The devil is in the details: indexes versus Amazon's X-Ray

Jan Wright

The X-Ray feature available on new editions of Amazon's ereaders could be viewed as a replacement for an index. Comparisons of the details available to readers in X-Ray and in a traditional index show that X-Ray is missing a lot of valuable information, and does not show us, as Amazon claims, 'the bones of the book.'

The devil in the white city by Erik Larson (2003) was a best-selling history of the 1893 World's Columbian Exposition in Chicago, Illinois. Larson combined the story of the architects designing and creating the beautiful buildings for the fair with the tale of a serial killer who used the crowds and festivities of the event to shield his murders. The first Ferris wheel was built for the fair, and many renowned characters passed through the grounds or helped create them, such as Buffalo Bill Cody, Frederick Law Olmsted, Susan B. Anthony, and Franz Boas.

The printed editions of *Devil* had a index of approximately 1,032 lines. The Kindle edition of the book does not include that index. For those who read the book on a Kindle purchased before November 2011, their only access to the contents is through a hyperlinked table of contents, or through the Search feature.

After the release of the Kindle Touch, however, Amazon implemented Kindle X-Ray, a new search and information feature that allows you to find information about characters, events, and topics in books. According to the Amazon website:

> Amazon invented X-Ray, a new feature that lets customers explore the 'bones of the book.' With a single tap, readers can see all the passages across a book that mention ideas, fictional characters, historical figures, places or topics that interest them, as well as more detailed descriptions from Wikipedia and Shelfari, Amazon's community-powered encyclopedia for book lovers.
>
> Amazon built X-Ray using its expertise in language processing and machine learning, access to significant storage and computing resources with Amazon S3 and EC2, and a deep library of book and character information. The vision is to have every important phrase in every book.[1]

As indexers, we have seen many concordance tools and proposed replacements for indexing come and go over the years. We have heard the question 'Wouldn't you just use Search?' a few too many times, and we all have our elevator speeches ready to explain (yet again) why indexes are a superior navigation tool. X-Ray will require us to have yet another version of the speech ready, so we need to examine its workings.

Is X-Ray yet another concordance tool, or is there some real value for readers in this technology? Let us see what kind of results it gives the reader, and compare them with the printed index.

X-Ray functionality

X-Ray, as of this writing, is only available on the Kindle Touch model. If successful, we are sure to see it implemented in the baseline software for other models of the Kindle as well. The new technology was introduced when the Touch was issued in November 2011, and some books were immediately available with the feature implemented. The feature is implemented by the use of a sidefile, a file that is downloaded with the book from Amazon. Books can be updated for X-Ray when the Kindle is connected, and books the reader owns might suddenly gain the X-Ray feature, as more and more files are added to the X-Ray system.

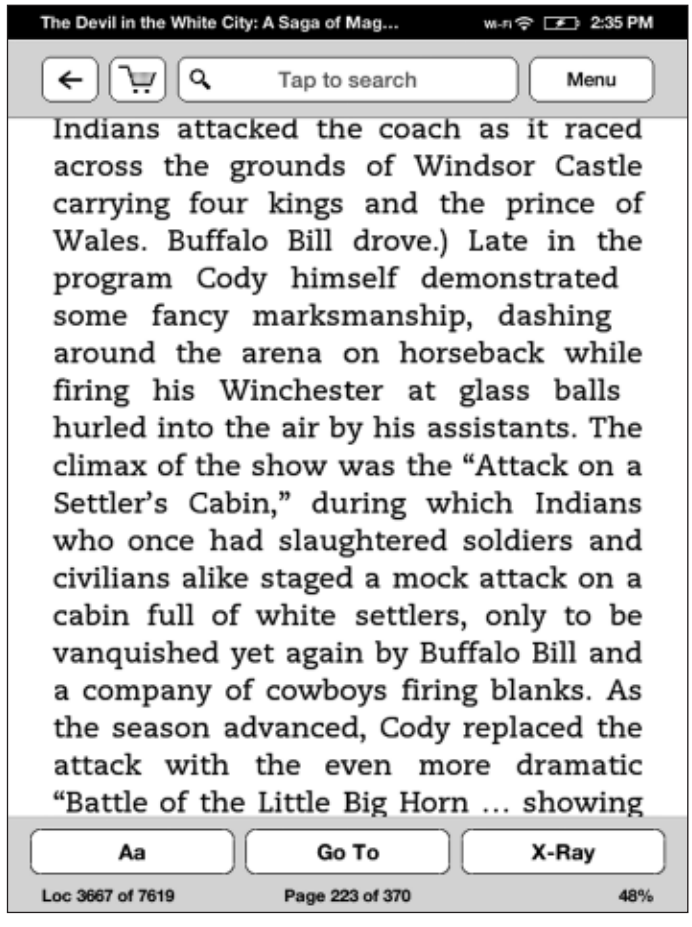

Figure 1 Accessing X-Ray

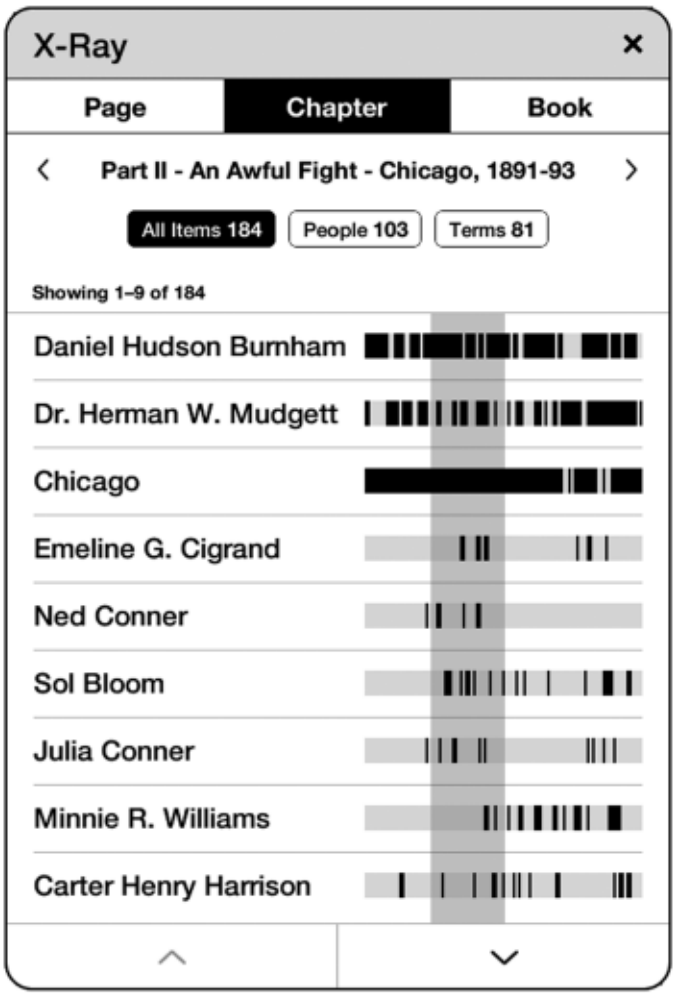

Figure 2 X-Ray's choices for page, chapter, or book filtering, as well as choices between people and term filtering. This screen shows choices for a chapter, and for all entries, both people and terms. The sparkline-like graph shows occurrences throughout the book. The vertical gray bar shows which occurrences are in the current chapter. Note this screen was shot directly after Figure 1.

The sidefile, if available, downloads automatically. This file contains links and information about the book, allowing the reader to access Wikipedia information, Shelfari information, and other terms that have been determined (by human or machine means, we do not know which) to have importance.

To access X-Ray, the reader taps near the top of the Kindle screen, and then chooses the X-Ray button in the lower right (see Figure 1). The feature then shows choices for the reader: examine this page, examine this chapter, or examine the entire book (Figure 2).

One immediate issue is that you cannot sort the results in alphabetical order. You can filter them by chapter or page, or by people or terms, but they are still in non-alphabetical order. The small up and down arrows at the bottom of the screen allow you to navigate through the list. The sparkline-like bars next to each subject show you the occurrences of the topic in the text, and allow you to tap on a vertical line to access the location. The results do not seem to be ordered by importance.

X-Ray has identified 341 items for this book, 205 people and 136 terms. This works out to be about a third of the number of lines in the printed index, so it provides about a third of the details.

Clicking on a topic (Buffalo Bill, or Europe) leads you to a screen showing additional information pulled from Wikipedia or Shelfari, and snippets of text from locations in the book where the topic appears. Let's explore Buffalo Bill (see Figure 4).

X-Ray's treatment of people

As you can see from Buffalo Bill's entry (Figure 4), X-Ray has some value in giving you background on a character and his role. And if the person has not appeared a lot in the text, it is relatively easy to read through and pick a location to read about him or her further in the text. The printed index has a heavy component of personal names, listing architects, planners, developers, relatives, and other important or interesting people who attended, had a part in, or were murdered during the fair.

William 'Buffalo Bill' Cody was one of the famous people involved, and the indexer included two sections related to him in the index: 'Buffalo Bill's Wild West Show' has 11

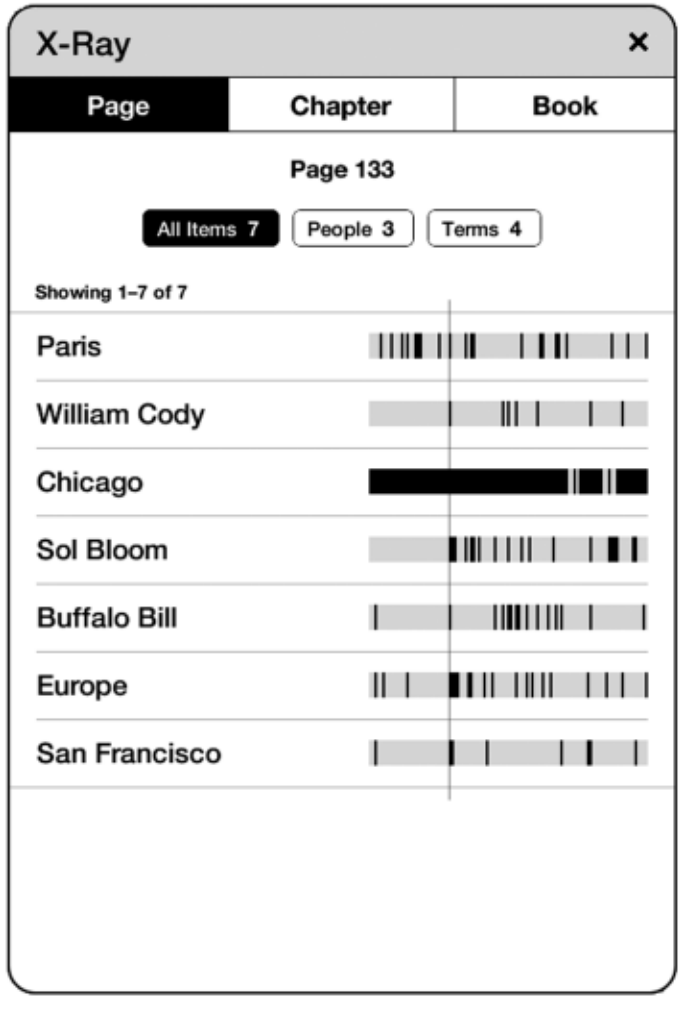

Figure 3 This screen shows all items filtered for just the current page

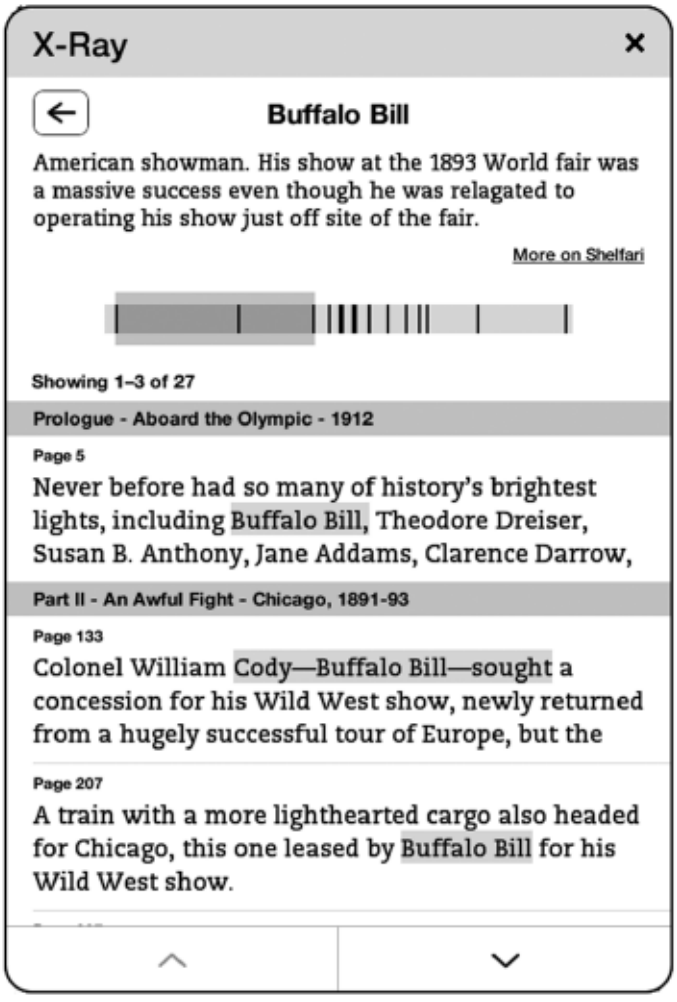

Figure 4 Buffalo Bill's entry in X-Ray, linked to Shelfari information and pulling it word for word. Shelfari's full entry is 'Colonel William Cody (Buffalo Bill)' but that information has not informed X-Ray.

locators, and 'Cody, William "Buffalo Bill"' has 13 locators. The indexer has kindly broken out the show's specific entries from Bill himself.

In X-Ray, when we search for Buffalo Bill, we see that the feature has failed to grasp that Buffalo Bill and William Cody are the same person. Re-examine Figure 3, and note that X-Ray has them listed separately. The Wild West Show does not appear at all in X-Ray. If we choose the X-Ray entry for William Cody, it is drawn from Wikipedia, and gives a much fuller description than the Shelfari one for Buffalo Bill (compare Figures 4 and 5).

X-Ray appears to be guilty of 'scattering' references. If you compare the sparklines, you can see you are getting differing locations.

Daniel Burnham, a major figure in the book, is thoroughly indexed in the print edition and has many X-Ray locators as well (547 instances). The indexer carefully noted his many accomplishments and events in the subheadings of the printed index, as well as distinguishing him from his son, Daniel Burnham. The distinction between father and son did not survive in X-Ray: there is only 'Daniel Burnham.'

Many people are referenced in X-Ray with only their first or last names, providing the reader with little or no information. In Figure 6, you can see the plight of a Mr McElroy, who appears only as 'McElroy' in X-Ray, and has no other information available. It sounds as though he is involved with a cannon, so surely he deserves a first name.

X-Ray's treatment of subjects

Like the Wild West Show, many subjects are ignored in X-Ray, and the subjects (or 'terms') that are included seem to have been developed by the Mad Hatter. 'Murder' does not appear, nor does 'serial murder' in any form. Architecture is not in X-Ray, although 'landscape architecture' made it in. Considering the egos of the architects involved, if they were alive, nasty notes would be in the mail. The Ferris wheel, which debuted at the fair, is not in X-Ray, although George Washington Gale Ferris is available. The Kindle screen however, cuts off his name so that the reader can see only the beginning: 'George Washington G . . .'. Any reader who did not know his first name would not pick this entry, and would have to resort to Search, or the entry for Mrs Ferris, no first name.

In the printed index, George G. W. Ferris, his wife Margaret, his Ferris Company, and the amazing wheel are all luxuriously detailed by the indexer.

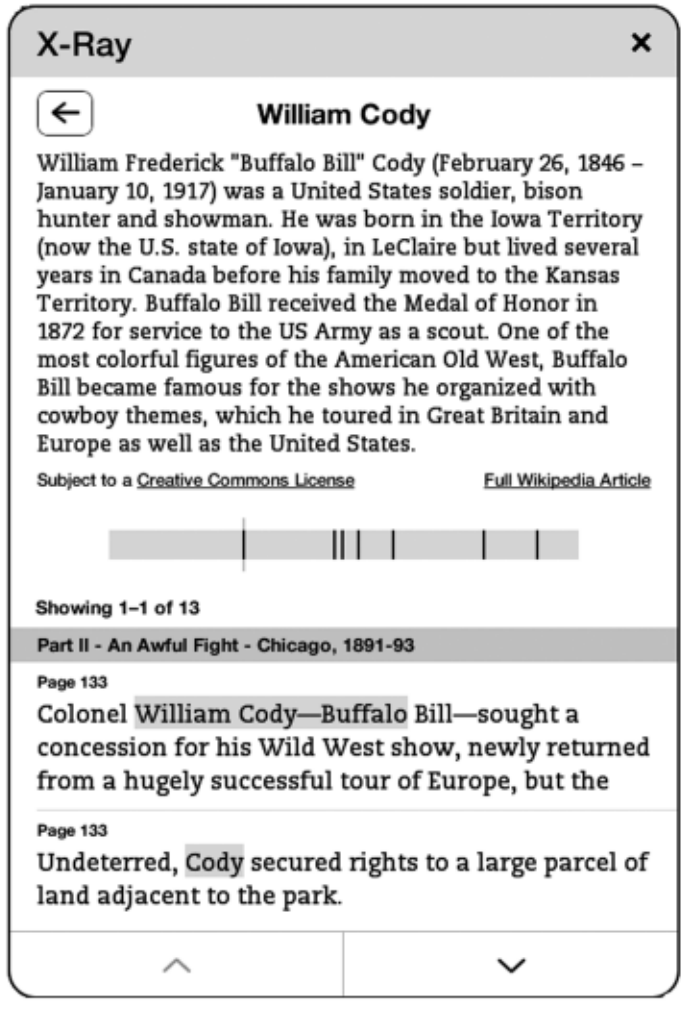

Figure 5 William Cody's fuller biography

Ferris Wheel, 208, 305, 373
in construction, 193-94, 218, 228, 236, 239, 240, 255, 280
final days of, 380-81
operation of, 279-81, 284, 287-88, 299, 300, 302, 327
romance on, 306-7
test runs of, 258-61, 269-73

What a lovely section! Any reader who has any romance in their soul will immediately turn to p. 306 with anticipation.

When filtered by 'terms' or subjects, instead of displaying the romance or the lurid instances of murder, X-Ray's version of the 'bones' of this book is very dry and uninviting (see Figures 7 and 8).

Jackson Park, the site of the fair, is not important enough to be in X-Ray, but 'belly dancers' and 'plate glass' are (Figure 8). Jackson Park was laid out by Frederick Law Olmsted and Calvert Vaux, the landscape architects who created New York's Central Park. The park has a large Wikipedia entry, and is quite important in landscape architecture's history. The indexer skipped the belly dancers, mentioned one location for the development of plate glass, and fully described Jackson Park:

Jackson Park, 35, 53-56
architects' visit to, 94-96

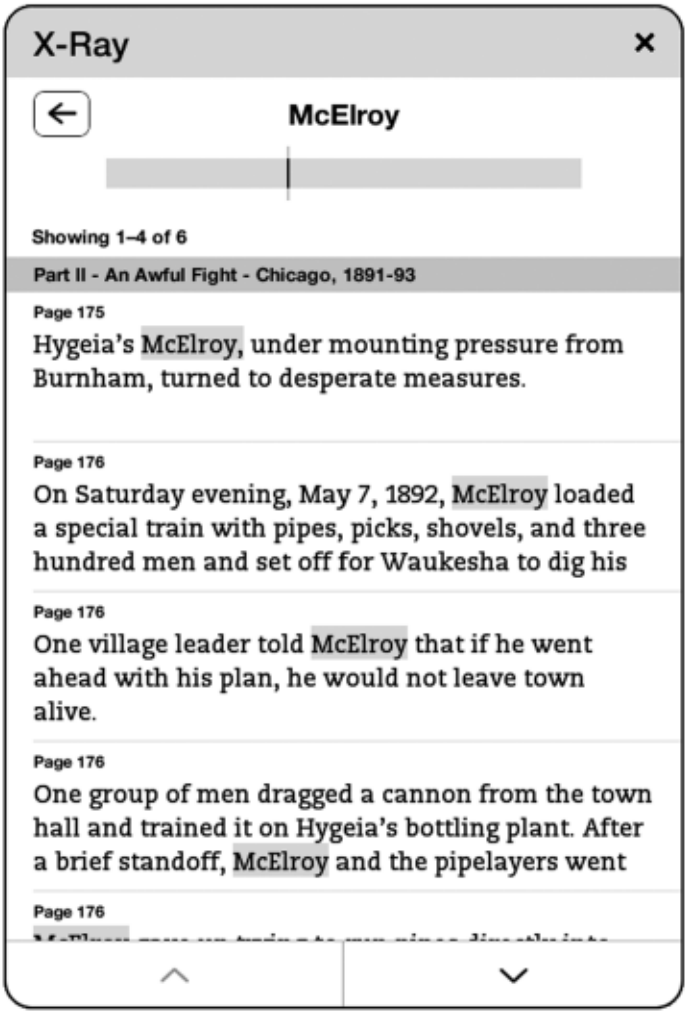

Figure 6 Note that poor Mr McElroy has no first name, no links to information about him, and no description

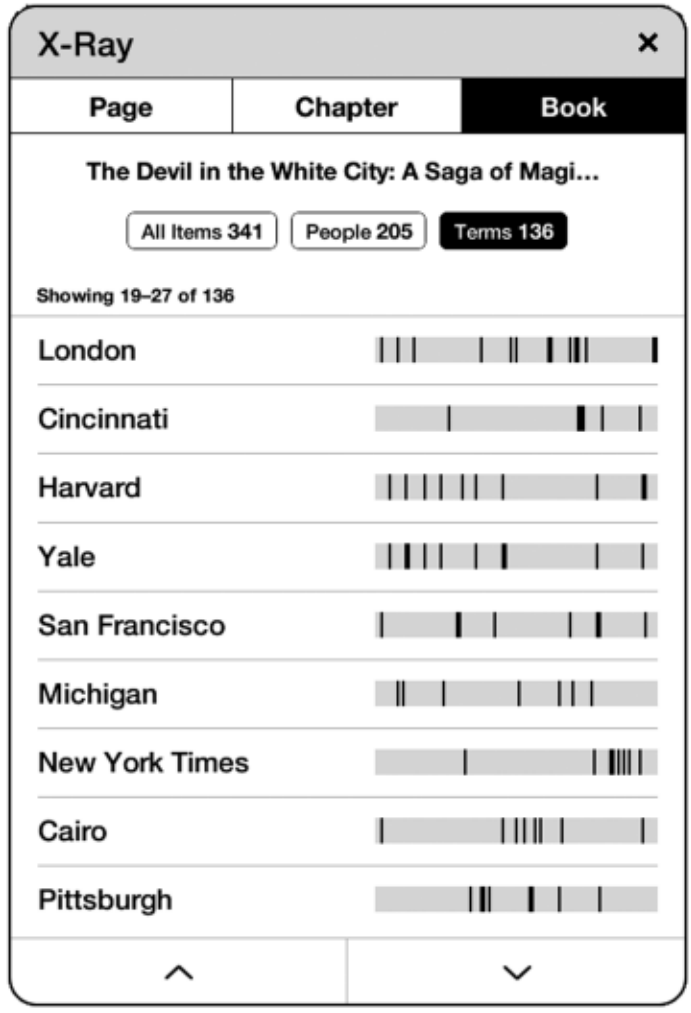

Figure 7 Most of the 'terms' appear to be state and city names

entrances to, 253
as fair site, 73-74, 82, 95-96; *see also* World's Columbian Exposition
permanent structures of, 375
plans for transformation of, 117-18
soil of, 129-30, 132, 144, 193

Again, the reader is probably going to be intrigued by 'soil,' and wonder why that was so important. Ah, landscape architecture! It is hard to immediately bring a white city into beautiful floral bloom if the soil is bad. If the soil is swampy, as Jackson Park was, buildings and heavy traffic could be a large concern. We can also see that Jackson Park has a relationship to the World's Columbian Exposition, important enough for a *see also* reference. An X-Ray-reliant reader could miss Jackson Park and its importance completely.

Shelfari's information

Amazon says that it is 'using its expertise in language processing and machine learning, access to significant storage and computing resources with Amazon S3 and EC2, and a deep library of book and character information. The vision is to have every important phrase in every book.' Thus far, we see no real deep access, nor every important phrase revealed.

Looking at the Shelfari[2] information for the book, it is

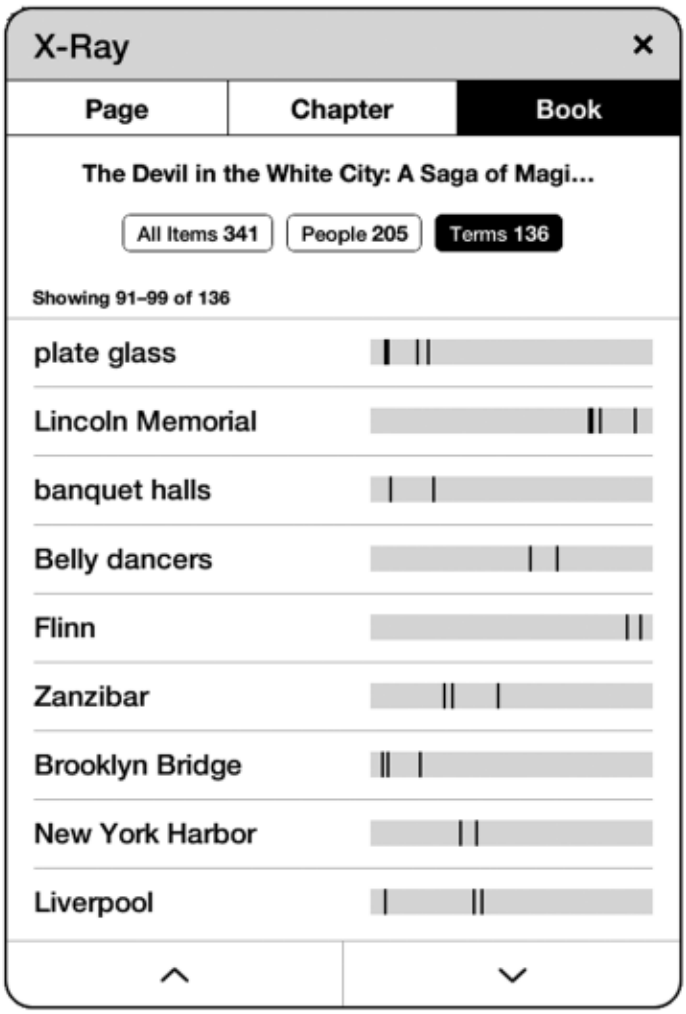

Figure 8 Plate glass and belly dancers do sound enchanting, but Flinn is actually a reference to a name included in the bibliography. Most of these items have very few occurrences. Every 'important phrase in every book' does not appear to be here.

clear that Shelfari knows that Buffalo Bill and William Cody are the same person, but that information is not informing X-Ray's topic choices. Shelfari's entry for the book makes no mention of the Ferris wheel, belly dancers, or plate glass, but does talk about 'swampy Jackson Park,' the Gilded Age, and murder. It seems as though X-Ray is not making full use of Shelfari's 'deep' knowledge for this particular book, and we have to wonder who chose to have the belly dancers make the cut, while leaving out the Ferris wheel. Shelfari also does not list Mr McElroy, so we have to wonder how he wound up in X-Ray.

Copyright, piracy, and file development?

The X-Ray feature has raised interesting issues beyond the reader's benefit or lack thereof. Who owns the X-Ray file: the author, Amazon, the publisher? Within days of its release, two hackers figured out how to create your own X-Ray file (sans sparkline graph) and download it to your Kindle. Will that be a violation of some form of copyright? Will Kindle remove these handmade files? Or replace them when they develop their own for a book?

Going forward, who is responsible for ensuring that a book has a good X-Ray file? Again, the author, the publisher, Amazon, or an entrepreneur who might develop them for sale on the web? This is all unclear ground.

A small ten-minute experiment reveals the details of the file. Having decided that Mr McElroy deserved to have at least some information, I edited a copy of the .asc file containing the data (Figures 9 and 10).

After I replaced the .asc file on my Kindle Touch, and reloaded the book, Mr McElroy became a full-fledged character with information (Figure 11, overleaf).

Ramifications for indexers

I was prepared to enjoy the X-Ray feature, especially while reading Steve Jobs' biography and playing with the feature thoroughly in the process. In the Jobs book, no characters went without their full names, and there were many more meaningful term entries. X-Ray also helpfully offered up search results from the index, allowing readers to see that there was indeed an index, that it was hyperlinked and that it was ready for use. For high-profile books, it is obvious that care has been taken to make the X-Ray experience a good one. And it could be good – it is easy to edit these files. But with lower-profile books, or older bestsellers like *Devil,* care

{"type":"character","term":"McElroy","desc":"","descSrc":"wiki","descUrl":"http://en.wikipedia.org/wiki/McElroy","locs":[[438060,92,16,8],[438233,184,34,7],[438795,99,24,7],[439045,247,199,7],[439553,196,0,7],[439891,123,101,7]]},

Figure 9 Raw X-Ray data for the McElroy entry before editing. We can see the term, that he is a character, and his locations. The Wiki information leads to a general search in Wikipedia for the word McElroy.

{"type":"character","term":"J. E. McElroy","desc":"J. E. McElroy, an entrepreneur heading up the Hygeia Mineral Springs Company. Hygeia contracted to pipe water to the fair.","descSrc":"wiki","descUrl":"http://en.wikipedia.org/wiki/McElroy","locs":[[438060,92,16,8],[438233,184,34,7],[438795,99,24,7],[439045,247,199,7],[439553,196,0,7],[439891,123,101,7]]},

Figure 10 I have filled out his initials, and given a brief description of his role in the book

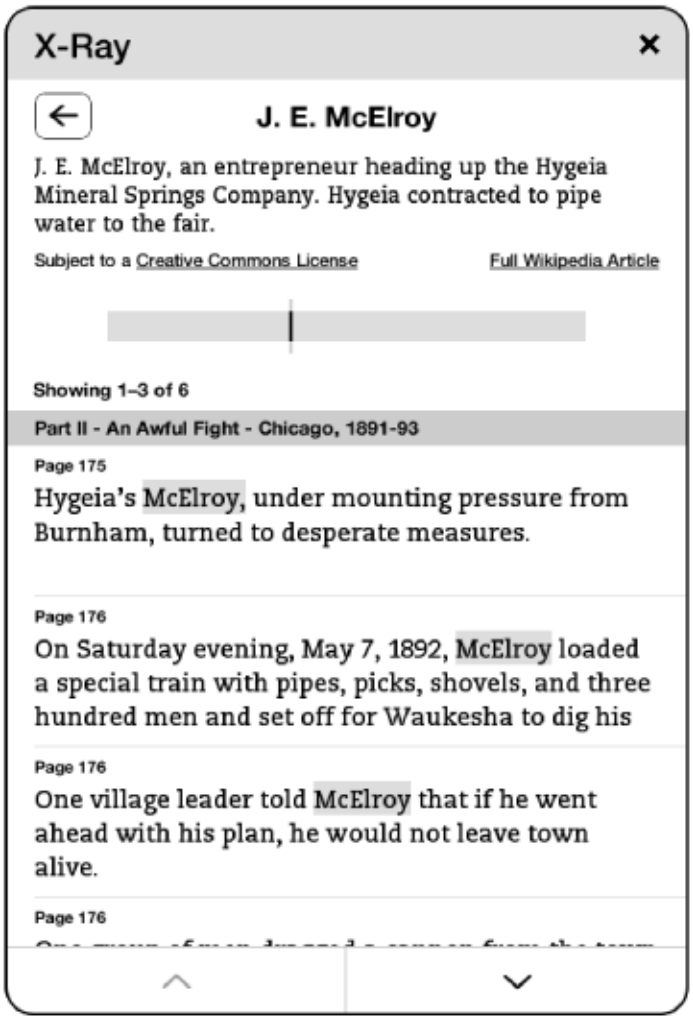

Figure 11 Now J. E. has come alive. So have, mysteriously, links to Creative Commons and Wikipedia. Unfortunately, the Wikipedia link leads still to a generic McElroy search: further editing could resolve this issue. The Creative Commons link leads to general terms and conditions for Creative Commons materials, and nothing about McElroy himself. I can only hope I have not violated something by editing this entry.

is not being taken, and this could leave the reader with far fewer navigational aids. Leaving the index out, and relying solely on Search and X-Ray, gives the reader a substandard experience for *Devil in the white city*. Since the print editions had an index, it is hard to imagine why it was not used to inform X-Ray choices for topics. Time and money may be the devils here.

Amazon Kindles are closed systems, using their own specified formats for ebooks, .mobi and .azw. The coolness factor of X-Ray may be appreciated and used at first by readers, but unless the details get better for all books, readers will be missing some of the picture. Indexes are one of the best navigation tools for a book, print or not: not only does the reader get directly to the piece of information needed, they also may learn a bit, find out differences and details, and get lured into other locations by romance and drama.

If the majority of X-Ray files are not being evaluated by humans before being shipped, it is likely that many books will suffer from lack of X-Ray detail. One promising thought is that Amazon might wake up to see that humans, perhaps indexers, could make this feature work a lot better, and offer them some work. Publishers might want to ensure that their book is shipped with a good X-Ray file, and not rely on one as poorly developed as this one. In both cases, opportunities for indexers lurk. If not, there are other ebook formats for us to work with, ones that will support indexes.

Throughout this spring of 2012, indexers will continue to be working with the International Digital Publishing Forum on the inclusion of real indexes in the Epub 3.0 ebook format. Some of the features we hope to incorporate are a pop-up index available to the reader at any point in the book, with an area to type in their search, and scrolling features to aid discovery. If implemented as outlined in the Charter Document (see Epub 2011a, 2011b), there could also be a reverse index, showing all the entries for a range of text, much like the X-Ray filter set to Page. But unlike X-Ray, these entries will be index entries, and they will distinguish between Mr Burnham and his son, and they will know that Buffalo Bill is William Cody. Mr McElroy will have his initials. And lastly, they will be sure to include the Ferris wheel and murder in all its lurid details.

If the devil is indeed in the details, X-Ray needs to get much better or it will be bedeviled.

Acknowledgments

Permission to use text from *The devil in the white city* was kindly granted by Erik Larson. Thanks also to Kate Mertes, indexer for the paperback edition of *Devil.*

Notes

1 From Amazon's explanation of its new X-Ray functionality: www.amazon.com/Kindle-Touch-e-Reader-Touch-Screen-Wi-Fi-Special-Offers/dp/B005890G8Y/ref=amb_link_359054382_3?pf_rd_m=ATVPDKIKX0DER&pf_rd_s=center-1&pf_rd_r=17F1Q2TRY2AV5QEY1Y68&pf_rd_t=101&pf_rd_p=1337548602&pf_rd_i=507846#xray

2 Shelfari page on *The devil in the white city*: www.shelfari.com/books/10065/The-Devil-in-the-White-City

References

Epub (2011a) Charter document for indexes in ePub 3.0. Available at: https://code.google.com/p/epub-revision/wiki/IndexesCharterProposal

Epub (2011b) Main page for Indexes in ePub 3.0. Available at: https://code.google.com/p/epub-revision/wiki/IndexesMainPage

Larson, E. (2003) *The devil in the white city: murder, magic, and madness at the fair that changed America.* New York: Crown.

Jan Wright has been indexing and taxonomizing since 1991, and has many years of experience in software documentation, help systems, and embedded indexing. An ASI member, she has won several awards from the Society for Technical Communication, and in 2009 won the ASI/H.W. Wilson Award for her index to Real World InDesign CS3*, the first time a technical manual had won. She is currently on the IDPF Indexes Working Group Charter team, and is co-chair of ASI's Digital Trends Task Force.*
Email: jancw@wrightinformation.com