
One of Michael Lesk’s opening comments is that Google was a “spinoff” of a digital library project and is now responsible for over half of all Web searches, at an average rate of more than 250 million searches per day (pp. xxvii–xxviii). This comment is indicative of the book as a whole, in a number of ways. For one thing, it is attention getting and engages the reader, and that is true of the book as a whole. It is a broad-based book that surveys digital library technologies as well as nontechnical issues, and in the process it manages to draw in the reader, perhaps even engendering a desire to join the author in a friendly discussion of one point or another.

It is worth noting that this book, the second edition, contains substantial updates to the first edition, published in 1997 under the title Practical Digital Libraries: Books, Bytes and Bucks. In the author’s own summary, the changes largely involve the impact of the Web and the newer digital library projects, although he also points out that at least half the lines in the book changed between editions. Certainly there is a lot of new material here. Although the chapter titles have stayed pretty much the same, there are two new chapters (one on user needs and one on the scope of digital libraries). Other indications of the extent of new material are that the second edition weighs in at 424 pages compared to 297 in the first edition and at that least half the references in the second edition are dated after 1997.

The goal of the book is to provide details of how things are done, or could be done, with respect to digital libraries. The first half of the book is the technology survey, and it indeed delves into details of digitization (for text, audio, images, and video), digital storage, and of searching that storage, retrieving information, and viewing digital displays. In addition, topics such as knowledge classification schemes, markup languages, metadata, hyperlinks, security, and cryptography are all presented and discussed. It could be overwhelming (how much detail do you really want to read about hash tables or regular expressions?), but Lesk keeps it interesting with examples and asides, such as the comment that “searching [online] is still often hit-or-miss” (p. 186).

The second half of the book discusses the development of digital libraries in a more diverse way. Several of the chapters cover questions in the areas of economics and legal constraints, several have sections that seem to be more of the technical detail variety (e.g., discussing performance measures or preservation of digital materials), one chapter surveys current digital library projects, and the concluding chapters look to the future. As Lesk acknowledges, on the technical side a lot is known about how to do things, for the most part, but on the broader, less technical side there are still a many unknowns. The chapters in the second half tend to end in admissions of lack of answers. Chapter 8’s summary says, “The systematic study of user needs is relatively little explored.” Chapter 9’s summary begins, “We do not yet have collection principles for a digital library.” The next chapter concludes with “Economics is emphatically not a solved problem.” And the chapter after that, on intellectual property rights, echoes the same sentiment: “This chapter has described another unsolved problem.”

Despite the understandable lack of answers, Lesk does provide considerable detail along with supporting data to flesh out most, if not all, of the major issues. When I mentioned the Google reference as indicative of the book as whole, another reason was that the comment contains some of the numbers behind the statement of Google’s popularity. The prevalence of supporting data throughout the book is a reflection of the author’s desire to supply practical information based on solid research.

There are several other reasons why I chose the Google comment as indicative. One is the association of a Web search engine with the concept of digital libraries, which leads to the general question of the relationship between the two. At various points within the book, the reader might get the distinct impression that when Lesk writes digital library he is in fact referring to the Web in general. In a few places, he refers to digital libraries as specific organized collections available on the Web. That implies that by definition there is no digital library without a Web presence, although just a few pages later he briefly mentions local digital libraries on compact disk–read only memory (CD-ROM). In other contexts he does talk about local digital storage of library material, but he does not seem to give much attention to any arrangement that is not on the Web.

In fact, this book does not address the possible definitions of digital library head-on, as other books (e.g., Chowdhury & Chowdhury, 2003) do. Ignoring this problem has the advantage of not bogging the reader down in balancing multiple definitions that might have conflicting implications. The disadvantage is that if the reader is not careful, it is easy to miss Lesk’s definition of a digital library. And sometimes Lesk himself seems to take a rather loose approach to what constitutes a digital library. For example, in the chapter on the scope of digital libraries, he opines that digital libraries will include not just published material but “email blogs... and everything else that turns up” (p. 372).

Returning to the Google connection, the remark in its own small way is an example of how Lesk’s style not only catches your attention but also perhaps encourages you to want to know more. In this case, for example, I am curious just how much, or how little, the founders of Google attribute to the fact that they received some initial funding through the Stanford Integrated Digital Library Project. After all, Google’s own Web site goes into some detail in describing the company’s history and sources of funding and never mentions a connection to the Digital Library Project (Google, 2005).

Of course, in a book of this scope there will always be room for someone to question why some topics were included and others minimized or omitted, or whether more or less explanation would be desirable. In the spirit of an open dialogue that I think the book encourages, I would like mention a few of the topics that, in my opinion, might have been presented differently.

One such topic is the book’s treatment of XML. The chapter discussing text formats and higher-level descriptive systems states that there are three standards worth discussing, namely, MARC, SGML, and hypertext markup language (HTML). There is no mention of XML at this point, even though XML, begun in 1996, is a descendant of SGML and has been deemed “one of the most important developments in document syntax in the history of computing” (Harold & Means, 2002). Instead, the topic is introduced later as part of a short discussion of the Semantic Web, with a conclusion that almost seems to discount both. I would think XML (and its many extensions) merits treatment as a solid topic in its own right because of its success in a wide variety of information processing applications.
In the area of searching and search technology, the book addresses the major issues, but there was at least one aspect for which more detail might have been provided: the limits of current search engines with respect to full-text searching. I do not mean concept searching, which is addressed, but the extent to which search engines actually even try to index all of a document. For example, do Web search engines index the entire content of Web-accessible portable document format (PDF) documents? A recent intriguing study of Google’s indexing of PDF documents suggests that Google stops indexing at some cut-off point, and in this case it was around the 62nd page of a 216-page document (Steward, 2005). In another test in the same study, when all the pages were accessible as individual PDF documents as well as individual HTML pages, Google still did not index all of the PDF pages.

As an aside, my usage of a Web reference in the previous paragraph exemplifies another point from Lesk’s book, and that is the quality issue. I do not know whether the experiment cited was peer-reviewed, nor whether the site providing that page exerts any quality control. All I know is that the site is a popular one for PDF information, and the author of the study is the one who maintains the PDF forum.

There are incredibly enormous benefits that follow from the ability to have full-text searching of digital materials available on the Web, as the book demonstrates. On the other hand, if one’s goal is to read a lengthy text such as a book, as opposed to searching through it, an electronic version is not very comfortable reading for many people. Lesk admits that “among the most important questions for digital libraries are those related to the willingness of people to read information from screens instead of from paper” (p. 189). But he obviously believes that, willingly or not, more people are reading more material on electronic screens, and concludes that “we can expect a major shift towards digital reading” (p. 375).

In another section pertinent to the issue of reading electronic screens, the author says that online reading of whole books has not been popular, and even the attempt to put e-books on handheld devices, on the assumption that portability might help, has been in general a failure. He posits a number of potential reasons why, without choosing one or another as key. In a possible nod to the fact that paper versions may indeed have some good points, he presents the technologist’s belief that “someday, we should have devices that are comparable in weight and readability to a sheet of paper” (p. 173). Later, Lesk suggests that paper versions of publications will continue to exist but electronic versions will become dominant, similarly to the relationship between, say, the recorded music industry and live music performances.

As a minor note on this issue, he mentions, in the context of e-books, that the reason behind Microsoft’s introduction of the Tablet personal computer (PC) was to see whether the tablet size improved readability. Actually, tablet computers (often called pen computers) had been around for nearly a dozen years, used by field workers and other mobile professionals in lieu of filling out paper forms. Microsoft introduced its Tablet PC operating system to tap into this market, and its primary distinguishing feature at that time was Microsoft’s version of “digital ink” as a data type with its own properties, and an inherent part of the operating system (see, for example, Walker, 2002). The digital ink data type was used for the purposes of note taking, writing, drawing, and the like. It is true that Microsoft also offered a free program for Tablet PCs for reading e-books that contained Microsoft’s ClearType technology for smoothing characters on liquid crystal displays (LCDs) to improve ease of reading. But this feature was not, I believe, the major impetus behind the Tablet PC operating system.

Finally, returning once again to the Google comment, there is one more reason why it is indicative, and that is the fact that Lesk mentions the Google connection at least three times in the book. This repetition is representative of other tidbits that are similarly repeated a number of times. I do not mean this as a criticism: Some repetition is inevitable because of the interrelated nature of the topics. It just happens to be something I think readers would notice. In some cases, the author adds “as previously mentioned,” and in other cases he does not, and in some cases a repeated point is slightly reworded.

One drawback to a little repetition, especially if the basic idea has been reworded, is that in some cases two similar statements appear to be in conflict. For example, at one point, Lesk says that the Web is about 170 terabytes of text, and the deep Web is “perhaps 400 times more” (p. 228). Later, he says the Web in 2003 was 150 terabytes, and in the “dark Web” there is “perhaps 10 times as much” (p. 287). The deep Web or dark Web caught my attention. Are the estimates of 10 times or 400 times simply from different sources or from different times? Are they based on different definitions, or does the large difference really mean that nobody knows with any degree of certainty? Or perhaps one of the two is a typographical error?

The bottom line is that Michael Lesk has pulled together a great deal of material to help the reader understand the many aspects of digital libraries. He presents evidence and conclusions, and alternative viewpoints when applicable, in an engaging manner. Reflective of the evolving nature of digital libraries and the unknowns involved, he poses thoughtful questions when there are no clear answers. One may wish to debate an item or two, but that is more a mark of an intelligent presentation than a criticism of the content.

Richard H. Veith
R. Veith Consulting
Port Murray, NJ 07865
E-mail: rveith@gti.net

Published online 00 XXXXXX 2006 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/asi.20358

References


en_corporate/history.html


AUTHOR QUERIES

AQ1: Please spell out the words in this paragraph that are acronyms.
AQ2: Personal computer correct? Capitalized as a product name?
AQ3: Is there a title for this work?
AQ4: Quotation marks in original title?