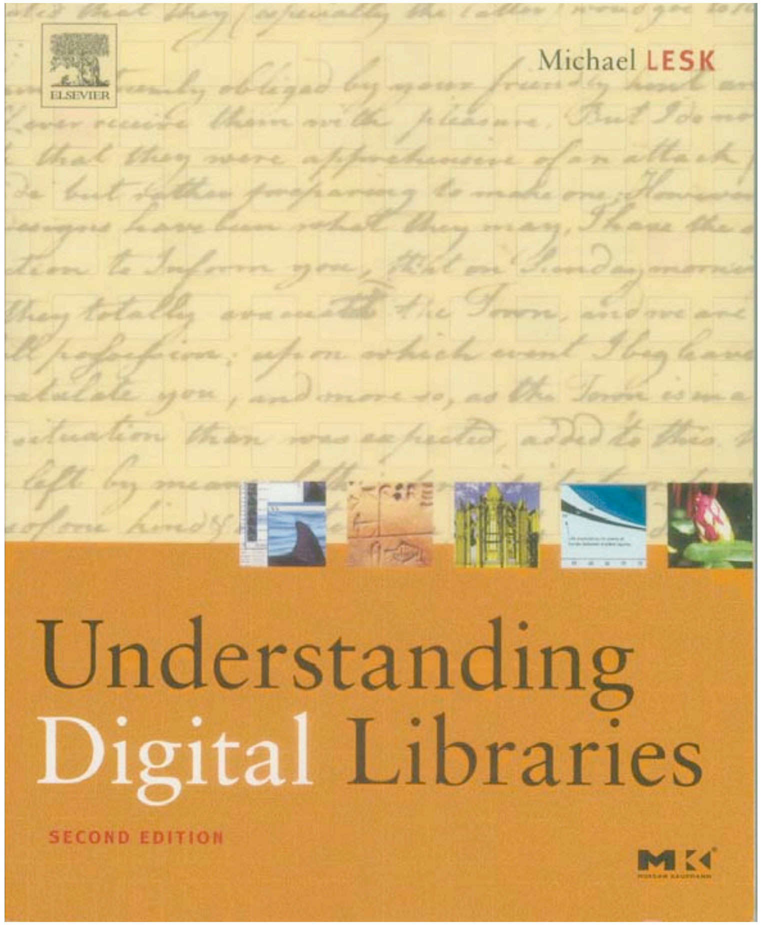
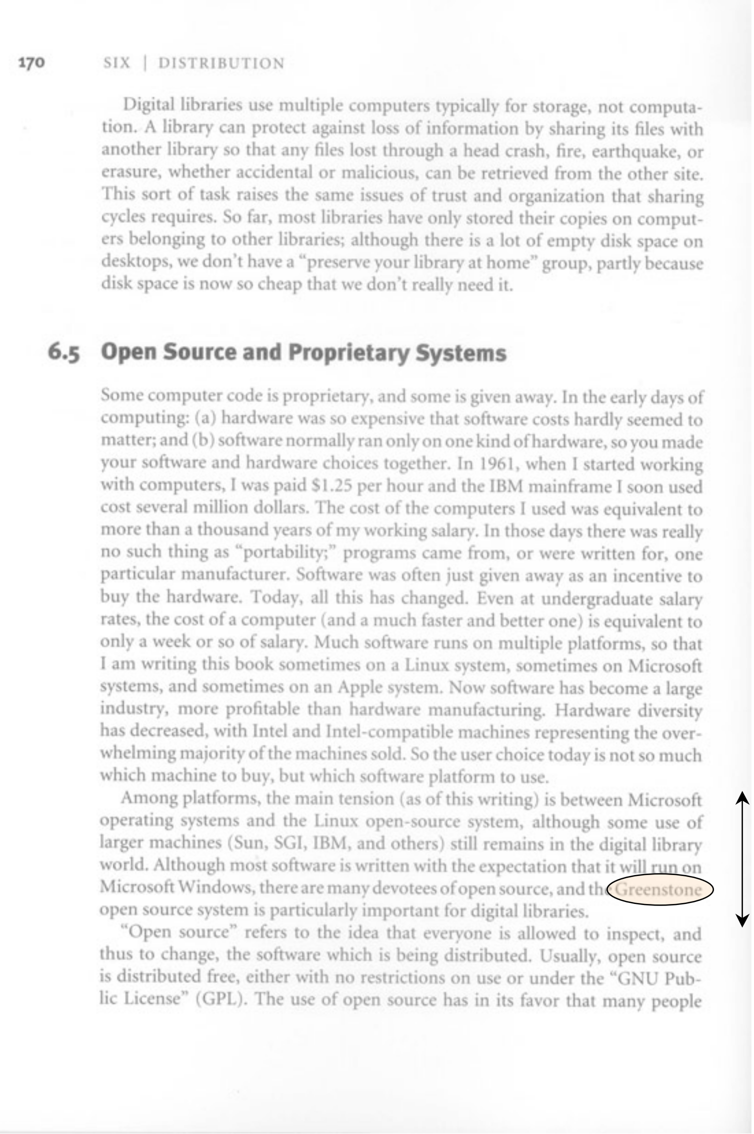
Endorsement by Mike Lesk—perhaps the leading U.S. guru in digital libraries



Understanding Digital Libraries: Second Edition, Morgan Kaufmann, 2005



170 SIX | DISTRIBUTION

Digital libraries use multiple computers typically for storage, not computation. A library can protect against loss of information by sharing its files with another library so that any files lost through a head crash, fire, earthquake, or erasure, whether accidental or malicious, can be retrieved from the other site. This sort of task raises the same issues of trust and organization that sharing cycles requires. So far, most libraries have only stored their copies on computers belonging to other libraries; although there is a lot of empty disk space on desktops, we don't have a "preserve your library at home" group, partly because disk space is now so cheap that we don't really need it.

6.5 Open Source and Proprietary Systems

Some computer code is proprietary, and some is given away. In the early days of computing: (a) hardware was so expensive that software costs hardly seemed to matter; and (b) software normally ran only on one kind of hardware, so you made your software and hardware choices together. In 1961, when I started working with computers, I was paid $1.25 per hour and the IBM mainframe I soon used cost several million dollars. The cost of the computers I used was equivalent to more than a thousand years of my working salary. In those days there was really no such thing as "portability;" programs came from, or were written for, one particular manufacturer. Software was often just given away as an incentive to buy the hardware. Today, all this has changed. Even at undergraduate salary rates, the cost of a computer (and a much faster and better one) is equivalent to only a week or so of salary. Much software runs on multiple platforms, so that I am writing this book sometimes on a Linux system, sometimes on Microsoft systems, and sometimes on an Apple system. Now software has become a large industry, more profitable than hardware manufacturing. Hardware diversity has decreased, with Intel and Intel-compatible machines representing the overwhelming majority of the machines sold. So the user choice today is not so much which machine to buy, but which software platform to use.

Among platforms, the main tension (as of this writing) is between Microsoft operating systems and the Linux open-source system, although some use of larger machines (Sun, SGI, IBM, and others) still remains in the digital library world. Although most software is written with the expectation that it will run on Microsoft Windows, there are many devotees of open source, and the Greenstone open source system is particularly important for digital libraries.

"Open source" refers to the idea that everyone is allowed to inspect, and thus to change, the software which is being distributed. Usually, open source is distributed free, either with no restrictions on use or under the "GNU Public License" (GPL). The use of open source has in its favor that many people contribute to and improve it, you can assure yourself of what it does and doesn't do, and of course it is free and not constrained by complex contracts. Disadvantages are that you may have to do your own support, not as many other people use it, and it changes more often and in unpredictable ways.

The best known open-source system is the Linux operating system, written originally by Linus Torvalds based on the design of Unix, and now maintained by a large community (although Torvalds is still the leader). Linux, effectively, competes with both Microsoft Windows and with various flavors of Unix (including other free versions such as FreeBSD). Nobody knows how many users of Linux there are: you can download it free and you don't have to report to anybody that you have it. Some estimates are that under 10% of servers are now Linux-based; others indicate that 30–40% or more of new servers are Linux (Gulker 2003, and Ewalt 2001). However, only about 1% of the machines accessing Google identify themselves as Linux, whereas about 90% say they are a variety of Microsoft Windows. Microsoft software, which comes bundled with almost all PCs sold, still dominates the end-user market. Compared to Windows, Linux users argue that their platform is more flexibile, less likely to crash, less vulnerable to viruses, and offers greater power and control to the users. Support is available from companies such as RedHat, and a lot of device-controllers and software are available for Linux. Microsoft would counter that far more software and device-controllers are available for Windows, and of course support for Windows is much more organized and well known. Perhaps the best evidence that Microsoft fears Linux, however, is that they helped fund a lawsuit by SCO which threatened to interfere with the sale and use of Linux (alleging that Linux contained lines of copyrighted code now belonging to SCO through a set of purchases of the original Unix code from AT&T).

Perhaps more important to the digital library community is the Greenstone open source package, available at `www.greenstone.org` in multiple languages and for multiple platforms. Like Linux, Greenstone code is open and available for inspection or use without charge. The owners of Greenstone do not charge per-seat fees, impose complex procedures for using their code to be sure that you are not exceeding the number of licenses you have, or engage in any of the other somewhat constraining activities which software companies feel they must do as a way of reducing software piracy. Greenstone was originally written by the University of Waikato in New Zealand; the leader of the project is Ian Witten.

Greenstone is distibuted under the GPL (GNU Public License), which basically says that you can use it freely, but if you redistribute it you must give others the right to redistribute the code you are sending them. The intent of GPL is to prevent a situation in which companies take open source code and resell it with limitations on the use and further distribution of the code. GPL has been around since 1991 and has been successfully used by a large number of projects.

Greenstone provides many facilities that a digital library would need. For details, you should read Witten's book (Witten and Bainbridge, 2003). To summarize, however, Greenstone enables users to build digital library collections and make them accessible to users, either locally on CD-ROM or over the Web. It includes text search, image display, hierarchical browsing, fielded data, and many other capabilities. Many projects around the world are using it. As with all open source projects, you can make whatever changes you want, you can find out exactly what the software does, and you will not be hassled about exactly how many people you have using it.

The alternatives to Greenstone as a way of distributing data are likely to be commercial database systems, not specific digital library systems. There really isn't any commercial software sold only for the purpose of supporting digital libraries, although a variety of data base packages can be used, and some library OPAC (online public access catalog) systems can be generalized to include full text. There are some specialized systems; for example, Olive Software is a leader in the problems relating to digital versions of historical newspapers. Perhaps the most significant, albeit very recent, commercial alternative is IBM DB2 Content Manager. For example, in June 2003, the Australian Broadcasting Corporation agreed to a $100 M deal with IBM to use DB2 Content Manager to store 100,000 old tapes of broadcast programs (see also Meserve, 2003).

Given the advantages of free software and unrestricted use, why hasn't open source spread more rapidly? One answer is simply the lack of advertising; Microsoft has recently announced that there will be a $150 M campaign for the 2003 version of Microsoft Office software. Nobody puts anything like the same effort into persuading people to use Linux, or OpenOffice, or Greenstone. Libraries planning to use open source also generally have to have a slightly higher level of technical sophistication, even with the advent of companies like RedHat. As with indexing, the more power you have, the more opportunity you have to dig your own hole and fall into it. Nevertheless, for many of us, falling into a hole we dug ourselves is less threatening than running into somebody else's brick wall, because you are more able to fix the situation by yourself.

6.6 Handheld Devices

In addition to networked computing, information can be distributed by putting it on special-purpose handheld devices. During 2000 there was a brief flurry of interest in the "e-book," the idea that people would read full books on special purpose machines. Online reading of whole books had not been popular, and