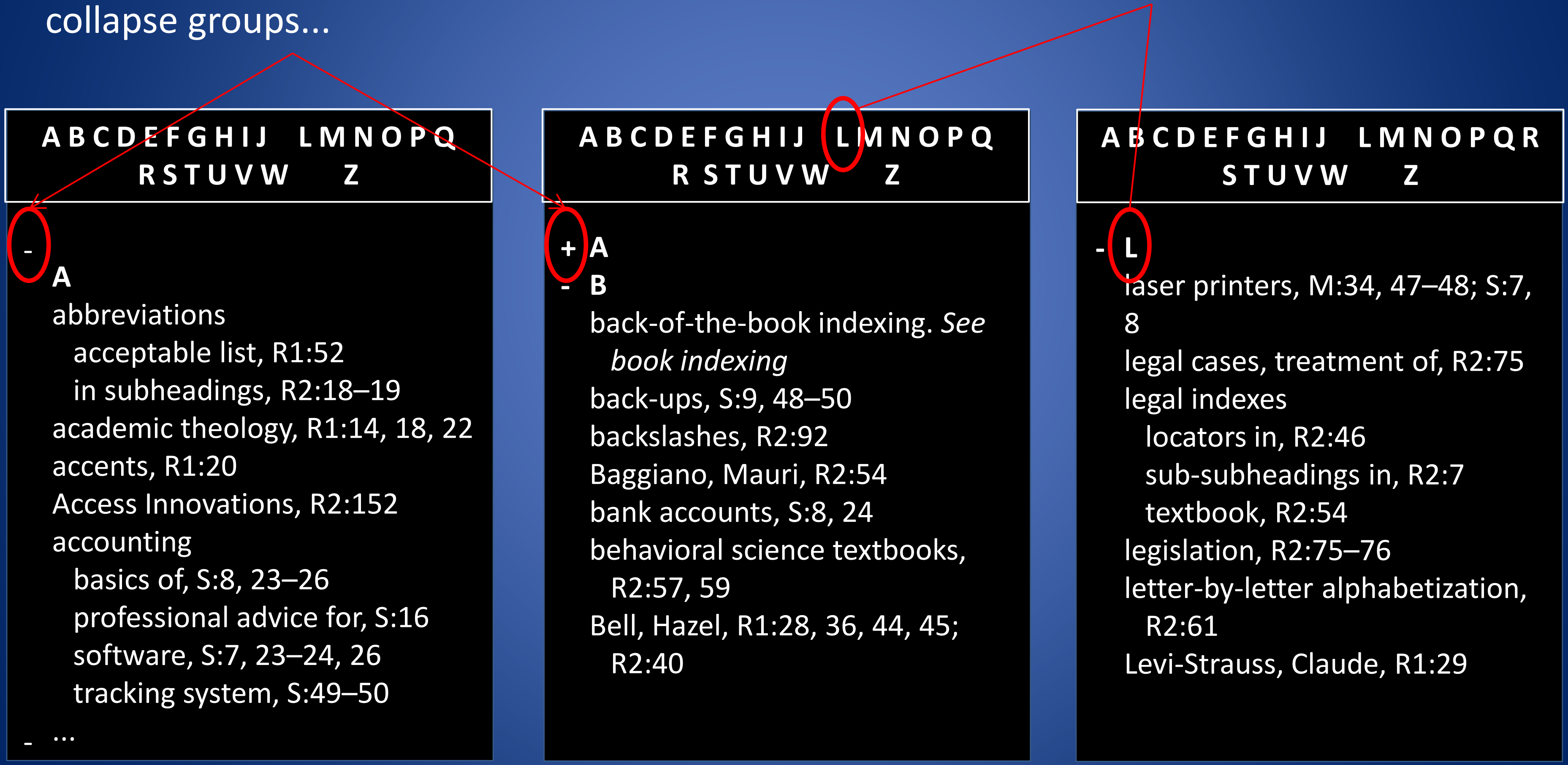
EPUB 3 indexes – how are they different?

- Better navigation
- New features
  - Index filtering
  - Range highlighting
  - Interactive generic cross references

Better Navigation

Index groups allow the user to expand and collapse groups...

...or easily access a group by clicking the letter to jump to it



A B C D E F G H I J L M N O P Q R S T U V W Z

-

A

abbreviations
  acceptable list, R1:52
  in subheadings, R2:18–19
academic theology, R1:14, 18, 22
accents, R1:20
Access Innovations, R2:152
accounting
  basics of, S:8, 23–26
  professional advice for, S:16
  software, S:7, 23–24, 26
  tracking system, S:49–50

- ...

A B C D E F G H I J L M N O P Q R S T U V W Z

+ A

- B

back-of-the-book indexing. *See book indexing*
back-ups, S:9, 48–50
backslashes, R2:92
Baggiano, Mauri, R2:54
bank accounts, S:8, 24
behavioral science textbooks, R2:57, 59
Bell, Hazel, R1:28, 36, 44, 45; R2:40

A B C D E F G H I J L M N O P Q R S T U V W Z

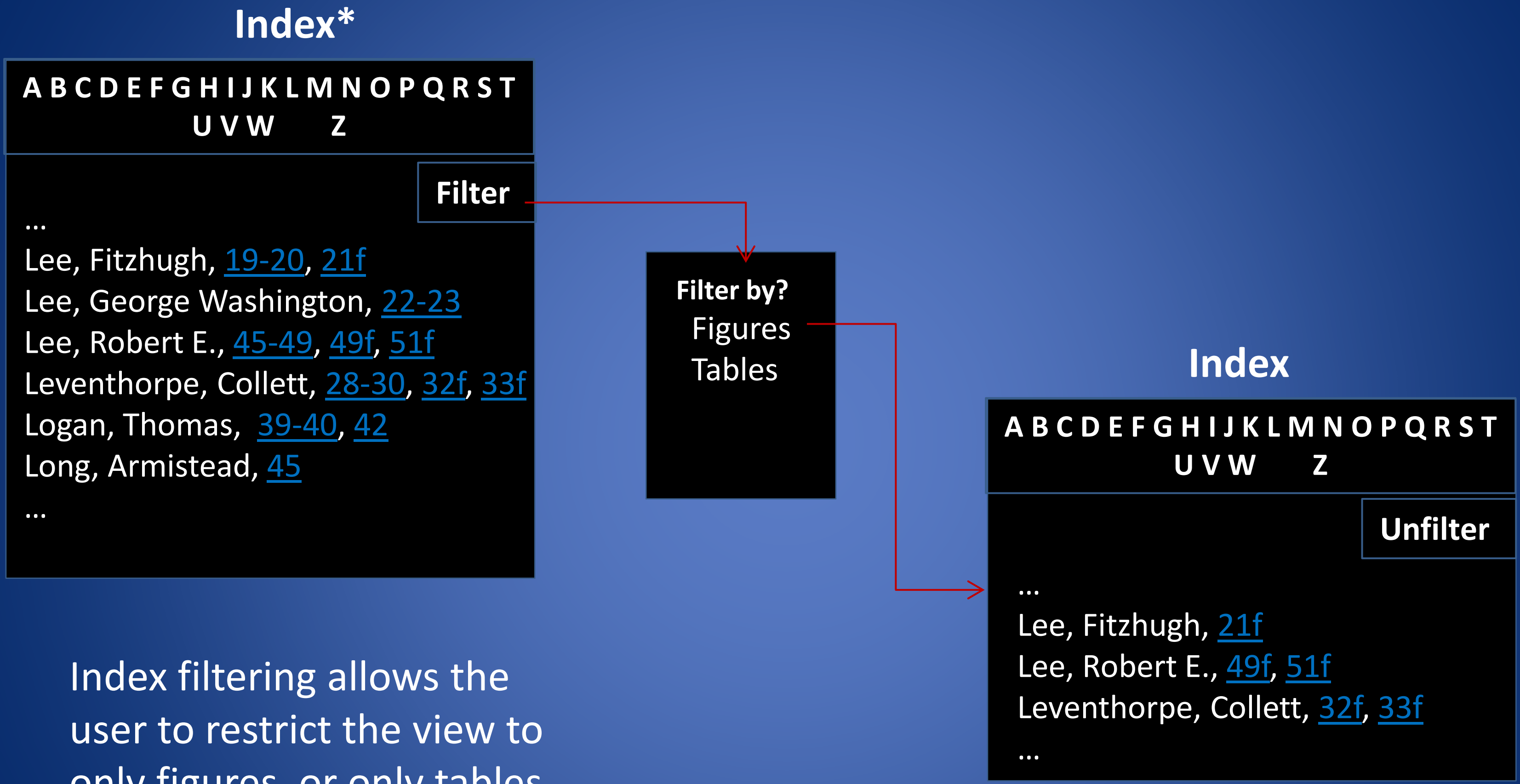
- L

laser printers, M:34, 47–48; S:7, 8
legal cases, treatment of, R2:75
legal indexes
  locators in, R2:46
  sub-subheadings in, R2:7
  textbook, R2:54
legislation, R2:75–76
letter-by-letter alphabetization, R2:61
Levi-Strauss, Claude, R1:29

(Combs)

[redacted]

Index Filtering



Index*

A B C D E F G H I J K L M N O P Q R S T U V W Z

Filter

…

Lee, Fitzhugh, 19-20, 21f

Lee, George Washington, 22-23

Lee, Robert E., 45-49, 49f, 51f

Leventhorpe, Collett, 28-30, 32f, 33f

Logan, Thomas, 39-40, 42

Long, Armistead, 45

…

Filter by?

Figures

Tables

Index

A B C D E F G H I J K L M N O P Q R S T U V W Z

Unfilter

…

Lee, Fitzhugh, 21f

Lee, Robert E., 49f, 51f

Leventhorpe, Collett, 32f, 33f

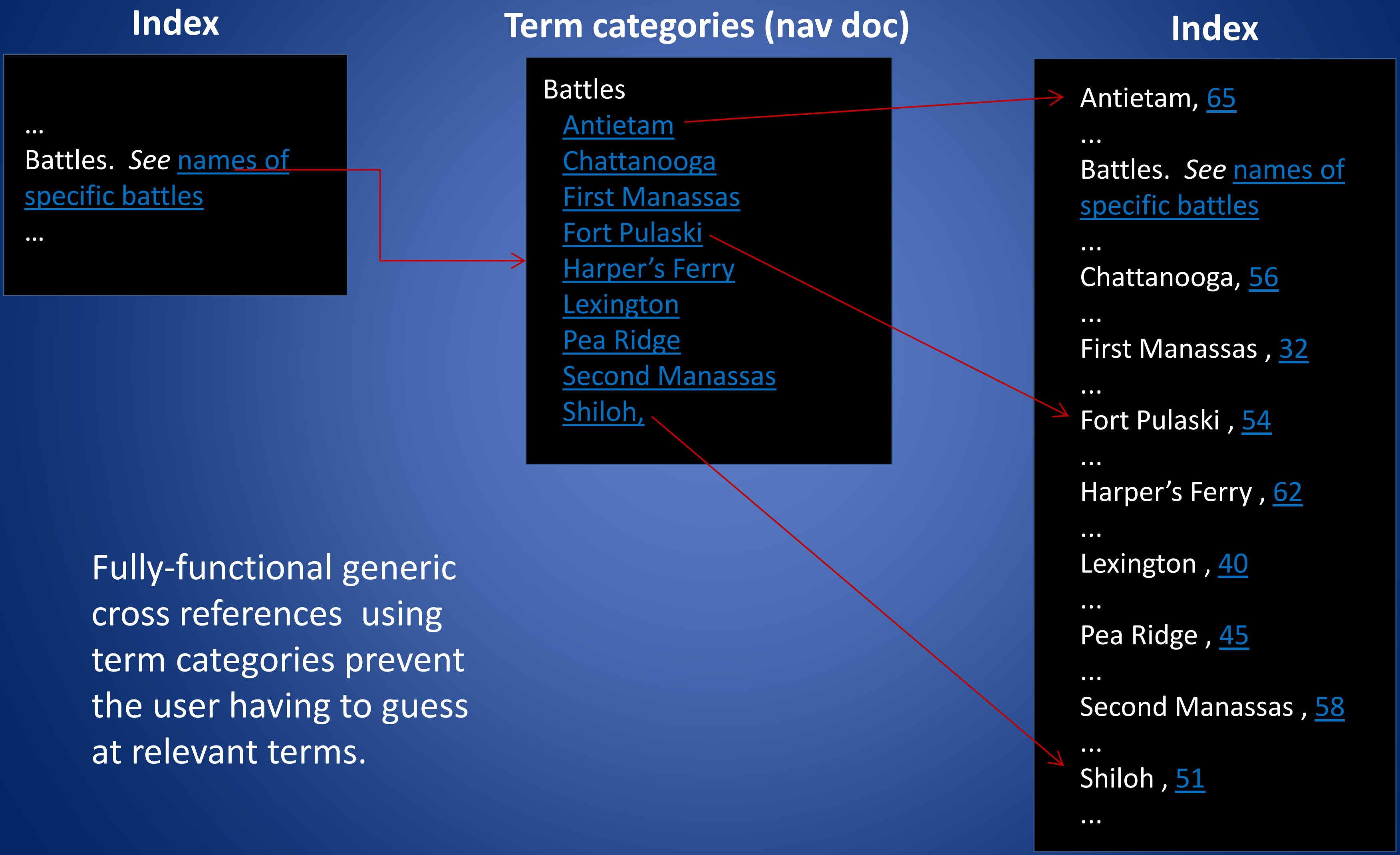
…

Index filtering allows the user to restrict the view to only figures, or only tables, for quick reference.

*** "f" following a locator indicates a figure**

(Combs)

Generic cross references



Index

…

Battles. *See* names of specific battles

…

Term categories (nav doc)

Battles

- Antietam
- Chattanooga
- First Manassas
- Fort Pulaski
- Harper's Ferry
- Lexington
- Pea Ridge
- Second Manassas
- Shiloh,

Index

Antietam, 65

…

Battles. *See* names of specific battles

…

Chattanooga, 56

…

First Manassas , 32

…

Fort Pulaski , 54

…

Harper's Ferry , 62

…

Lexington , 40

…

Pea Ridge , 45

…

Second Manassas , 58

…

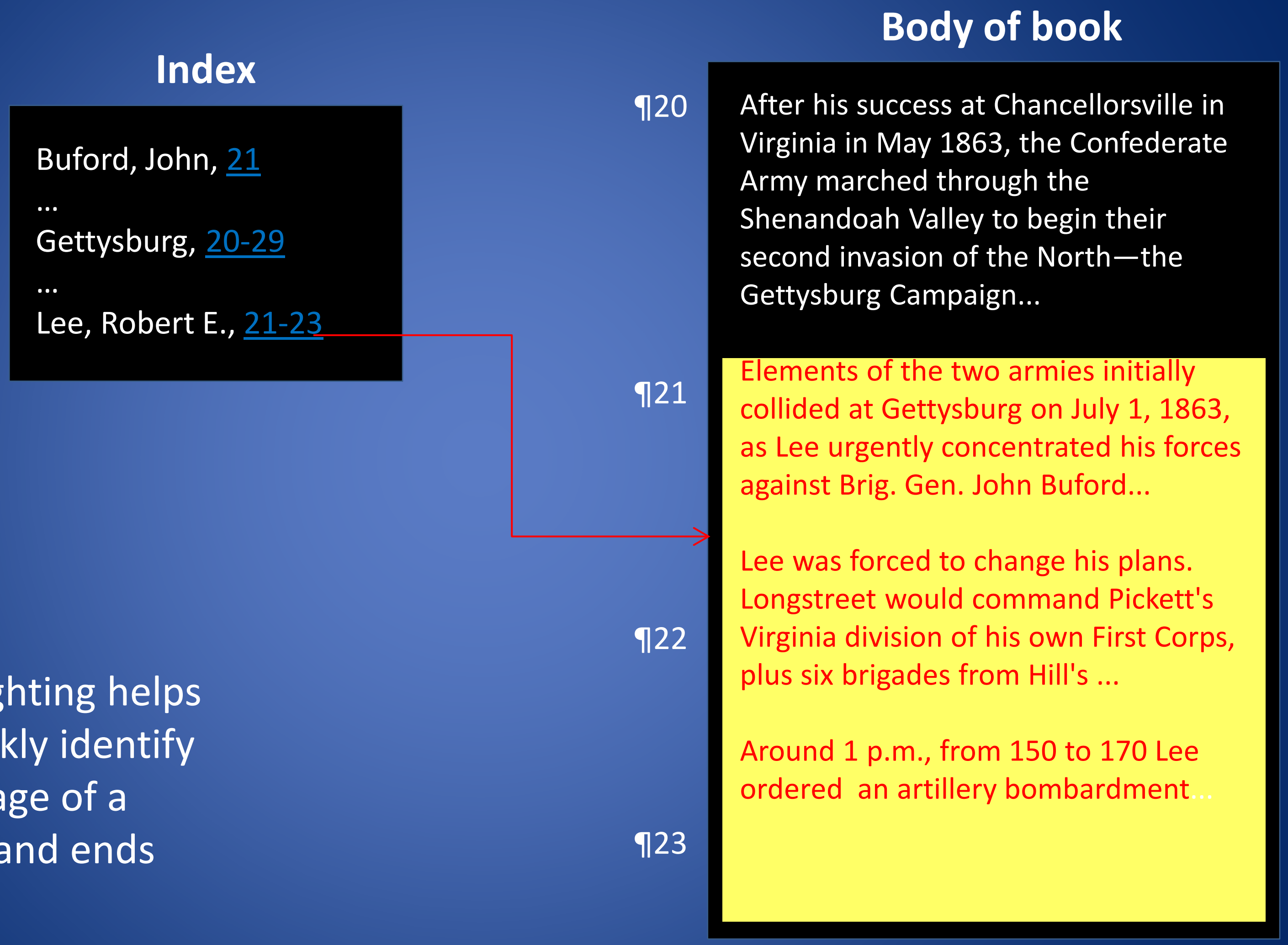
Shiloh , 51

…

Fully-functional generic cross references using term categories prevent the user having to guess at relevant terms.

(Combs)

Range Highlighting



Index

Buford, John, 21
…
Gettysburg, 20-29
…
Lee, Robert E., 21-23

Body of book

¶20 After his success at Chancellorsville in Virginia in May 1863, the Confederate Army marched through the Shenandoah Valley to begin their second invasion of the North—the Gettysburg Campaign...

¶21 Elements of the two armies initially collided at Gettysburg on July 1, 1863, as Lee urgently concentrated his forces against Brig. Gen. John Buford...

Lee was forced to change his plans. Longstreet would command Pickett's
¶22 Virginia division of his own First Corps, plus six brigades from Hill's ...

Around 1 p.m., from 150 to 170 Lee ordered an artillery bombardment...

¶23

Range highlighting helps the user quickly identify where coverage of a topic begins and ends

Software tools for ebook indexing

Software options

- Hyperlinked (or tagged) indexes
- Embedded indexing

Hyperlinked indexes

- Standalone indexes are created and linked to standard XHTML anchors in the text.
- Target in text:

```
id="s2.1“
Eg, <div id="s2.1">s2.1</div>
```

- Index hyperlink

```
<a href="../Text/FreelanceIndexing.xml#s2.1">2.1</a>
```

- The link title/anchor text is 2.1.

Embedded indexes

- Embedded indexes are those in which the index entry is inserted into the text to which it refers.
- Embedded indexing can be done in:
  - Page layout programs, eg, InDesign
  - Word processing programs, eg, MS-Word and Libre Office
  - XML editors

InDesign

- InDesign Creative Cloud natively outputs EPUB indexes. This was not the case with the earlier version, where indexes were lost when EPUB formats were generated.
  - This is the result of work done by American Society for Indexing members and InDesign staff.
  - Kerntiff Publishing System (KPS) has created plugins to aid indexers in handling the interface between dedicated indexing software and InDesign. See http://www.luciehaskins.com for resources.

XML editors

- Embedded index entries contain wrappers for the entry, eg,

  ```
  <index><primary>videos</primary><secondary>slow moving images in</secondary></index>
  ```

  ```
  videos
      slow moving images in
  ```

- Need to generate the index to see it, then edit the entries in the text.
- Small errors are easy to introduce.
- File management is crucial when embedding.

XML editors – some of the problems

- Once an entry has been created, the only way to create subentries or sub-subentries is to repeat the entire entry from its main heading .
- The index must be recreated each time you want to check whether the entries have been inserted correctly.
- It is not possible to edit the index without going back to individual entries in the HTML coding.
- The HTML coding for each index entry is long. When there are a lot of entries together it is difficult to scan them and even to identify the text.

Software choices

- The software indexers use will largely depend on the approaches publishers are taking to creating ebooks.
- Self-publishers will have to make software decisions and learn technology themselves. Advising them will be one of the challenges for indexers in the future.

Approaches to indexing ebooks

Ebook indexing decisions, some

- Indexing for small (or variable-sized) screens
- Text targets (granularity of location display)
- Locator link text (what the user clicks on)

Indexing for small (or variable-sized) screens

- You may choose to use:
  - Shorter entries and subentries to avoid turnover lines
  - Shorter subentry lists so main entries don’t scroll off the screen
- The reading system may be able to make it easier for users to view subentries.

Text targets – page break references

- Page numbers may retain value for legacy books for citation and comparison with printed versions, and for books in PDF format. They can be included as anchors without being displayed.

Text targets – paragraphs or sections

- Paragraphs and/or sections provide a logical grouping within the text.
- Indexing to paragraphs is usually easier to create, edit, update and translate than indexing that is scattered throughout the text.

Pinpoint indexing

- Exact locations might work better for
  - specific terms such as names
  - short discussions
  - users who like to get straight to the core of the matter.

Text targets

- The software approach you are using may limit the text targets you can use.
- If you rely on IDs that are already embedded in the text, these will be limited, eg, to paragraph level. Putting unique IDs at every word would add too much data to the content files.
- Embedded entries can be added at paragraph or word level. It may be easier to keep track of them if they are grouped.
- Indexing decisions about text targets will largely depend on book-level decisions.

Granularity of location display

	Entries linked to unique IDs in content	Entries embedded in content
When you click an index entry, how close do you get to the right content?	Can point to paragraph or word level. Must choose a concrete page if a page number is displayed on screen. Page number could be wrong if content shifts. Indexer chooses displayed number. If you rely on already embedded anchors, these will be limited, eg, to paragraph level. Unique IDs at every word adds too much data to the content files, which must remain within limits for EPUB.	Can point to paragraph or word level. Module chooses page number to display. Can be wrong if index has not been regenerated. Most modules still work as snapshots, and need to be regenerated every time content changes.

Locator link text

- Locator links (the bits of the index that users click on) can be:
  - Page numbers
  - Paragraph and/or section numbers
  - Running numbers (eg, 1, 2, 3)
  - Symbols (eg, *, *, *)
  - Main entry or subentry text.
- These alternatives are similar to those that have been considered for website indexes over the last decade. (Browne and Jermey)

Introduction to Indexing experiment

- I created a mini-EPUB ebook based on five articles from my website.

  free at http://www.lulu.com

- I indexed to section and subsection level, and created two versions of the index. In one the numbers are visible; in the other the main entries or subentries are the clickable links.

Introduction to Indexing experiment

- The 'cleaner' one, in which the entries are links, is more attractive.

  indexing standards

- The one with paragraph numbers displayed is more useful.

  indexing standards 3, 6.1

  - The numbers provide information about the possible length of the content being targeted.
  - The numbers help the user home in on the required text when they get to the location in the text.
  - Including the paragraph numbers in the index means that you can have more than one entry per main entry or subentry.

Publisher decisions

Commissioning ebook indexes

- Things to consider
  - audience
  - software for creation
  - reading devices
  - planned reuse

Commissioning ebook indexes

- Planned reuse
  - Mashups – combining many books into one.
  - Subsets – separating one book into many mini-books.
  - Translations – will you translate the index or re-index.
  - Later editions – both hyperlinking and embedded indexing aid reusability, but care needs to be taken.

Indexes as a marketing tool

- Indexes for discovery

  ‘I'm bored with ebook samples… Let's start with the index…Give me a sense of what amount of coverage I can expect on every topic right there in the sample…’ (Joe Wikert)

Ebook indexing future

- Trinity: If you tell me we'll make it I'll believe you.
- Neo: We'll make it. We have to.