Understanding the IDPF EPUB3 Indexes Specification

David K. Ream, Leverage Technologies, www.levtechinc.com
Michelle Combs, Carpe Indexum, www.carpeindexum.com

ASI representatives to the International Digital Publishing Forum
Co-chairs, EPUB3 Indexes Working Group

Introduction

- ASI's Digital Trends Task Force concluded that an encoding standard was the best way to promote indexes in eBooks
  - Gives reading systems the ability to better integrate indexes with the content
  - Gives publishers a common way to encode an index's structure
- DTTF proposed the International Digital Publishing Forum (IDPF) as best avenue to standard development/acceptance
- ASI board approved IDPF membership

Timeline 2011 – 2014

- Submit charter proposal with four use cases to IDPF [2011]
- Establish working group [2012]
- Define structural components of an index
- Write specification [2013]
- Develop schema, coding examples
- Final edits and submission [2014]
- Currently awaiting final IDPF membership vote

EPUB3

- "Owned" by IDPF, approved May 2011
- Built on other standards: HTML5, XML, CSS, Zip, …
- Working groups develop sub-specifications: EPUB3, Fixed Layout, Indexes, Dictionaries, etc.
- `epub-type` attributes specify structural elements in content documents
- Content documents plus other information form an EPUB "package"

In Scope for Indexes Spec

- Define the specification within the EPUB3 protocols
- How are one or more indexes identified to the reading system
- How indexes are represented by multiple files
- Provide a consistent way of encoding the structure and content
- Provide some implementation suggestions

Out of Scope for Indexes

- Index content, format, style, order
- How/whether a Reading System (R/S) should
  - Use the tagging for rendering
  - Provide access to the index(es)
  - Deliver additional functionality
- Broader EPUB3 issues, such as the interaction between searching and index(es)
- What should be used as locators
- What type of linking is used

Specification Information

- epub-types and which HTML elements they can be applied to
- epub-types can sometimes be implied from context reducing verbosity
- Often multiple ways to encode epub-types
- Ancillary files: schema, samples, etc.

Encoding – Index and Notes

- Several `epub-type` values are involved:
  - index
  - index-head-notes
  - index-legend

Encoding – Index

Entire index is wrapped in "**index**"

document contains only a single index

```
<body epub:type="index">
```

document contains index + other content

```
<body>
  …
  <section epub:type="index">
  …
</body>
```

Encoding – Head notes

```
<body epub:type="index">
 <header epub:type="index-headnotes">

  <h1>Subject Index</h1>
  <p>Alphabetization is word-by-word:
   New York comes before Newtown.</p>

 </header>

 …
 [index entries]
</body>
```

Encoding – Legend

```
<header epub:type="index-headnotes">
 <p>The following abbreviations are
  used in this
  index.</p>
 <dl epub:type="index-legend">
  <dt>Civ. R.</dt><dd>Civil Rule</dd>
  <dt>Crim. R.</dt><dd>Criminal
  Rule</dd>
  <dt>§</dt><dd>Statute</dd>
 </dl>
</header>
```

Encoding – Entries

- Requires several `epub-type` values:
  - index-group
  - index-entry-list
  - index-entry
  - index-term
  - index-editor-note

Encoding – Entries

index without group breaks

```
<section epub:type="index">
  [head notes if present]
  [entries] ...
</section>
```

Encoding – Entries

index with group breaks

```
<section epub:type="index">
 <section epub:type="index-group">
  <h1>A</h1>
  [entries beginning with "A"] ...
 </section>
 <section epub:type="index-group">
  <h1>B</h1>
  [entries beginning with "B"] ...
 </section>
 ...
</section>
```

Encoding – Entries

One entry with two subentries (locators not shown):

```
<ul epub:type="index-entry-list">
 <li epub:type="index-entry">
  <span epub:type="index-term">Black, John</span>
  <ul epub:type="index-entry-list">
   <li epub:type="index-entry">
    <span epub:type="index-term">birth</span>
   </li>
   <li epub:type="index-entry">
    <span epub:type="index-term">death</span
   </li>
  </ul>
 </li>
</ul>
```

Encoding – Entries

epub-types may also be implied when inside index or index-group

```
<ul>
 <li>
  <span epub:type="index-term">Black, John</span>
  <ul>
   <li>
    <span epub:type="index-term">birth</span>
   </li>
   <li>
    <span epub:type="index-term">death</span>
   </li>
  </ul>
 </li>
</ul>
```

Encoding – Editor's note

```
<li epub:type="index-entry">
   <span epub:type="index-term">
      Heston, Charlton</span>
   <span epub:type="index-editor-note">
      Actor (1923-2008) in numerous
      American films.</span>
</li>
```

Encoding – Locators

- Requires several `epub-type` values:
  - index-locator
  - index-locator-list
  - index-locator-range
  - content semantics

Encoding – Locators

Page number

```
<span epub:type="index-term">
 telephone</span>
<a epub:type="index-locator">35</a>
```

Section number

```
<a epub:type="index-locator">13.27</a>
```

Image ()

```
<a epub:type="index-locator"><img
alt="phone" src="phone-icon.png"/></a>
```

Encoding – Locators

Term itself used as locator

```
<span epub:type="index-term">
 <a epub:type="index-locator">
  telephone</a>
</span>
```

Encoding – Locators

Multiple locators wrapped in `index-locator-list` (note that `index-locator` is implied on <a>'s)

```
<span epub:type="index-term">
 telephone</span>
<ul epub:type="index-locator-list">
 <a href="…">35</a>
 <a href="…">42</a>
 <a href="…">113</a>
</ul>
```

Encoding – Locators

Range locator is wrapped in `index-locator-range` (note that `index-locator` is implied on <a>'s)

```
<span epub:type="index-locator-range">
   <a href="chap2.xhtml#p076">76-79</a>
</span>
```

or

```
<span epub:type="index-locator-range">
   <a href="chap2.xhtml#p076">76</a>-
   <a href="chap2.xhtml#p079"> 79</a>
</span>
```

Encoding – Locators

Locators with semantics

```
<a href="chap14.xhtml#fig7"
 epub:type="index-locator figure">
 35f</a>
```

Other semantics available in EPUB3:

- footnote
- table
- appendix
- etc.

Encoding – Links

- <a> element `href` attribute
  - Typical web and eBook linking method
  - Requires anchor (`id`) in the content at the destination
- Canonical Fragment Identifiers (CFIs)
  - Typical bookmark and annotation method
  - Provides directions to destination, so no anchor is needed in content
- Eventually may be tools that can convert one format into the other

Encoding – Cross-references

- Requires several `epub-type` values:
  - index-preferred (see, voir, véase)
  - index-related (see also, voir aussi, véase también)
  - index-category (see *names of specific battles*)
  - index-categories [occurs in the navigation document]

Encoding – Cross-references

Term with single *see* cross reference

```
<span epub:type="index-term">
 Peiking</span>
 <a epub:type="index-xref-preferred
   index-term">Beijing</a>
```

Encoding – Cross-references

Term with *see also* references

```
<span epub:type="index-term">Sugars</span>
  <a epub:type="index-locator">48</a>
  <span epub:type="index-xref-related">See also
    <a href="..." epub:type="index-term">
    glucose</a>;
    <a href="..." epub:type="index-term">
    sucrose</a>.
  </span>
```

Encoding – Cross-references

Term with *see* reference to a category of terms ...

```
<span epub:type="index-term">battles</span>
<span epub:type="index-xref-preferred">See
   <a epub:type="index-term-category"
    href="nav.xhtml#battles">names of
    specific battles</a>
</span>
```

Encoding – Cross-references

… points to a list in the navigation document (nav.xhtml)

```
<nav epub:type="index-term-categories">
  <ul>
    <li id="battles">battles
      <ol>
        <li><a href="index.xhtml#chan">
        Chancellorsville</a></li>
        <li><a href="index.xhtml#man1">
        First Manassas</a></li>
        ...
      </ol>
    </li>
  </ul>
</nav>
```

Use Cases

- How users might want to interact with the index:
  - Chapter-like index
  - Index locator search
  - Index term search
  - Stand-alone indexes

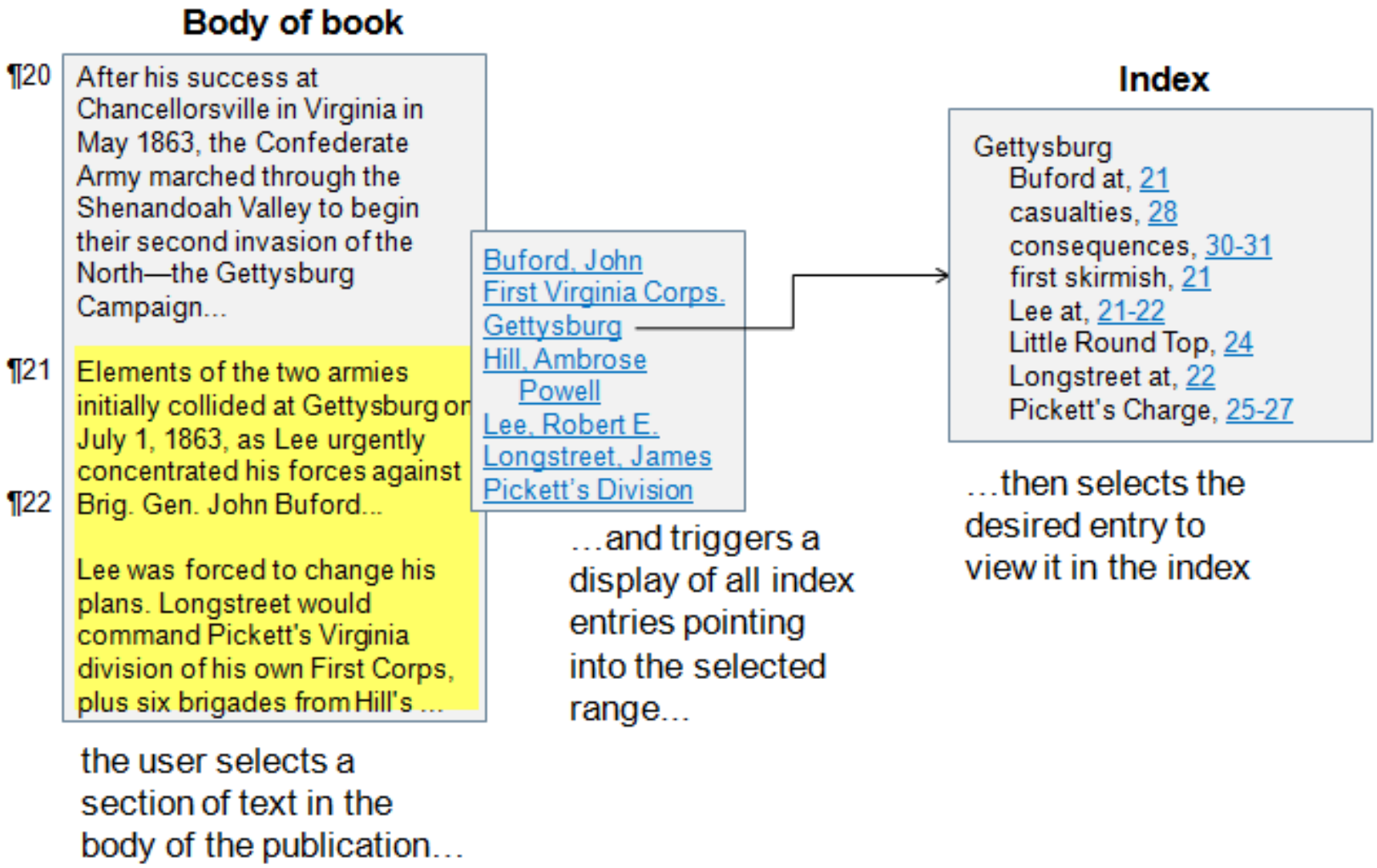
Use Case 1: Chapter-Like

- Traditional approach
- Index(es) appear as last chapter(s) in the book
- User clicks on group break text if present
- User browses through index to desired term
- User clicks on locator to jump to content or web
- User clicks on cross reference to jump within index

Use Case 2: Index Locator Search

- In print books, indexes point to the content but the content doesn't refer to the index
- In eBooks, indexes can create a "two-way street" allowing the user to go from content to index:
  - User highlights section of content
  - Reading system displays all index entries that point to that content
  - User chooses term of interest and jumps to index

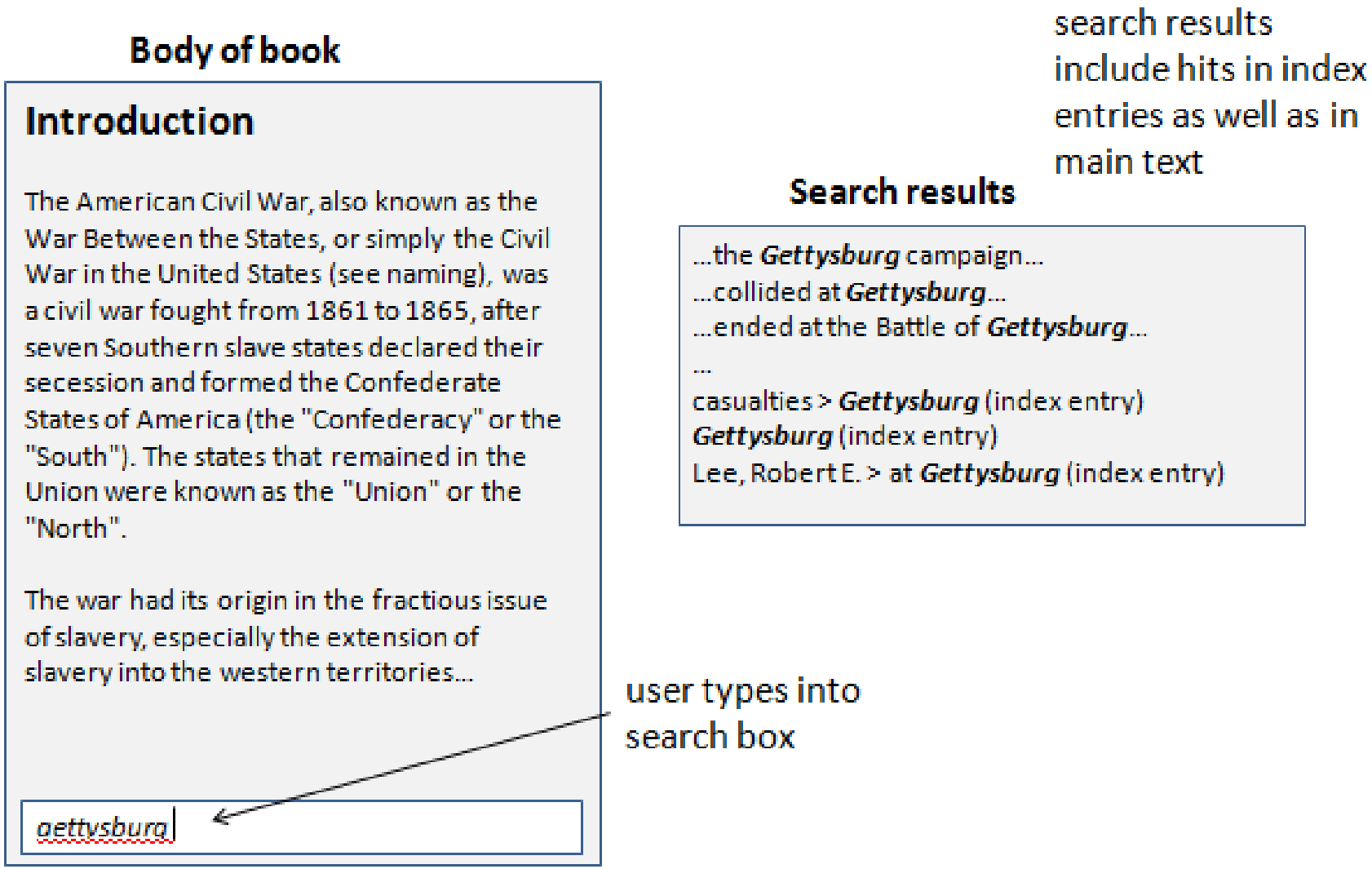
Use Case 2: Index Locator Search



Use Case 3: Index Term Search

- User is viewing content
- User starts to type term
- Index main headings are matched as well as full-text matches in the content
- User can elect to browse the index entries for these headings

Use Case 3: Index Term Search



Use Case 4: Stand-alone Index

- Master or cumulative index to one or more volumes published separately, such as
  - legal books with base volumes and supplements
  - journals with issues collected by year
- Index(es) can be updated independently from the EPUB(s) to which they refer.
- Currently EPUB3 does not support linking between EPUBs.

Miscellany

- Other EPUB files that are affected by the spec:
  - package document
  - navigation document
  - table of contents

Miscellany

Metadata is included in the package document or content documents identifying index content to the R/S

```
<metadata>
  ...
  <dc:type>index</dc:type>
  ...
</metadata>
```

Miscellany

Each index document must be listed in the manifest in the package document:

```
<manifest>
  <item href="index01.xhtml"
  properties="index"/>
  <item href="index02.xhtml"
  properties="index"/>
  ...
</manifest>
```

Miscellany

Indexes that span multiple files are identified as *collections* in the package document:

```
<collection role="index">
  <link href="subjectIndexA.xhtml" />
  <link href="subjectIndexB.xhtml" />
  <link href="subjectIndexC.xhtml" />
  ...
</collection>
```

Miscellany

The landmarks in the navigation document allows the R/S to create a navigation bar for the user interface:

```
<nav epub:type="landmarks">
 <ol>
  … [front matter links]
  <li><a epub:type="toc" href="#toc">
   Table of Contents</a></li>
  … [body and appendix content links]
  … [index(es) content links]
 </ol>
</nav>
```

Miscellany

index documents landmarks structure

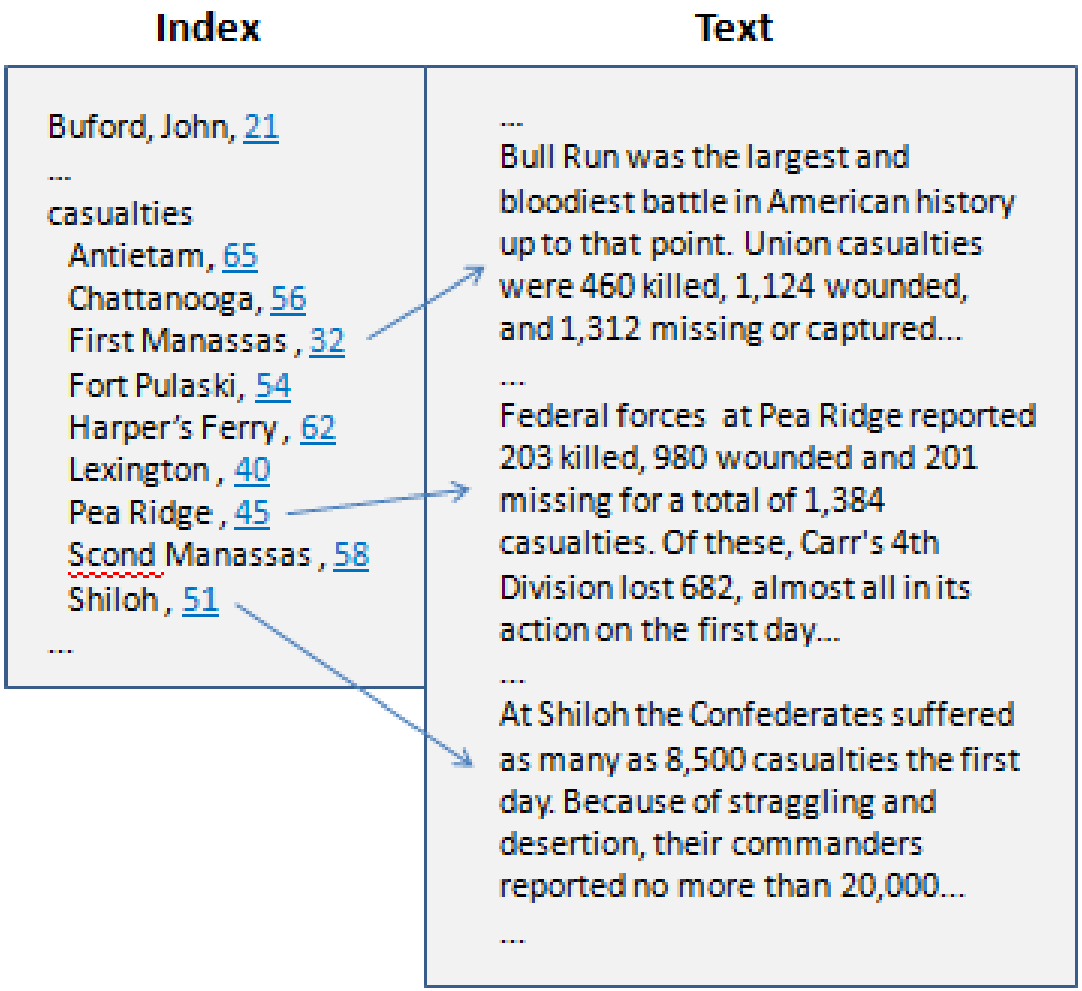
```
<li>
  <a epub:type="index" href="index.xhtml#sbjidx">
  Subject Index</a>
  <ol hidden="">
    <li><a epub:type="index-group"
      href="index.xhtml#A">A</a></li>
    <li><a epub:type="index-group"
      href="index.xhtml#B">B</a></li>
    ...
  </ol>
</li>

<li><a epub:type="index" href="index.xhtml#authidx">
  Author Index</a></li>
```

R/S Suggestions: Navigation

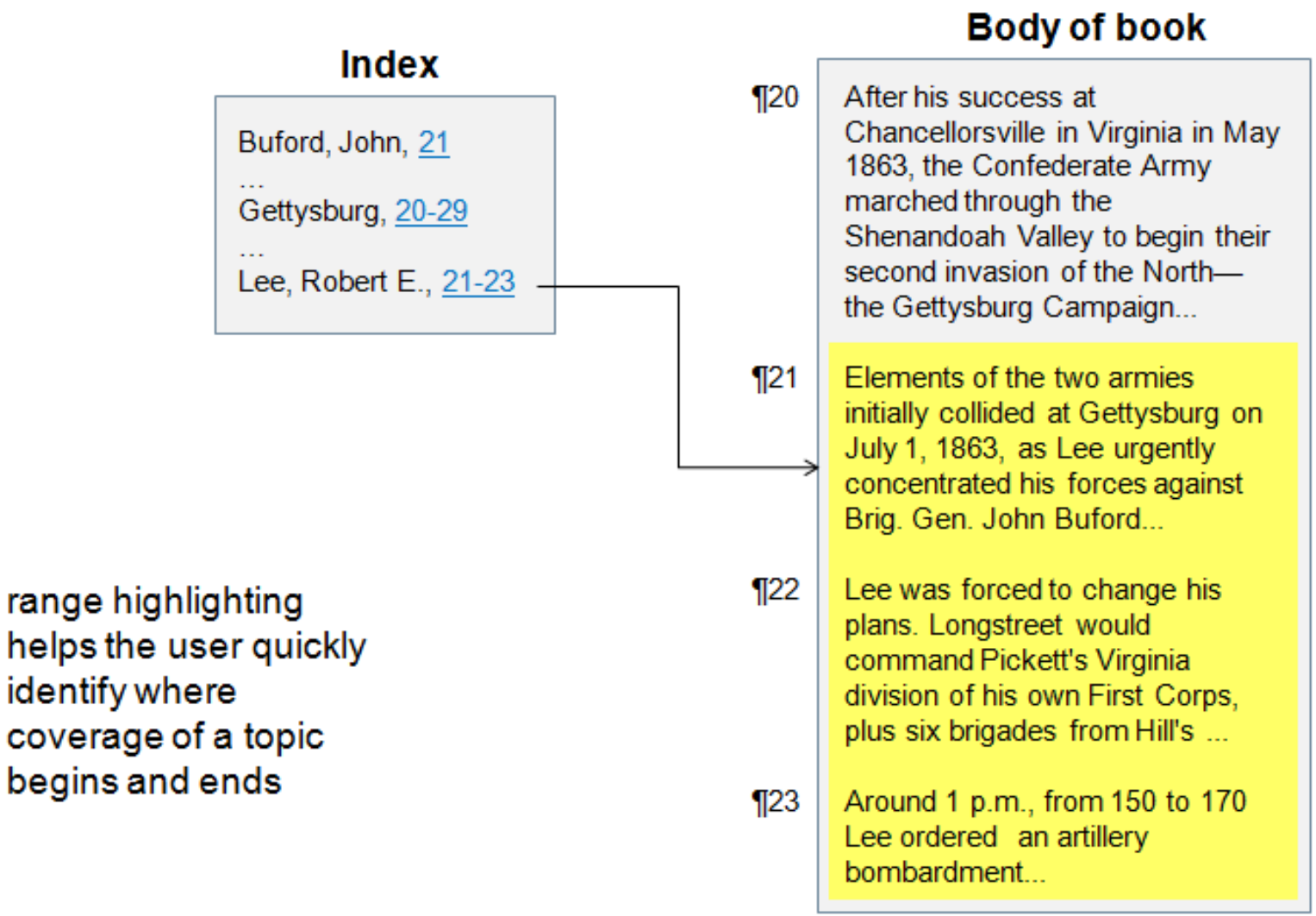
- Show available indexes not just table of contents:
  - menu, search button
  - navigation bar (from landmarks)
- Index(es) appear as option when search is invoked
- Easier back and forth flipping between content and index
- Highlighting of content ranges based on index locator range

Back and Forth Viewing



Side-by-side display allows user to quickly skip through text and locate comparative or related information

Range Highlighting



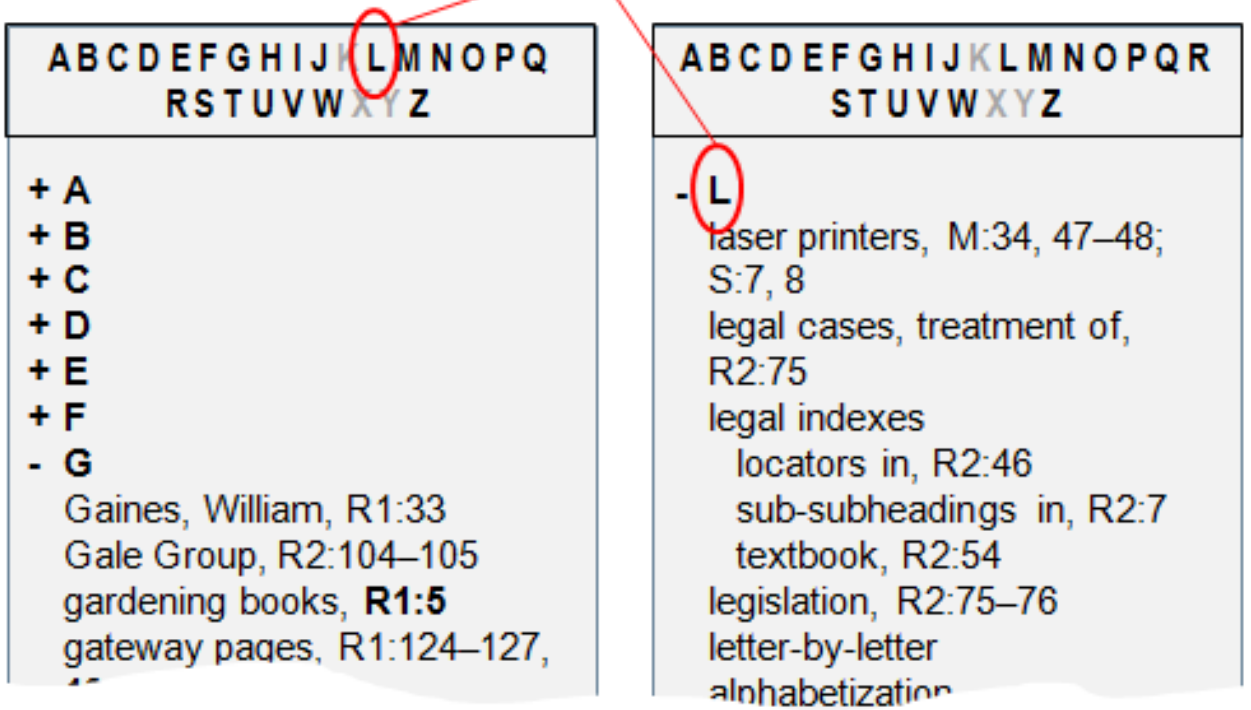
range highlighting helps the user quickly identify where coverage of a topic begins and ends

R/S Suggestions: Index Browsing

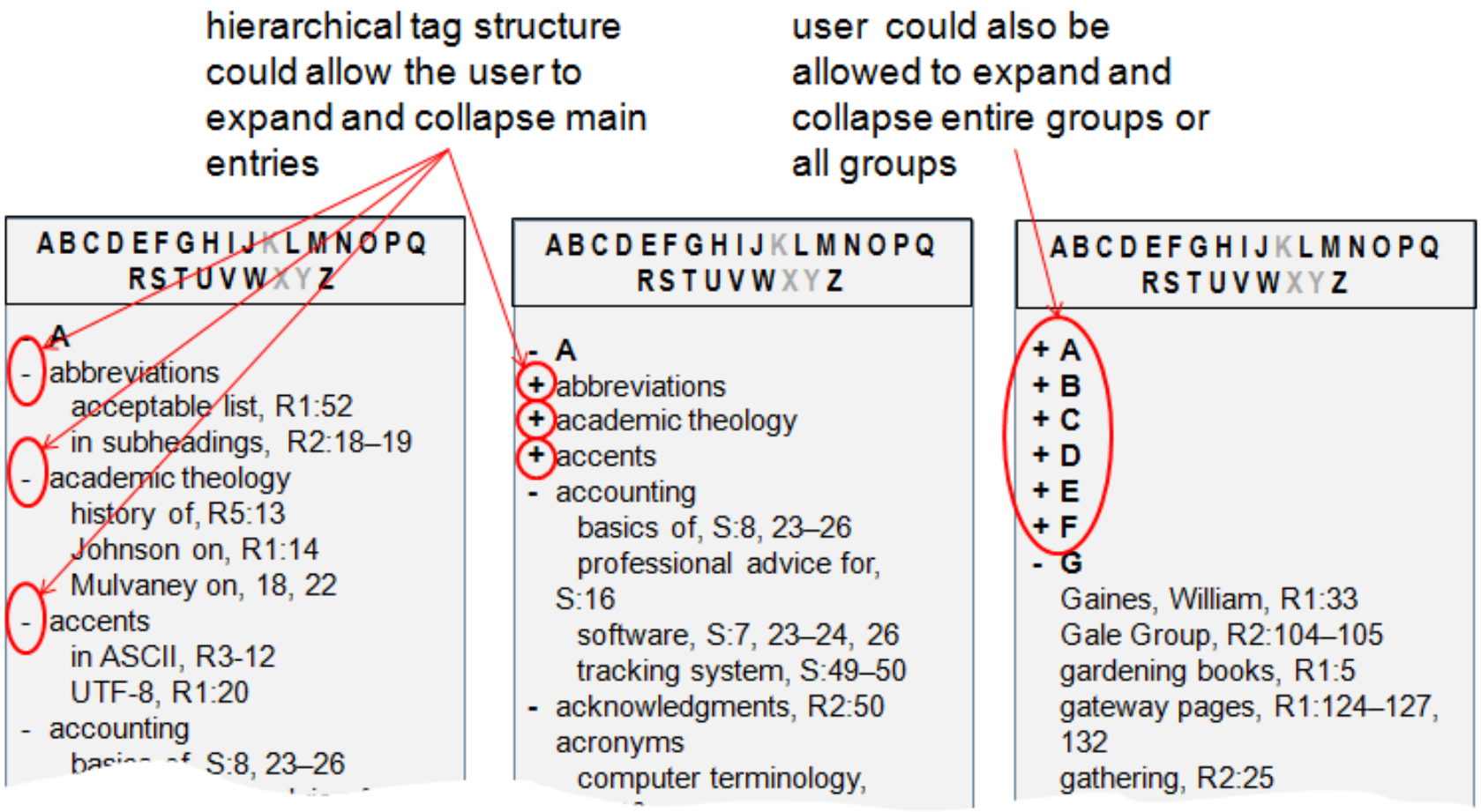
- Group letters are easily accessible
- Legend can be easily recalled
- Type-ahead matching for heading levels
- Clickable heading level breadcrumbs
- Collapsible heading levels
- Pop-up information about locators
- Filter by categories of heading (battles)
- Filter by type of locator (tables, figures)
- Generic cross reference target lists (categories)

Access Index Groups

user can easily access an index group by clicking the letter to jump to it



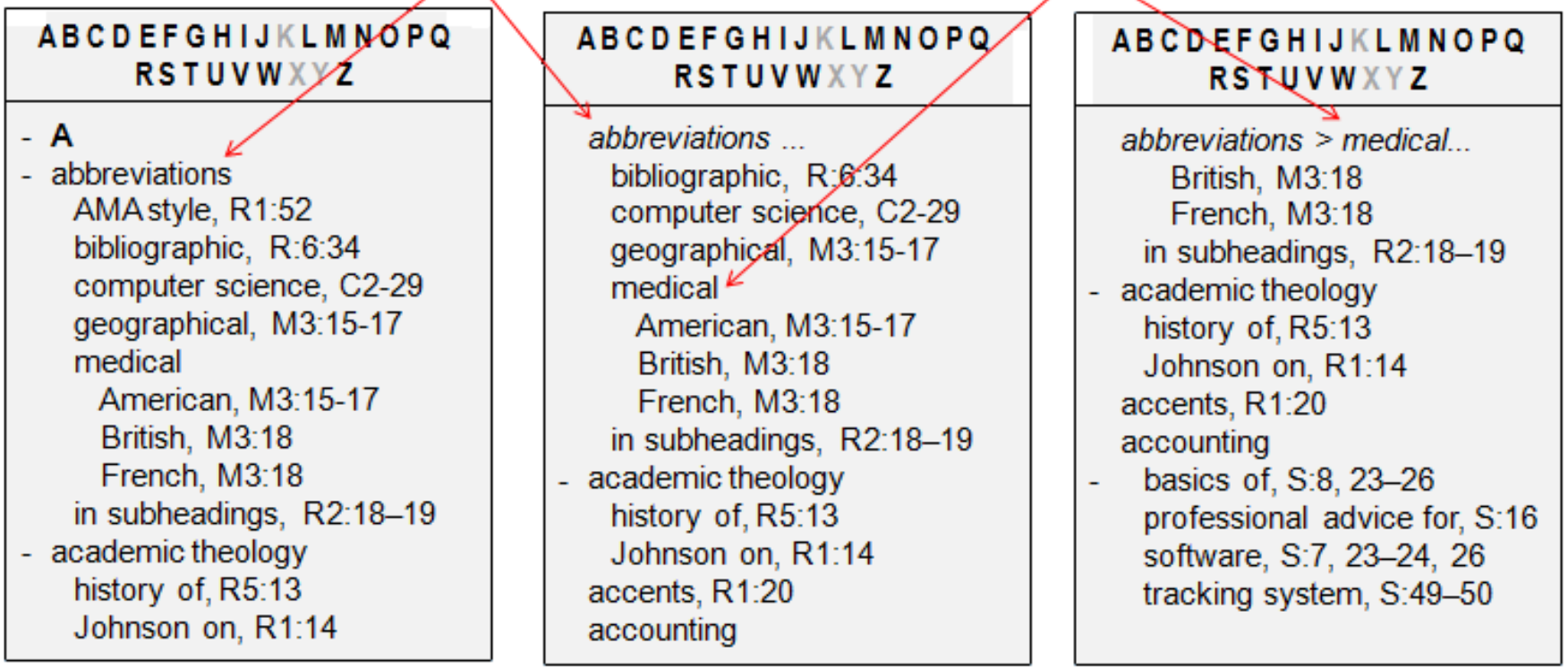
Expand/Collapse Groups



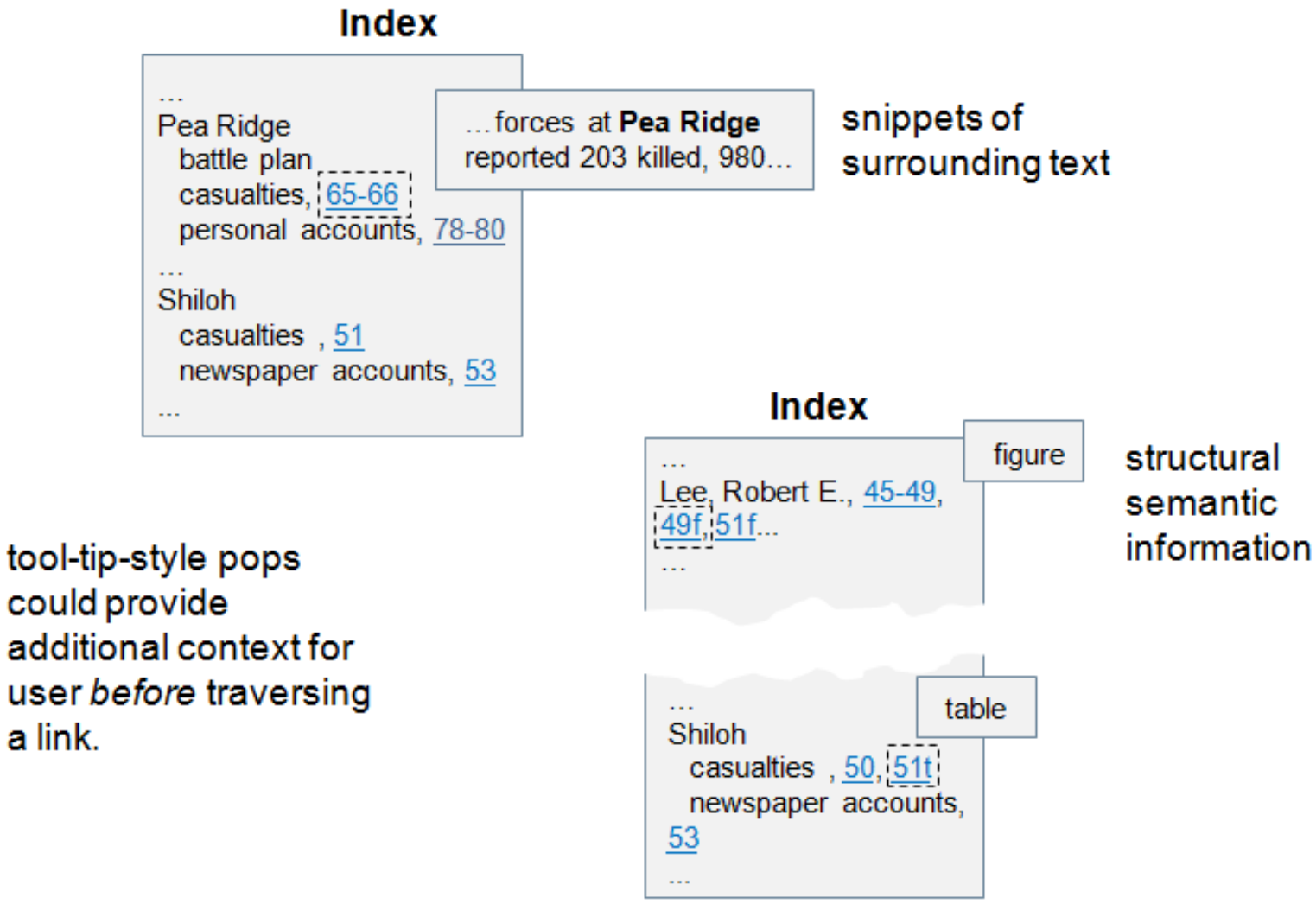
Parent Term Display

as user scrolls down through the index, current parent terms could be persistently displayed as "breadcrumbs" consisting of

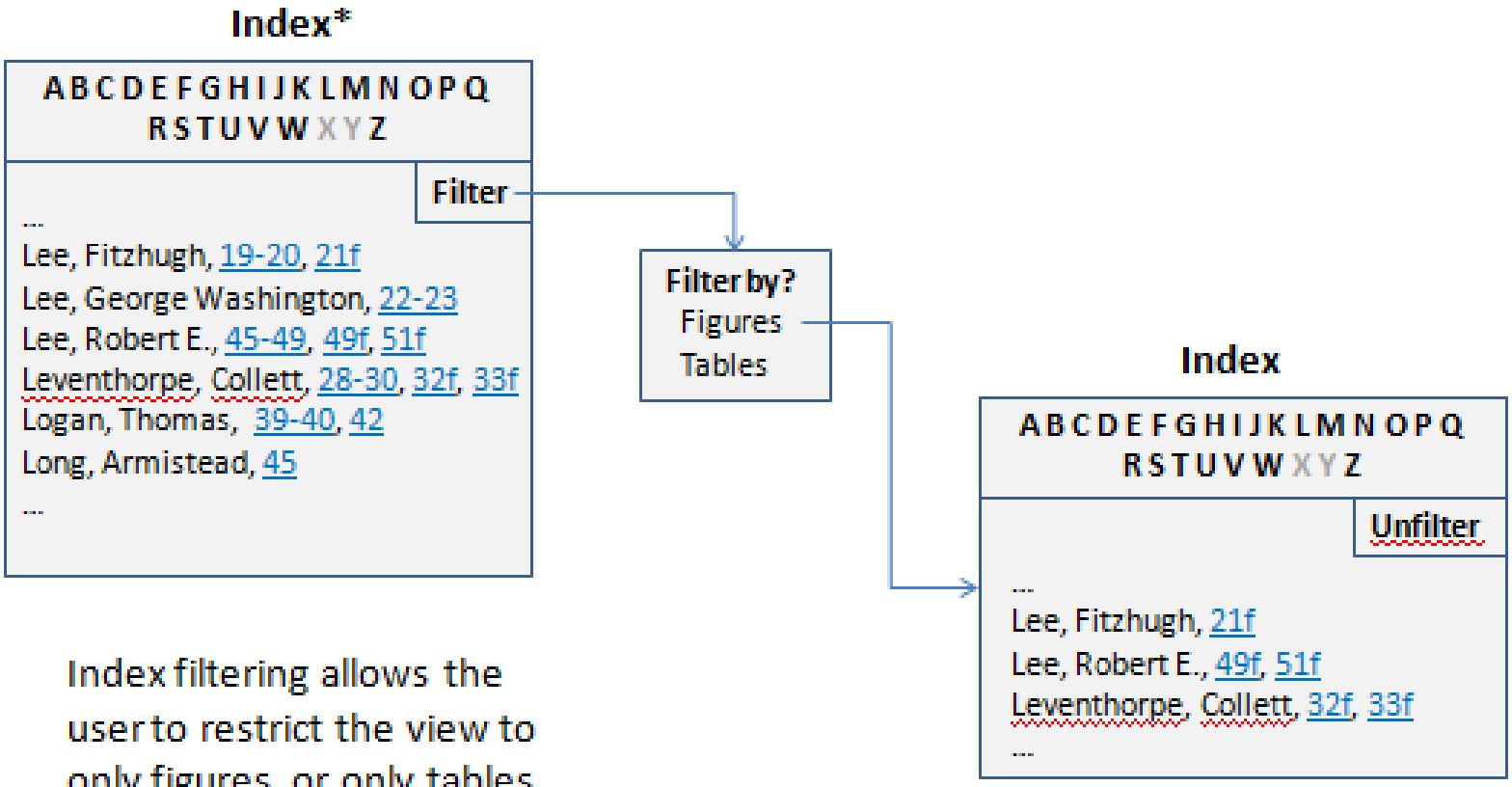
main entry alone... ...or main plus subentry, as applicable



Contextual Information



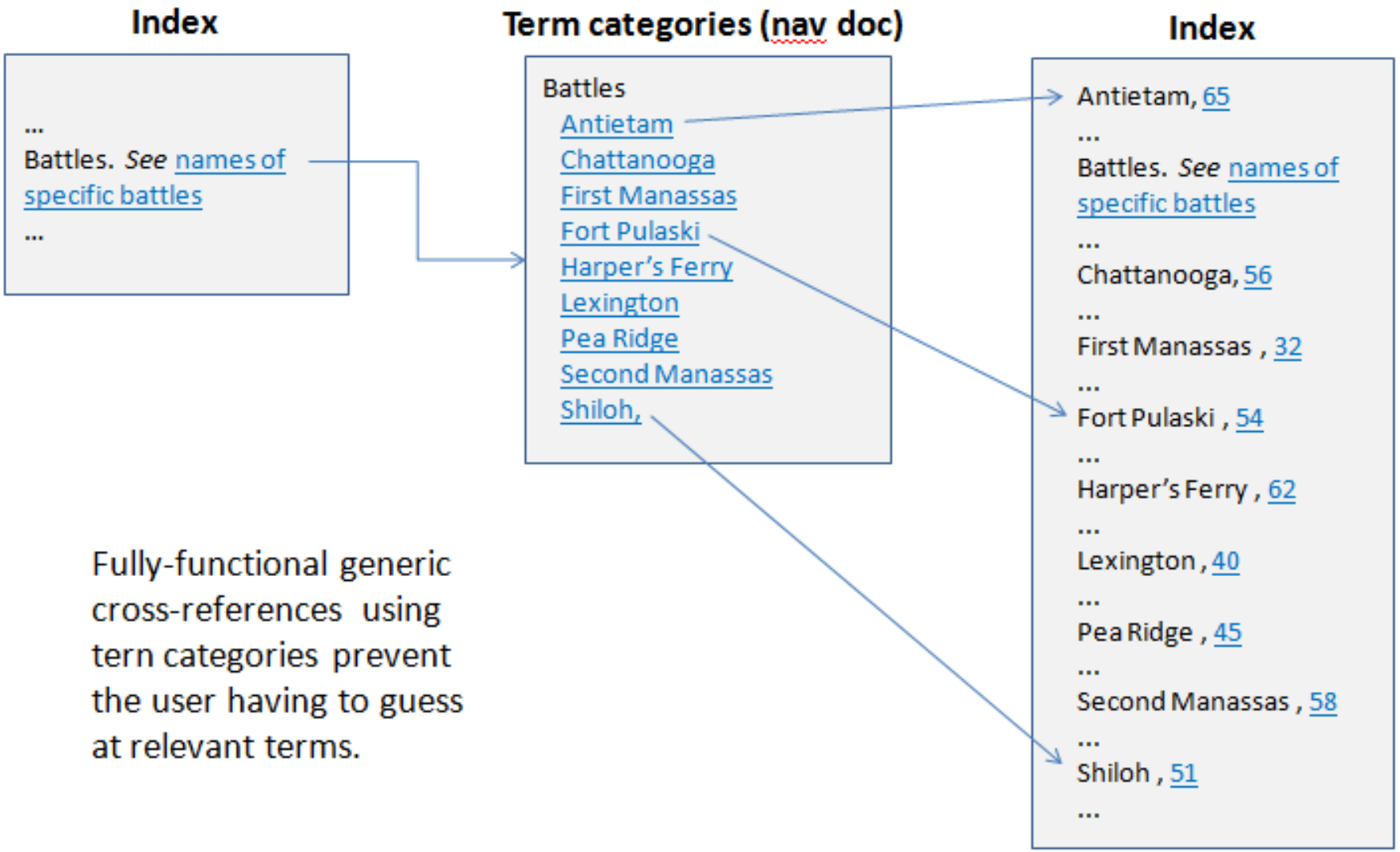
Filtering



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**

Generic Cross-Reference Assist



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

R/S Suggestions: User Settings

- Limit search to:
  - only the content
  - only the index
  - both
- Search results
  - show index hits first
- Filters
  - save filters used

Tools/Software: Indexer

- Embedded entry approach
  - allows for automated creation of links in index and matching anchors in content
  - Adobe InDesign, Word, XML
  - additional scripts required to extract entries, sort/output index, insert anchors into content
  - will vary depending on publisher workflow

Tools/Software: Indexer

- Stand-alone indexes approach
  - CINDEX, Macrex, SKY Index, TExtract
  - InDesign Scripts (Kvern, Kerntiff) for Adobe InDesign
  - HTML/Prep (Leverage Technologies) [EPUB3-compliant]
- Indexing software enhancements needed:
  - better tagging output
  - record/store IDs or link values
  - support for collections
  - support for generic cross references (categories)
  - record/store locator semantics (figure, table, …)

EPUB Tasks

- Break index file(s) into chunks
- Build collections
- Generate TOC entries
- Update navigation and package documents
  - collections, landmarks, metadata, etc.
- Insert or create any term categories
- Apply Dublin Core metadata
- Convert links to CFIs if desired
- Legacy: convert EPUB2 indexes to EPUB3

Distribution & Discovery

- Index crawler
  - index preview
  - main heading extraction for
    - heading match between books
    - books-for-sale searching
- Mashups creation

Summary: What Should Indexers Do Now?

- Get familiar with the standard
  - What it's for
  - What it does and doesn't do
- Take apart an EPUB
  - What's inside an ePub? (video) http://www.lynda.com/InDesign-tutorials/Whats-inside-EPUB-file/123435/125649-4.html
  - Anatomy of an ePub (PDF) http://www.asindexing.org/wp-content/uploads/2014/04/EPUB_Anatomy.pdf

Summary: What Should Indexers Do Now?

- Stay aware of related updates/changes in indexing software and e-readers
- Talk to your publishers about indexes in eBooks
  - Executive Summary: Indexes in Ebooks http://www.levtechinc.com/pdf/ExecutiveSummaryForPublishers.pdf

Links

- Indexes Specification
  http://www.idpf.org/epub/idx
- Indexes Working Group wiki
  http://code.google.com/p/epub-revision/wiki/IndexesMainPage
- DTTF resources
  http://www.asindexing.org/about-indexing/digital-trends-task-force/