Orientation: The world of digital libraries

Digital libraries are changing the world.

Example One: Supporting Human Development

Kataayi is a grassroots cooperative organization based in the village of Kakunyu in rural Uganda. In recent years its enterprising members have built ferro-cement rainwater catchment tanks, utilized renewable energy technologies such as solar, wind, and biogas, and established a local industry making clay roofing tiles—among many other projects. But amid such human resourcefulness, information resources are scarce. The nearest public phone, fax, library, newspapers, and periodicals are found in the district town, Masaka, 20 km distant over rough roads. Masaka boasts no e-mail or Internet access. The difficulty of getting there effectively discourages local inhabitants from taking advantage of the information and communication technologies that we take for granted in developed countries.

The Kataayi community believe that an information and communication center will have a major development impact in their area. They laid the groundwork by acquiring a computer and solar power generation equipment. They established an e-mail connection via cellular phone and set up a computer training program. They constructed a brick building to house the center (Figure 1.1). And they gathered several books. But they need more information resources—lots more. They were looking for books covering such topics as practical technology, fair-trade marketing, agriculture, environmental conservation, spirituality, and social justice issues.

Then they discovered digital libraries. The Humanity Development Library is a compendium of some 1,200 authoritative books and periodicals on just such topics, produced by many disparate organizations—UN agencies and other international organizations. In print, these books would weigh 340 kg, cost $20,000, and occupy a small library bookstack. Instead, the collection takes the form of a digital library and is distributed on a single CD-ROM throughout the developing world at essentially no cost. Related digital library collections cover topics including disaster relief, agriculture, the environment, medicine and health, food and nutrition; more are coming. These digital libraries will increase Kataayi’s information resources immeasurably, at a minuscule fraction of the cost of paper books.
Example Two: Pushing on the Frontiers of Science

Leave this local community and enter a very different one that operates internationally and on a far larger scale. For over a decade, physicists have been using automated archives to disseminate the results of their research. The first archive, in high-energy physics, began in the early 1990s. It targeted a tiny group of fewer than 200 physicists working on a particular set of research problems, who wanted to communicate their progress. Within months the clientele had grown fivefold. Numerous other physics databases sprang into existence. Within a few years these archives served tens of thousands of researchers; by the year 2000 they processed 150,000 requests per day and the number of papers passed the half-million mark in 2008.

The physics archival digital libraries are entirely automated. To submit a research paper, contributors fill out title, author, and abstract on an electronic form and transmit the full text of the paper. Upon receipt, which is instantaneous, the paper immediately and automatically becomes part of the archive, permanently accessible to others. The contributions are not reviewed or moderated in any way, except for a quick scan to ensure that they are relevant to the discipline. The chaff is simply ignored by the community. The upshot is that research results are communicated on a dramatically accelerated timescale, and the expense and waste of hard-copy distribution is eliminated.

For some areas of physics, online archives have already become the dominant means of communicating research progress. Many people believe that the scheme has effectively replaced commercial publication as a way of conveying both topical and archival research information. Why don’t researchers in every discipline follow suit? Soon, perhaps, they will. Proponents of these online archives forecast the imminent demise of commercially published research journals and believe that communicating research results using “chemicals adsorbed onto sliced processed dead trees” will rapidly become a quaint anachronism. On the other hand, many disagree: they argue that peer review is still highly valued in most scientific disciplines, and that even in the small, specialized communities where they are used, online archives augment rather than replace peer-reviewed journals.
Example Three: Preserving a Traditional Culture

The physics archive is centered at the Los Alamos National Laboratory in New Mexico. Only 58 km away as the crow flies, but light-years distant in other respects, is the Zia Pueblo, home of one of a score of Native American tribes in New Mexico (Figure 1.2). By 1900 the population had fallen to less than 100, and the tribe was expected to die out during the 20th century. With the return of some land, and medicine and education provided by U.S. government programs, fortunes have improved and the people now number 600. But today the Zia Pueblo face a major problem: the loss of their language and traditional culture. Young people are not learning the Zia Pueblo traditions or Keresan, its language. This is a common complaint in traditional societies, overexposed as we all are to the deafening voice of popular commercial culture blaring everywhere from television, radio, and advertising billboards.

To preserve the Zia language and traditions, a digital library has been proposed. It will include an oral history compilation, with interviews of tribal elders conducted in their native language. It will include an anthology of traditional songs, with audio recordings, musical scores transcribed from them, and lyrics translated by a native speaker. It will include video recordings of tribal members performing Pueblo dances and ceremonies, along with a synopsis describing each ceremony and a transcription and translation of the recorded audio. The goal is to produce a multimedia artifact, the purpose of which is not so much to archive the material as to make it publicly available and to involve members of the tribe in collecting and disseminating it.

Figure 1.2: The Zia Pueblo village
Example Four: Exploring Popular Music

Turn from this small, esoteric group in New Mexico to the wide-ranging, disorganized, eclectic panoply of music that is played in the Western world today. In all human societies, music is an expression of popular culture. Different generations identify strongly with different musical styles and artists. People’s taste in music reflects their personality and sense of identity: teenagers, in particular, regard their musical preferences as being strongly bound up with who they are. Music is a medium that is both popular and international. Pop music culture transcends social boundaries, be they national borders or socioeconomic groupings. Yet music also exhibits strong cultural diversity: folk music is specific to a particular country or region, and different styles characterize local ethnic groupings.

Imagine a digital music library that reflects popular taste, a library that people from all walks of life will want to use. From an immense music collection you can retrieve tunes in many ways: by humming a theme, by recalling words from the title or lyrics, by giving the composer’s name, or you can specify any combination of these. Flexible browsing facilities allow you to meander through the collection, listening to tunes rendered by a synthesizer, or indeed to live recordings. Almost any song you can think of is there, often in dozens of different versions.

Experimental versions of such libraries already exist. A huge volume of musical material is already on the Web in the form of MIDI files, the musical representation used by synthesizers. It is easy to locate and download hundreds of thousands of files covering a wide range of styles, from classical symphonies to current pop songs, from jazz classics to ethnic folk songs. In a very real sense, these reflect popular taste, comprising whatever people have decided to spend their time entering. You will find a score of versions of The Beatles’ Yellow Submarine and Bach’s Air on a G-string. All this music can be indexed by automatically identifying melodic themes and extracting text containing the title, artist, composer, and lyrics. Contentious copyright issues can be avoided by leaving all source material on its home site: what the library provides is not a repository but a catalog and means of access. And the Web is a prolific source of other musical resources, from record stores to guitar tablatures for popular tunes. Having found a tune, you can listen to samples of recordings by different artists, obtain a CD, watch a video, or buy sheet music.

The scope of digital libraries

These four examples, at different stages of development and deployment, hint at the immense range of digital libraries. From the perspective of ordinary people, libraries often seem scholarly and esoteric. But they are not necessarily so. Practical topics are of interest to practical people like Kataayi’s members. Academic libraries have as their purpose research and education: high-energy physicists already base their research activity on electronic document collections. Digital libraries offer unique ways of recording, preserving, and propagating culture in multimedia form. Collections that reflect popular taste in music (or film, or TV) have already become mass-market consumer products, with delivery to teenagers on miniature, mobile, Web-capable, pocket devices.

An application that makes a sustained market for a promising but underutilized technology is often called a killer app. The term was coined in the mid-1980s for the Lotus spreadsheet, then the major
driving force behind the business market for IBM PCs. (VisiCalc had previously played a similar role in the success of the Apple II.) The World Wide Web is often described as the Internet’s killer app. The killer app for digital libraries may well be music collections; in turn, Section 1.4 shows that as far as the developing world is concerned, digital libraries themselves may be killer apps for computer technology.

1.1 Libraries and Digital Libraries

Is a digital library an institution or a piece of technology? The term digital library, like the word library, means different things to different people. Many people think of libraries as bricks and mortar, a quiet place where books are kept. To professional librarians, they are institutions that arrange for the preservation, collection, and organization of material, as well as for access to it. And a library’s material is not just books: there are libraries of art, film, sound recordings, botanical specimens, and cultural objects. To researchers, libraries are networks that provide ready access to the world’s recorded knowledge, wherever it is held. Today’s university students of science and technology, sadly, increasingly think of libraries as the World Wide Web—that is, they misguidedly regard the Web as the ultimate library.

A digital library is not simply a “digitized library.” We hope that you are reading How to Build a Digital Library because you are thinking of building one. However, we do not imagine that you are the director of the New York Public Library, contemplating replacing that magnificent edifice with a computer (Figure 1.3). Nor do we want you to think, even for a moment, of burning your books at home and sitting by the fireside on winter evenings absorbed in a flat-panel computer display. (Some say that, had books been invented after computers were, they would have been hailed as a great advance.) Rather, we hope that you are inspired by a vision, perhaps something like the examples we began with, of achieving new human goals by changing the way that information is used in the world. Digital libraries are about new ways of dealing with knowledge—preserving, collecting, organizing, propagating, and accessing it—not about deconstructing existing institutions and putting them in an electronic box.

In this book, a digital library is defined as

a focused collection of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization, and maintenance of the collection.

This broad interpretation of “digital objects” (not just text) is reflected in the examples above. Beyond audio and video, we want to include 3D objects, simulations, dynamic visualizations, and virtual reality. The second and third parts of the definition of a digital library deliberately accord equal weight to user (access and retrieval) and librarian (selection, organization, and maintenance). The librarian’s functions are often overlooked by digital library proponents, who generally have a background in technology and approach their work from this perspective rather than from the viewpoint of library or information science.

But selection, organization, and maintenance are central to the notion of a library. If data is characterized as recorded facts, then information is the set of patterns, or expectations, that underlie the
You could go on to define knowledge as the accumulation of your set of expectations, and wisdom as the value attached to knowledge. Not all information is created equal, and wisdom is what librarians add to the library by making decisions about what to include in a collection—difficult decisions—and by following up with appropriate ways of organizing and maintaining the information. Indeed, it is exactly these features that distinguish digital libraries from the anarchic mess that we call the World Wide Web.

Digital libraries do tend to blur what has traditionally been a sharp distinction between user and librarian. (The collections in the examples above were not, in the main, created by professional librarians.) Nevertheless, it is important to keep in mind the distinction between the two roles. Digital library software supports users as they search and browse the collection; equally, it supports librarians as they strive to provide appropriate organizational structures and to maintain them effectively.

Digital libraries are libraries without walls. But they do need boundaries. The very notion of a collection implies a boundary: the fact that some things are in the collection means that others must lie outside it. And collections need a kind of presence, a conceptual integrity, that gives them cohesion and identity: that is where the wisdom comes in. Every collection should have a well-articulated purpose, which states the objectives it is intended to achieve, and a set of principles, which are the
directives that will guide decisions on what should be included and—equally important—what should be excluded. These decisions are difficult.

Digital collections often present an appearance that is opaque: a screen (typically, a Web page) with no indication of what, or how much, lies beyond. Is it a carefully selected treasure or a morass of worthless ephemera? Are there half a dozen documents or many millions? At least physical libraries occupy physical space, present a physical appearance, and exhibit tangible physical organization. When standing on the threshold of a large bricks-and-mortar library, you gain a sense of presence and permanence that reflects the care taken in building and maintaining the collection inside. No one could confuse it with a dung heap! Yet in the virtual world the difference is not so palpable.

We draw a clear distinction between a digital library and the World Wide Web: the Web lacks the essential features of selection and organization. We also want to distinguish a digital library from a Web site—even one that offers a focused collection of well-organized material. Existing digital libraries invariably manifest themselves in this way. But a Web site that provides a wealth of digital objects, along with appropriate methods of access and retrieval, should not necessarily be considered a “library.” Libraries are storehouses to which new material can easily be added. Many well-organized Web sites are created manually through hand-crafted hypertext linkage structures. But just as adding new acquisitions to a physical library does not involve delving into the books and rewriting parts of them, it should be possible for new material to become a first-class member of a digital library without any need for manual updating of the structures used for access and retrieval.

What establishes a new acquisition in tangible format in the collection structure of a physical library is partly where it is placed on the shelves, but more important is the information about it that is included in the library catalog. In digital libraries, we call the equivalent “cataloging” information *metadata* (data about data) and it figures prominently in the digital libraries described in this book.

### 1.2 The Changing Face of Libraries

Libraries are society’s repositories for knowledge: temples, if you like, of culture and wisdom. Born in an era when agriculture was humankind’s greatest preoccupation, libraries experienced a resurgence with the invention of printing in the Renaissance, and really began to flourish when the industrial revolution prompted a series of inventions that mechanized the printing process—the steam press, for example.

Libraries have been around for more than 25 centuries, although only one individual library has survived more than 5 centuries, and most are far younger. The exception is a collection of more than 2,000 engraved stone slabs, or *steles*, situated in Xi’an, an ancient walled city in central China with a long and distinguished history. The collection was established in the Song dynasty (ca. 1100 A.D.) and has been gradually expanded with new work since that time. Each stele stands 2 or 3 meters high and is engraved with a poem, story, or historical record (Figure 1.4). For example, Confucius’s works are here, as is much classic poetry, and an account of how a Christian sect spread eastward to China along the Silk Road. Chinese writing is an art form, and this library gathers together the works of many outstanding calligraphers over a period of two millennia. It also contains the heaviest books in the world.
We think of the library as the epitome of a stable, solid, unchanging institution, and indeed the silent looming presence of 2,000 enormous stone slabs (often called the “forest of steles”) certainly projects a sense of permanence. But this is an exception. Over the years, libraries have continuously evolved. Originally intended for storage and preservation, libraries have refocused to place users at the center, with increased emphasis on information exchange.

Ancient libraries were useful to only the small minority of people who could read, and they were accessible within stringent limitations imposed by social conditions. Medieval monastic and university libraries held chained copies of books in public reading areas. Other copies were available for loan, although substantial security was demanded for each volume borrowed.
The public library movement took hold in the 19th century. Still, the libraries of the day had bookstacks that were closed to the public: patrons perused the catalog and chose their books, which were then handed out over the counter. In continental Europe, most libraries still operate this way. However, progressive 20th-century librarians came to realize the advantage of allowing readers to browse among the shelves and to make their own selections, and the idea of open-access libraries became widely adopted in English-speaking countries, marking the fulfillment of the principle of free access to the contents of libraries by all—the symbolic snapping of the links of the chained book.

Today we stand on the threshold of the digital library. The information revolution not only supplies the technological horsepower that drives digital libraries, but fuels an unprecedented demand for storing, organizing, and accessing information—a demand that, for better or worse, is economically driven, rather than curiosity-driven as in days gone by. If information is the currency of the knowledge economy, digital libraries will be the banks where it is invested. Indeed, Goethe once said that visiting a library was like entering the presence of great wealth that was silently paying untold dividends.

In the beginning

The fabled library of Alexandria is widely recognized as the world’s first great library—although long before it, Assyrian king Assurbanipal (668–626 B.C.) established a comprehensive, well-organized collection of tens of thousands of clay tablets; and long before that, Chinese written records began, having a history extending at least as far back as the 18th century B.C. Created around 300 B.C., the Alexandrian library grew at a phenomenal rate and, according to legend, contained some 200,000 volumes within 10 years.

The work of the acquisitions department was rather more dramatic than in the libraries of today. During a famine, for example, the king refused to sell grain to the Athenians unless he received in pledge the original manuscripts of some leading authors. The manuscripts were diligently copied and the copies returned to the owners, while the originals went into the library. By far the largest single acquisition occurred when Mark Antony stole the rival library of Pergamum and gave it lock, stock, and barrel (200,000 volumes) to Cleopatra as a love token; she passed it over to Alexandria for safekeeping. These acquisitions increased the collection to 700,000 volumes. More than 2,000 years would pass before any other library would attain this size, notwithstanding technological innovations such as the printing press.

Tragically, the Alexandrian library was destroyed. Julius Caesar set fire to the ships in the harbor of Alexandria in 47 B.C. and the fire spread to the shore. Much of the library survived this, but it was willfully laid waste (according to the Moslems) by Christians in 391 A.D. or (according to the Christians) by Moslems in 641 A.D. In the Arab conquest, Amru, the captain of Caliph Omar’s army, would apparently have been willing to spare the library, but the fanatical Omar is said to have disposed of the problem of information explosion with the immortal words, “If these writings of the Greeks agree with the Koran, they are useless and need not be preserved; if they disagree, they are pernicious and ought to be destroyed.”
Moving ahead a thousand years, let us peek at what was happening in a library at a major university near the center of European civilization a century or two after Gutenberg’s introduction of the movable-type printing press around 1450. (The printing press was invented in China much earlier, around five centuries before Gutenberg.) Trinity College, Dublin, one of the oldest universities in Western Europe, was founded in 1592 by Queen Elizabeth I. In 1600 the library contained a meager collection of 30 printed books and 10 handwritten manuscripts. This grew rapidly, by several thousand, when two Trinity College Fellows mounted a shopping expedition to England, and by a further 10,000 when the library received the personal collection of Archbishop Ussher, a renowned Irish man of letters, on his death in 1661.

At the time, however, even this collection was dwarfed by that of Duke August of Wolfenbüttel, Germany, whose collection had reached 135,000 imprints by his death in 1666 and was the largest contemporary library in Europe, acclaimed as the eighth wonder of the world. These imprints were purchased in quires (i.e., unbound) and shipped to the duke in barrels. The duke had them bound in 31,000 volumes with pale parchment bindings that you can still see today. Incidentally, this collection inspired Casanova, after spending seven days visiting the library in 1764, to declare that “I have sometimes thought that the life of those in heaven may be somewhat similar to [this visit].” Coming from the world’s most renowned lover, this is high praise indeed.

Returning to Ireland, another great event in the development of Trinity College occurred in 1801, when an act was passed by the British Parliament decreeing that a copy of every book printed in the British Isles was to be donated to the Trinity College Library. This privilege extends to this day and is shared by five other libraries—the British Library, the University Libraries of Oxford and Cambridge, and the National Libraries of Scotland and Wales. This “legal deposit” law had a much earlier precedent in France, where King François I decreed in 1537 that a copy of every book published was to be placed in the Bibliothèque du Roi (long since incorporated into the French National Library). Similarly, the Library of Congress receives copies of all books published in the United States. But we digress.

There were no journals in Archbishop Ussher’s collection. The first scholarly journals appeared just after his death: the Journal des Scavans began in January 1665 in France, and the Philosophical Transactions of the Royal Society began in March 1665 in England. These two have grown, hydra-like, into hundreds of thousands of scientific journals today, although some are now being threatened with replacement by electronic archives.

In the 18th century, printing really took hold. For example, more than 30,000 titles were published in France during a 60-year period in the mid-1700s. Some 300 years after Gutenberg developed the printing press in order to make the Bible more widely available, the press became the vehicle for disseminating the European Enlightenment—an emancipation of human thinking from the weight of authority of the church.

In the United States, President John Adams created a reference library for Congress when Washington became the new national capital in 1800. He began by providing $5,000 “for the pur-
chase of such books as may be necessary for the use of Congress, and for putting up a suitable apartment for containing them therein.” The first books were ordered from England and shipped across the Atlantic in 11 hair trunks and a map case. The library was housed in the new Capitol until August 1814, when—in a miniature replay of Julius Caesar’s exploits in Alexandria—British troops invaded Washington and burned the building. Some 3,000 volumes were lost in that fire. Another fire destroyed two-thirds of the rebuilt collection in 1851. Unlike Alexandria, however, the Library of Congress has regrown—indeed its rotunda is a copy of the one built in Wolfenbüttel two centuries earlier. Today the Library of Congress contains approximately 22 million volumes.

The information explosion began to hit home in Ireland in the middle of the 19th century. In 1835, work started on the production of a printed catalog for the Trinity College Library (Figure 1.5), but by 1851 only the first volume, covering letters A and B, had been completed. The catalog was finally finished in 1887, but only by restricting the books that appeared in it to those published up to the end of 1872. Other libraries, however, were wrestling with much larger volumes of information. By the turn of the century, the Trinity College Library had around a quarter of a million books, while the Library of Congress had nearly three times that number. Both were dwarfed by the British Museum (now part of the British Library), which at the time had nearly 2 million books, and the French National Library in Paris, which had over 2.5 million.

The Alexandrian principle

In an early statement of library policy, an Alexandrian librarian was reported as being “anxious to collect, if he could, all the books in the inhabited world, and, if he heard of, or saw, any book worthy of study, he would buy it”—and two millennia later this was formulated as a self-evident principle of librarianship: it is a librarian's duty to increase the stock of his library. When asked how large a library should be, librarians answered, “Bigger. And with provision for further expansion.”

Only recently has the Alexandrian principle begun to be questioned. In 1974, following a 10-year building boom then unprecedented in library history, the *Encyclopedia Britannica* noted that “even the largest national libraries are … doubling in size every 16 to 20 years” and gently warned that “such an increase can hardly be supported indefinitely.” And the struggle continues. In the 20th century’s last decade the national libraries of the United Kingdom, France, Germany, and Denmark all opened new buildings. The ones in London and Paris are monumental in scale. Standing on the bank of the Seine River, the Bibliothèque Nationale de France consists of four huge towers designed to look like open books, surrounding a sunken garden plaza (Figure 1.6). The reading rooms occupy two levels around the garden, with bookshelves encircling them on the outer side.

Sustained exponential growth cannot continue. A collection of essays published in 1976 entitled *Farewell to Alexandria: Solutions to Space, Growth, and Performance Problems of Libraries* dwells on the problems that arise when growth must end. Sheer limitation of space has forced librarians to rethink their principles. Now they talk about “aggressive weeding” and “culling,” “no-growth libraries,” the “optimum size for collections,” and some even ask, “Could smaller be better?” In a striking example of aggressive weeding, the library world was rocked in 1996 by allegations that the San Francisco Public Library had surreptitiously dumped 200,000 books, or 20 percent of its collection,
— Resolutie van de staten generael der Vereenigde Nederlanden, dienende tot antwoord op de memo-
rie by de ambassadeurs van sijne majesteyt van Vranckrijck.

— Tractaet van vrede gemaectt tot Nimwegen op den 10 Augusty, 1678, tusschen de ambassadeurs
van [LOUIS XIV.] ende de ambassadeurs vande staten generael der Vereenigde Nederlanden.

— Nederlantsche absolutie op de Fransche bely-
denis.
Amsterdam, 1684. 4º. Fag. H. 2. 50. N°. 22.

— Redenen dienende om aan te wijsen dat haar
ho. mog. [niet] konnen verhindert werden een
vredige afkomst te maken op de conditien by
memorien van den grave d' Avaux van de 5 en 7
Juny, 1684, aangeboden.
[s. l.] 1684. 4º. Fag. H. 2. 86. N°. 3.

— Redenen om aan te wijsen dat de bewuste werving
van 16000 man niet kan gesustineert werden te
zullen hebben konnen strekken tot het bevorderen
van een accommodement tusschen Vrankrijk en
Spaigne.

— D' oude mode van den nieuwen staat van oor-
logh.
Fag. H. 2. 96. N°. 3.

— Aenmerkingen over de althans swevende ver-
schillen onder de leden van den staat van ons
vaderlant.

— Missive van de staten generael der Vereenigde
Nederlanden, ... 14 Maert, 1684.

— Missive van de staten generael der Vereenigde
Nederlanden, ... 11 July, 1684.
Fag. H. 3. 44. N°. 69.

— Resolutie vande staten generael der Vereenigde
Nederlanden, ... 2 Maart, 1684.
Fag. H. 3. 44. N°. 9.

— Extractuyt de resolutien van de staten gene-
rael, ... 31 Maert, 1684.
Fag. H. 3. 44. N°. 15.

— Antwoord van de staten generael der Vereenigde
Nederlanden op de propositie van wegen sijne
churf. dooorl. van Ceulen, Maert 23, 1684, gedaen.
into landfills, because its new building, although lavishly praised by architecture critics, was too small for all the books.

The notion of focused collections is replacing the Alexandrian model of an ideal library that is vast and eternally growing. The notion of service to library users is replacing the idea of a library as a storehouse of all the world’s knowledge. These movements will surely be reinforced by the experience of the World Wide Web, which amply illustrates the anarchy and chaos that inevitably result from sustained exponential growth. The events of the last quarter century have even shaken librarians’ confidence in the continued existence of the traditional library. Defensive tracts with titles like *Future Libraries: Dreams, Madness and Reality* deride “technolust” and the empty promises of the technophiles.

**Early technodreams**

Let us, for a moment at least, consider the technophiles. Around 1936, science fiction writer H. G. Wells promoted the concept of a “world brain” based on a permanent world encyclopedia that
“would be the mental background of every intelligent [person] in the world. It would be alive and growing and changing continually under revision, extension and replacement from the original thinkers in the world everywhere,” and he added sardonically that “even journalists would deign to use it.”

In 1945, Vannevar Bush, the highest-ranking scientific advisor in the U.S. war effort, urged us to “consider a future device for individual use, which is a sort of mechanized private file and library … a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility” (Figure 1.7).

In 1960 J. C. R. Licklider, head of the U.S. Department of Defense’s Information Processing Techniques Office, envisioned that human brains and computing machines would be tightly coupled together and supported by a “network of ‘thinking centers’ that will incorporate the functions of present-day libraries together with anticipated advances in information storage and retrieval.”

Toward the end of the 20th century we became accustomed to hearing similar pronouncements from the U.S. Presidential Office, rising above the road noise of the information superhighway.

**The library catalog**

Wells, Bush, Licklider, and other visionary thinkers were advocating something very close to what we might now call a virtual library. To paraphrase the dictionary definition, something is virtual if it
exists in essence or effect though not in actual fact, form, or name. A virtual library is a library for all practical purposes, but a library without walls—or physical books.

In truth, a virtual representation of books has been at the core of libraries right from the beginning: the catalog. Even before Alexandria, libraries were arranged by subject and had catalogs that gave the title of each work, the number of lines, the contents, and the opening words. In 240 B.C. an index was produced to provide access to the books in the Alexandrian library that was a classified subject catalog, a bibliography, and a biographical dictionary all in one.

A library catalog is a complete model that represents, in a predictable manner, the universe of books in the library. Catalogs provide a summary of library contents. Today we call this metadata. And it is highly valuable in its own right. As a late 19th-century librarian wrote, “Librarians classify and catalog the records of ascertained knowledge, the literature of the whole past. In this busy generation, the librarian makes time for his fellow mortals by saving it. And this function of organizing, of indexing, of time-saving and thought-saving, is associated peculiarly with the librarian of the 19th century.”

Other essential aids to information-seeking in libraries are published bibliographies and indexes. Like catalogs, these are virtual representations—metadata—and they provide the traditional means of gaining access to journal articles, government documents, microfiche and microfilm, and special collections.

A possible interpretation of “digital library” that is quite different from the concept developed in this book is a system designed to automate traditional library functions by helping librarians to manage physical books. Online public access catalogs (OPACs) are standard in libraries today. And many other functions are automated: acquisitions, loans, recalls, interlibrary services, and library planning. However, this kind of library automation system is not closely related to the digital libraries we describe in the examples that open this chapter.

**The changing nature of books**

The technophile visionaries whose dreams we shared above were not talking about a virtual library in the sense of an automated catalog. They wanted the full text of all documents in the library to be automated, not just a metadata surrogate. They took it for granted that books are adequately represented by the information they contain: the physical object is of no consequence.

The information in library catalogs and bibliographies can be divided into two kinds: the first having reference to the contents of books; the second treating their external character and the history of particular copies. Intellectually, only the abstract content of a book—the information contained therein—seems important. But the strong visceral element of books cannot be neglected and is often cited as a reason why book collections will never become “virtual.”

Portable writing began with the clay tablet, invented by the Sumerians in 3500 B.C. A well-known example is the Phaistos Disk from the ancient Minoan civilization (Figure 1.8a). A thousand years later, Egyptians began writing on scrolls made of papyrus plants (a practice that was adopted by various cultures over the centuries, see Figure 1.8b). In other places, animal skins were used for
Figure 1.8: Ancient devices for portable writing: (a) clay tablet; (b) scroll; (c) palm leaves (from Thanjavur Maharaja Serfoji’s Sarasvat Mahal Library, Thanjavur, Tamil Nadu, 1995); (d) concertina; (e) codex
parchment or vellum; elsewhere, bamboo, silk, and even palm leaves served as writing surfaces (Figure 1.8c). But lengthy scrolls become unwieldy—some, like the Great Harris Papyrus composed by King Ramses IV, were as long as 135 feet. A simple solution was to fold the scroll into a concertina (Figure 1.8d). Folded scrolls are more compact, making them easier to handle while reading and to store afterwards. Information can be readily accessed: when seeking particular passages, readers need not unroll and reroll the document. An unfolded concertina book is essentially a scroll, providing backward compatibility.

The book form or “codex” superseded scrolls and concertinas (Figure 1.8e). Around 200 B.C., Greeks and Romans began to write on wax tablets backed with wood, sometimes connected with cords (resembling a ring binder). Later the Romans substituted the more durable vellum for the wood panels.

Because of its numerous virtues, the codex quickly became the preferred format for all literary works. Writers can use both sides; readers can access pages randomly. Codices are easy to read, store, carry, and search. The content is well protected, and thicker pigments can be used for decoration and illustration. The only real disadvantage is that readers can’t view more than two pages at once. Books quickly became the standard document format, although concertinas are still in use today for brochures and maps.

Surprisingly, the evolution of computer output paralleled the book format’s development. Early printers used paper rolls (scrolls); later, line printers used fanfold paper (concertinas). The latter innovation was spurred by accelerated print mechanisms, but old-timers still recall the advantages of fanfolds over unruly rolls of paper. Indeed, paper rolls were eventually perforated so that they could be folded, boxed, and perused more easily. Today we print on pages and bind them into books. The parallel isn’t confined to print technology. Early computer display monitors scrolled, as do Web pages, which originated in the early 1990s. Adobe Reader and Microsoft Word Print Preview provide a paginated concertina-like view: readers scroll through the document page by page, using a scrollbar or Page Up and Page Down keys.

Bibliophiles love books as much for the statements they make as objects as for the statements they contain as text. Indeed, early books were works of art. The steles in Xi’an are a monumental example, studied as much for their calligraphic beauty as for the philosophy, poetry, and history they record, a priceless, permanent record of earlier civilizations. The Book of Kells in Ireland, produced by Irish monks at the scriptorium of Iona about 1,200 years ago, is one of the masterpieces of Western art. Figure 1.9 shows part of a page and illustrates the extraordinary array of pictures, interlaced shapes, and ornamental details. Giraldus Cambrensis, a 13th-century scholar, fancifully wrote that “you might believe it was the work of an angel rather than a human being.”

Beautiful books have always been highly prized for their splendid illustrations, for colored impressions, for heavily decorated illuminated letters, for being printed on uncommon paper or uncommon materials, for their unusual bindings, and for their rarity and historic significance. In the castle library of Königsburg there are 20 books bound in silver, richly adorned with large and beautifully engraved gold plates. Whimsical bindings abound: a London bookseller had Fox’s History of King James II bound in fox skin.
Catalogs and bibliographies are metadata: virtual information about books. In the kind of virtual library sketched in the early technodreams above, the very concept of the book as an individual physical entity seems to be at risk. However, technology has advanced to the point where it need not be: surrogates can substitute for physical works. A picture of the cover may be displayed as a “tangible”—or at least memorable—emblem of the physical book itself. Users can browse the collection using graphical techniques of virtual reality. Maybe they will even be able to caress the virtual cover, smell the virtual pages. But although it is unlikely that readers will love simulated books the way that bibliophiles love real ones, what really matters in libraries is information, even knowledge. Ask the Kataayi community.

1.3 Searching for Sophocles

Pergamos, a digital library at the University of Athens, is representative of many interesting resources that have appeared on the Web over the last decade. It contains four collections spanning
historical archives, folklore, the theater, and sheet music; and more collections are being developed. The name is an alternate spelling for Pergamum, which, as already mentioned, was the site of a significant library in its day. Here we use Pergamos to exemplify features of a real-world digital library.

The system provides an interface in Greek, English, and French. Figure 1.10a shows the home page (in English), which lists the collections available. For each collection a fitting image is shown, along with a summary of what is provided: a description of the resource, the time span it covers, who owns the material, their contact information, and access rights for the user. Some collections, such as the Historical Archive collection, include subcollections: printed documents and photography from the university archives, the Senate Secretariat, and five of the university’s faculties. These can be accessed individually or collectively.

Suppose a user is interested in finding out which of the seven surviving plays by Sophocles have been performed at the university over the years. He brings up the library’s search page shown in Figure 1.10b and selects the Theatrical collection from a nested list of possibilities. Next he enters sophocles in the search box. The result of the query, shown in Figure 1.10c, is disappointing: only one matching document is found. This is an unexpectedly sparse result for a writer of such prominence, and our user mentally runs a few checks. Was there a typing mistake in the query? Could the uncapitalized initial letter have prevented matching? In both cases the answer is No: the name is correctly spelled, and in this digital library search is case-insensitive. Why then was only one match returned?

It is hard not to notice the size of a physical library. But establishing equivalent cues in the digital realm is tricky. Perhaps, wonders our user, the Theatrical collection is rather small after all? To get a handle on its scope and coverage, he clicks the Navigate button in the interface’s main navigation bar, which appears on each page as a row of words beneath the masthead. A few clicks bring up Figure 1.10d, which lists all items in the collection alphabetically by title. There are 633 items in all, and items 1–20 are presented in summary form: title of the show, thumbnail of the program, table displaying author, director, group, theater, and period. Not all items in the list have all these components. The page also contains navigation links: next page and direct entry to a particular page. The sort order can be changed from ascending to descending, and the sort field can also be chosen.

Clicking through a page or two, our user quickly realizes that author names, titles, and so forth are predominately entered in the Greek alphabet. When the user repeats the search in Greek, Σωφοκλῆς, twelve matches are returned (Figure 1.10e), which accords better with the user’s expectations. He scrolls through the returned matches (Figure 1.10f) and selects the play Αντιγόνη (Antigone). Figure 1.10g shows the digitized version of the actual program produced for this show. The thumbnail shown on the left is quick to load, and the main area of the page is taken up with a larger version. A zoom control allows the user to alter the display resolution in order to see different levels of detail. There is a link to the second page (and, where relevant, subsequent pages) of the program. Digital libraries like this can be accessed by anyone, from any corner of the world. This makes it impossible to predict the uses that people will want to make of the material. For instance, research students might study patterns in scribbled annotations left on the paper documents, which would normally be viewed as ephemera, in order to help investigate a hypothesis they have developed. Most visitors are
Figure 1.10: The Pergamos Digital Library: (a) home page (in English); (b) searching the Theatrical Collection for *sophocles*;
Figure 1.10, cont’d: (c) search result; (d) browsing titles alphabetically;
Figure 1.10, cont’d: (e) searching for Σοφοκλῆς; (f) search result;
more interested in theatrical aspects. They might feel that the highest levels of magnification are overkill, but to researchers seeking faded pencil annotations they are a godsend.

Countless digital libraries on the Web adopt the basic form presented in this walk-through. They offer searching and browsing operations to locate items of interest, and ultimately show digital representations of each document, along with ancillary information (metadata). In this example the documents are digitized images (theater programs), to which access is provided through fields of descriptive text, such as title, author, and date. In the case of text-based documents like word-processed reports or Web pages, the full text can be searched as well.

1.4 Digital Libraries in Developing Countries

It sometimes happens that technological advances in developing countries leapfrog those in developed ones. This occurs because established infrastructure, a strong and necessarily conservative force, is absent. For example, in developing countries alternative energy sources, such as solar energy, are widely used in place of traditional power generation and distribution, and these countries have experienced far higher levels of mobile phone growth than developed countries have. Digital
libraries provide another example, compensating for the failure of traditional distribution mechanisms to address local requirements and to deliver information where and when it is needed.

Many current technology trends are not benefiting developing countries—indeed, some bring serious negative consequences. Just as industrialization and globalization have increased the gulf between haves and have-nots, so information and communications technology is creating a chasm between “knows” and “know-nots.” By and large, developing countries are not participating in the information revolution, although knowledge is critical for development. The knowledge gap between rich and poor is widening.

In the developing world, digital libraries provide perhaps the first really compelling raison d’être for computing technology. Priorities in these countries include health, food, hygiene, sanitation, and safe drinking water. Though computers are not a priority, simple, reliable access to targeted information meeting these basic needs certainly is. Digital libraries give system developers a golden opportunity to help reverse the negative impact of information technology on developing countries.

**Disseminating humanitarian information**

Traditional publishing and distribution mechanisms have tragically failed the developing world. Take medicine, a field of great importance in this context. Whereas a U.S. medical library subscribes to about 5,000 journals, the Nairobi University Medical School Library, long regarded as a flagship center in East Africa, received just 20 journals in 1998 (compared with 300 a decade before). In Brazzaville, Congo, the university has only 40 medical books and a dozen journals, all published before 1993, and the library in a large district hospital consists of a single bookshelf filled mostly with novels.

By decoupling production and distribution costs from intellectual property charges, digital libraries offer a lifeline. A wealth of essential humanitarian material is produced by international organizations such as the United Nations, as well as national units like the U.S. Peace Corps. Being produced by internationally oriented, nonprofit organizations, funded by all people on the planet, this information is—at least in principle—in the public domain: it could be made freely available in the form of networked digital libraries. While those 5,000 medical journals cannot be distributed for free because they are produced by commercial publishers, this problem does not arise in many areas of physics, as we have seen. The world is changing, and the rate of change will accelerate.

**Disaster relief**

Natural disasters like earthquakes and hurricanes and human-made ones like terrorist attacks and nuclear accidents demand an immediate and informed response. Disaster relief situations are complex and are addressed by a broad range of players in a variety of organizations acting in parallel. They present an overwhelming need for information: information that is tailored for the problem at hand, organized so that it can be accessed effectively, and distributed even in the absence of an effective network infrastructure. The response to a crisis is characterized by the generation of large amounts of
unstructured multimedia data that must be acquired, processed, organized, and disseminated sufficiently rapidly to be of use to crisis responders.

Digital library technology allows the rapid creation of organized collections of information, graced with comprehensive searching and browsing capabilities. Intelligence specific to the nature of a disaster, the geographical region, and the logistic resources available for the relief effort can all be gathered into a built-to-order digital library collection that combines targeted knowledge with general information about medicine and sanitation.

**Preserving indigenous culture**

Libraries and their close relatives, museums, have always been involved in preserving culture. These institutions collect literature and artifacts and use them to disseminate knowledge and understanding of different times and cultures. Digital libraries, however, open up the possibility of flexible and coherent multimedia collections that are both fully searchable and browsable in multiple dimensions—and permit more active participation by indigenous people in preserving and disseminating their own culture, as is illustrated by the example of the Zia Pueblo. The principal participants here are by definition the indigenous people themselves: the technological world assumes the role of catalyst, midwife, and consumer—once indigenous culture has been recorded, it will find a fascinated, sympathetic, and perhaps influential audience elsewhere.

Information about indigenous culture takes many guises: oral history, in the form of narration and interviews; artifacts, in the form of images and descriptions; songs, in the form of audio recordings, music transcriptions, and lyrics; and dances and ceremonies in the form of video, audio, written synopses, and interpretations. Multimedia digital libraries allow such information to be integrated, recorded, browsed, and searched, all within a uniform user interface.

Because language is the vehicle of thought, communication, and cultural identity, a crucial feature of digital libraries for culture preservation is the ability to work in local languages. This strengthens individual cultures, promotes diversity, and reduces the dominance of English in the global information infrastructure.

**Locally produced information**

In digital library applications for culture preservation, the relevant information is, of necessity, readily available locally. But there are countless other scenarios that involve creating and distributing locally produced information collections. At first glance one might think that the Internet includes such a wealth of content that surely there must be something of benefit to everyone. However, this ignores not only the problem of language—most information is available only in major languages like English—but also the importance of the kind of content that only local communities can generate.

Teachers prepare educational material that addresses specific community problems, and they adapt published material to employ local examples. Indigenous people have medicinal knowledge based on
local plants or long-acquired knowledge of the cultivation and protection of local species. Such knowledge is vital: more than half of the world’s most frequently prescribed drugs are derived from plants or synthetic copies of plant chemicals, and this trend is growing.

Local groups assemble information collections that describe and reflect neighborhood conditions, providing new material for sociocultural studies, fostering cultural exchange while retaining diversity, and increasing international understanding. Web sites for community and social development might include information on health problems endemic to a particular African community, or information on commodity prices for a particular good traded in Brazilian markets, or examples of curricular projects suitable for use in Indian schools.

The development of content that addresses the specific needs of a particular community stimulates the demand for information technology among that community. Getting learners to produce their own content is one of the best ways to exploit information technology in learning situations. It not only improves the learning experience, but also creates material that benefits the community. Teachers and students can work together to create their own content that has value for the community, and for the nation as well.

Effective human development blossoms from empowerment. As the Chinese proverb says, “Give a man a fish and he will eat for a day; teach him to fish and he will eat for the rest of his days.” Dissemination of information that originates in the developed world is useful to developing countries, as Kataayi members will testify, but a more effective strategy for sustained long-term human development is to foster the capability for creating information collections, rather than the collections themselves. This allows developing countries to participate actively in the information society, rather than observing it from outside. It stimulates the creation of new industry, and it helps ensure that intellectual property remains where it belongs, in the hands of those who produce it.

**The technological infrastructure**

Computers are not so hard to come by in developing countries as one might think. Their extraordinarily rapid rate of obsolescence, coupled with the developed world’s voracious appetite for the latest and greatest, makes low-end machines essentially free: instead of clogging landfill sites many (although certainly not enough) find their way to developing countries. A 1998 World Bank survey of developing countries found 3 to 30 PCs per 1,000 people, depending on the poverty level. Using an estimated growth rate of 20 percent per year, we conclude that at the turn of the millennium there were 50 million PCs in developing countries, serving a population of four billion.

A more serious obstacle is that network access varies widely around the globe. Whereas in 1998 more than a quarter of the U.S. population was surfing the Internet, the figure for Latin America and the Caribbean was 0.8 percent, for Sub-Saharan Africa 0.1 percent, and for South Asia 0.04 percent. Schools and hospitals in developing countries are poorly connected. Even in relatively well-off South Africa, many hospitals and 75 percent of schools have no telephone line. In African universities, up to 1,000 people can depend on just one workstation. The Internet is failing the developing world. While global satellite communication networks may eventually bring relief, this takes time and money.
Because of the difficulty of network access, the structure and organization of digital libraries should be separated from their distribution media. Physical distribution of information on recordable devices can provide an attractive alternative to networks. Compact disk read-only memory, CD-ROM, is a practical format for areas with little Internet access. The 650-MB capacity of a CD-ROM can hold a useful volume of information, such as the 1,200 fully illustrated and fully indexed books in the Humanity Development Library. Most of the space in a collection like this is consumed by pictures: several times as many books could be included if they were not so lavishly illustrated. CDs are giving way to digital versatile disks, DVDs, which can hold from 5 to 20 GB of data. A year’s supply of the 5,000 medical journals mentioned above could fit, fully indexed, on a single DVD. And save lives.

1.5 The Pen Is Mighty: Wield It Wisely

Collecting information and making it widely available to others has far-ranging social implications, and those who build digital libraries must act responsibly by making themselves aware of the legal and ethical issues that surround their particular application. Copyright is the place to begin.

Digital libraries are far more accessible than physical ones. And this brings its own problems: access to the information in digital libraries is generally less controlled than it is in physical collections. Putting information into a digital library has the potential to make it immediately available to a nearly unlimited audience.

This is great news. For the user, information becomes available wherever you are, worldwide. For the author, the potential audience is greater than ever before. And for publishers, new markets open up that transcend geographical limitations. On the other hand, authors and publishers worry that fewer copies of a work will be sold if networked digital libraries enable worldwide access to an electronic copy of it. Their nightmare is that the number of copies sold could be as few as one. How many books will be published online if the entire market can be extinguished by the sale of one electronic copy to a public library?

Copyright law

In copyright law, possession of a copy of a document certainly does not constitute ownership. Although there may be many copies, each document has only one copyright owner. This applies not just to physical copies of books, but to computer files, too, whether they have been digitized from a physical work or created electronically in the first place—“born digital.” When you buy a copy of a document, you can resell it, but you certainly do not buy the right to make further copies and redistribute them. That right rests with the copyright owner.

Copyright subsists in the work, rather than in any particular embodiment of it. A work is an intangible intellectual object, of which a document is a physical manifestation. Lawyers use the word subsists, which in English means to remain or to continue in existence, because copyright has no separate existence without the work. Copyright protects the way ideas are expressed, not the ideas themselves. Two works that express the same idea in different ways are independent in copyright law.
Who owns a particular work? The creator is the initial copyright owner, unless the work is made for hire. If the work is created by an employee within the scope of her employment, or under a specific contract that explicitly designates it as being made for hire, the employer or contracting organization owns the copyright. Any copyright owner can transfer, or “assign,” copyright to another party through a written contract. Typically, an author sells copyright to a publisher (or grants the publisher an exclusive license) and the publisher reproduces, markets, and sells the work.

The copyright owner has the exclusive right to do certain things with the work: thus copyright is sometimes referred to as a “bundle” of rights. In general, there are four rights, although details vary from country to country. The reproduction right allows the owner to reproduce the work freely. The distribution right allows the owner to distribute it, but this is a one-time right: once a copy has been distributed, the copyright owner has no right to control its subsequent distribution. For example, if you buy a book, you can do whatever you want with your copy, including reselling it. The public lending right compensates authors for public lending of their work—although an exception is granted for not-for-profit and educational use, which do not require the copyright holder’s consent. The remaining rights, called other rights, include permitting or refusing public performance of the work, and making derivative works like plays or movies.

Copyright law is complex, arcane, and varies from one country to another. The British Parliament adopted the first copyright act in 1710; the U.S. Congress followed suit in 1790. Although copyright is national law, most countries today have signed the Berne Convention of 1886, which lays down a basic international framework. According to the Convention, copyright subsists without formality, which means that (unlike patents) it’s not dependent on registering a work with the authorities or depositing a copy in a national library. It applies regardless of whether the document bears the international copyright symbol ©. You automatically have copyright over works you create (unless they are made for hire). Some countries—including the United States—maintain a copyright registry even though they have signed the Berne Convention, which makes it easier for a copyright holder to take legal action against infringers. Nevertheless, copyright still subsists, even if the work has not been registered.

The Berne Convention decrees that it is always acceptable to make limited quotations from protected works, with acknowledgment and provided it is done fairly. The United States has a copyright principle called “fair use” that allows material to be copied by individuals for research purposes. The U.K. equivalent, which has been adopted by many countries whose laws were inherited from Britain in colonial times, is called “fair dealing” and is slightly more restrictive than fair use.

Making copies of copyrighted works for distribution or resale is prohibited. That is the main economic point of the copyright system. The Berne Convention also recognizes certain moral rights. Unlike economic rights, these cannot be assigned or transferred to others; they remain with the author forever. They give authors the right to the acknowledgment of their authorship, and to the integrity of their work—which means that they can object to a derogatory treatment of it.

The public domain

Works not subject to copyright are said to be in the “public domain,” which comprises the cultural and intellectual heritage of humanity that anyone may use or exploit. Often, works produced by the
government are automatically placed in the public domain, or else the government sets out generous rules for their use by not-for-profit organizations. This applies only in the country of origin: works produced by the U.S. government are in the public domain in the United States and its territories, but are subject to U.S. copyright in other countries.

Copyright does not last forever; when it expires, the work passes into the public domain, free of all copyright restrictions. No permission is needed to use it in any way, incorporate any of its material into other works, or make any derivative works. You can copy it, sell it, excerpt it—or digitize it and put it on the Web. The author’s moral rights still hold, however, so you must provide due attribution.

As internationally agreed in the Berne Convention, the minimum copyright duration is life plus 50 years, that is, until 50 years after the author dies. Of course, it is often difficult to find out when the author died. One way is to consult the authors’ association in the appropriate country, which maintains links to databases maintained by authors’ associations around the world.

The duration of copyright has an interesting and controversial history. Many countries specify a longer term than the minimum, and this changes over the years. The original British 1710 act provided a term of 14 years, renewable once if the author was alive; it also decreed that all works already published by 1710 would get a single term of 21 further years. The 1790 U.S. law followed suit, with a 14-year once-renewable term. Again, if an author did not renew copyright, the work automatically passed into the public domain. In 1831, Congress extended the initial period of copyright from 14 to 28 years, and in 1909 it extended the renewal term from 14 to 28 years, giving copyright a maximum duration of 56 years.

From 1962 onward, Congress enacted a series of copyright extensions, some of one or two years, others of 19 or 20 years. In 1998, it passed the Sonny Bono Copyright Term Extension Act, which extended the term of existing and future copyrights by 20 years. (The act is named in memory of former musician Sonny Bono, who, according to his widow, believed that “copyrights should be forever.”) The impetus behind the changes was the desire of large, powerful corporations to protect a minuscule number of cultural icons; opponents call them the “Mickey Mouse” copyright extensions. Many parts of the world (notably the United Kingdom) have followed suit by extending their copyright term to life plus 70 years—and in some countries (e.g., Italy) the extension was retroactive, so that works already in the public domain were suddenly removed from it.

The upshot is that copyright protection ends at different times depending on when the work was created. It also begins at different times. In the United States, older works are protected for 95 years from the date of first publication. Through the 1998 Copyright Extension Act, newer ones are protected from the “moment of their fixation in a tangible medium of expression” until 70 years after the author’s death. Works made for hire—often ones belonging to corporations—are protected for 95 years after publication or 120 years after creation, whichever comes first.

The original copyright term was one-time renewable, but few copyright holders renewed. Focusing again on the United States, in 1973 more than 85 percent of copyright holders failed to renew their copyright, which meant that, at the time, the average term of copyright was just 32 years. Today there is no renewal requirement for works created before 1978: copyright is automatically given for a period of 95 years—tripling the average duration of protection.
No copyrights will expire in the 20-year period from 1998 to 2018. To put this into perspective, 1 million patents will pass into the public domain during the same period. The effect of this extension is dramatic. Of the 10,000 books published in 1930, only a handful (less than 2 percent) are still in print. If the recent copyright extensions had not occurred, all 10,000 of these books would by now be in the public domain, their copyright having expired in 1958 (after 28 years) if it was not renewed, or 1986 (after a further 28 years) if it was renewed. Unfortunately, that is not the case.

If you want to digitize one of these 1930 books and make it available on the Internet, you must first determine its copyright status. If copyright expired in 1958 without renewal, it is already in the public domain. The official registry of copyright renewal is not available online, but new initiatives like the Copyright Renewal Database at Stanford University and the Registry of Copyright Evidence by the Online Computer Library Center (OCLC) are trying to help streamline the task. With or without the help of these newer initiatives, for works still under copyright, you will have to contact the copyright holders. If the book is out of print (98 percent of cases), they would likely be perfectly happy to give you permission. But you’d have to track these people down.

As Lawrence Lessig wrote in his book *Free Culture* (from which much of the above information was taken),

> Now that technology enables us to rebuild the library of Alexandria, the law gets in the way. And it doesn’t get in the way for any useful copyright purpose, for the purpose of copyright is to enable the commercial market that spreads culture. No, we are talking about culture after it has lived its commercial life. In this context, copyright is serving no purpose at all related to the spread of knowledge. In this context, copyright is not an engine of free expression. Copyright is a brake.

**Relinquishing copyright**

As we have explained, the Berne Convention grants copyright without formality, without registration. Anything you write, every creative act that is “fixed in a tangible means of expression”—be it a book, an e-mail, or a grocery list—is automatically protected by copyright until 50 years after you die (according to the Berne Convention’s minimum restrictions) or, today in the United States, 70 years after you die (assuming you did not write your grocery list as a work made for hire). People can quote from it under the principle of fair use, but they cannot otherwise use your work until such time as copyright expires, unless you reassign the copyright.

Authors who wish to relinquish copyright must take active steps to do so. In fact, it’s quite difficult. To facilitate it, a nonprofit organization called the Creative Commons has developed licenses that people can attach to the content they create. Each license is expressed in three ways: a legal version, a human-readable everyday-language description, and a machine-readable tag. Content is marked with the CC mark, which does not mean that copyright is waived but that freedoms are given to others to use the material in ways that would not normally be permissible by default under copyright.

The freedoms allowed by licensing all go beyond traditional fair use, but their precise nature depends on the choice of license. One license permits any use so long as attribution is given. Another permits only noncommercial use. A third permits any use within developing nations. Or any educational use. Or any use except for the creation of derivative works. Or any use so long as the same freedom is
given to other users. Most important, according to the Creative Commons, is that these licenses express what people can do with your work in a way they can understand and rely upon without having to hire a lawyer. The idea is to help reconstruct a viable public domain.

The term *copyleft* is sometimes used to describe a license that imposes the same terms on any derivative work: typically the rights to use, modify, and redistribute. The intention of copyleft is to use the facilities of copyright legislation to preserve these freedoms. So, if you take a copyleft work, make changes and distribute a modified version, then other people automatically acquire the same rights (to change and distribute) that you had received.

A prominent figure that helped establish this form of licensing is Richard Stallman. In 1984 he left MIT and founded the GNU Project and set about developing—and persuading over time others to help develop with him—a comprehensive suite of general purpose software. (The acronym, GNU, is itself a self-referential joke, standing for GNU’s Not Unix.) The produced software was made available under the terms of the GNU General Public License (GPL), encapsulating the notion of copyleft, and at the time a radical departure from how mainstream software was distributed. Subsequently the project extended the idea to other forms of work. For instance, the GNU Free Documentation License is the counterpart to GPL, originally designed for the manuals and documentation that accompany GNU software. Others are free to make use of the GNU licenses for their own work, and it is a popular choice for many open source projects.

**Digital rights management**

A treaty adopted by the World Intellectual Property Organization (WIPO) in 1996 addresses some of the copyright issues raised by digital technology and networks in the modern information era. It decrees that computer programs should be protected as literary works and that the arrangement and selection of material in databases is protected. It provides authors of works with more control over their rental and distribution rights than the Berne Convention does.

In order to understand the motivation behind this treaty, and its implications, reflect on the enormous repercussions of current innovations in electronic publishing. Publishers are accelerating production of digitized versions of their titles (e-Books) and selling them online. They continue to rely on conventional retailers for the sale of paper copies, but they are experimenting with combining e-Books with pre- and post-release paper copies. E-Book technology is being forced to standardize.

New sales models are beginning to emerge. Highly effective search engine advertising is leveling the market, providing more opportunities for small publishers and self-publishing. Physical bookstores are finding themselves bypassed. E-Books also provide preview options (flip through the pages at no cost), and can be rented on a time-metered or absolute duration basis (like video rentals), sometimes with an option to purchase. Publishers are also experimenting with print-on-demand technology.

Content owners (publishers) are adopting technical means to implement policies governing access to the information they sell. The term “digital rights management” (DRM) refers to the control and protection of digital content, including text documents, images, video, and audio. DRM technology limits what users can do with content—even when they possess it. Access can be restricted to a
particular computer: this means no lending to friends, no sharing between your home computer and your office, and no backing up on another machine. It also precludes resale, eliminating the second-hand market. Furthermore, expiry dates can be imposed, precluding permanent collections and archives. These measures go far beyond the traditional legal bounds of copyright.

In the entertainment industry, DRM schemes are used to counter perceived threats of piracy. Some complain that these schemes are concerned solely with content owners’ rights, not with users’ rights. DRM does not grant “permissions,” which is what copyright authorizes, but instead enforces absolute, mechanical “controls.” For example, the motion picture industry establishes control by compelling manufacturers to incorporate encryption into their products because it holds key patents on DVD players.

The WIPO Treaty introduces an important but controversial requirement that countries must provide effective legal measures against the circumvention of DRM schemes. One of the first national laws implementing the WIPO Treaty was the U.S. Digital Millennium Copyright Act (DMCA), which, among many other things, makes it unlawful to publish information that exposes the weaknesses of technical protection measures. In Europe, the European Council has approved the treaty on behalf of the European Community and has made directives that largely cover its subject matter.

Such legislation jeopardizes basic rights legally enshrined in the concept of copyright. DRM allows reading rights to be controlled, monitored, and withdrawn, and DMCA legislation makes it illegal for users to seek redress by taking matters into their own hands. In scholarly publishing, DRM is already well advanced. Academic libraries license access to content in electronic form, often in tandem with purchase of print versions. Because they form the entire market, they have been able to negotiate reasonable conditions with publishers. However, libraries have far less power in the consumer book market. For example, one can envisage a scenario where publishers establish a system of commercial pay-per-view libraries for e-Books and refuse public libraries access to books in a form that can be circulated.

Copyright and digitization

Many digital library projects involve digitizing documents. If the work to be digitized is in the public domain, or it attempts to faithfully reproduce a work in the public domain, you may digitize the work without securing anyone’s permission. Of course, the result of your digitizing efforts is not protected by copyright either, unless you produce something more than a faithful reproduction of the original.

If material has been donated to your institution for digitizing, and the donor is the copyright owner, you can certainly go ahead, provided the donor gave your institution the right to digitize—perhaps in a written form, such as “the right to use the work for any institutional purpose, in any medium.” Even without a written agreement, it may be reasonably assumed that the donor implicitly granted the right to take advantage of new media, provided the work continues to be used for the purpose for which it was donated. You do need to ensure, of course, that the donor is the original copyright owner and has not transferred copyright. For example, you cannot assume permission to digitize letters sent to the donor but written by others.
If you want to digitize documents and the above considerations do not apply, you should consider whether you can go ahead under the concept of fair use. This is a difficult judgment to make. You need to reflect on the copyright owner’s concerns and address them. Institutional policies about who can access the material, backed up by practices that restrict access appropriately, can help. Finally, if you conclude that fair use does not apply, then you must obtain permission to digitize the work or acquire access to it by licensing it.

Thus, building a digital library requires serious attention to copyright. Digital library projects must be undertaken with a full understanding of ownership rights and with full recognition that permission to convert materials that are not in the public domain is essential. Because of the potential for legal liability, any prudent library builder should consider seeking professional advice. A full account of the legal situation is far beyond the scope of this book, but the “Notes and sources” section at the end of the chapter (Section 1.8) contains sources for practical information about copyright. The sources include information on how fair use can be interpreted and on the issues involved in negotiating copyright permission or licensing.

Looking at the situation from an ethical, rather than legal, perspective sheds light on the fundamental issues. It is unethical to steal: deriving profit by distributing a book for which someone else has rightful claim to copyright is wrong. It is unethical to deprive someone of the fruit of their labor: giving away electronic copies of a book for which someone else has rightful claim to copyright is wrong. It is unethical to pass someone else’s work off as your own: making a digital library collection without proper acknowledgment is wrong. It is unethical to willfully misrepresent someone else’s point of view: modifying documents before including them in the collection is wrong, even if the original authorship is acknowledged.

**Collecting from the Web**

The legal status of documents published on the World Wide Web is murky. Because activities like Web searching and archiving are in a state of rapid transition, it is impractical, and also inappropriate, for legal regulation to try to keep up with the change. If any legislation is needed, it should be designed to minimize harm to interests affected by technological change while at the same time enabling and encouraging effective lines of development. Legislators are adopting a “wait and see” policy, while leading innovators strive to ensure that what they do is reasonable and accords with the spirit—if not necessarily the letter—of copyright law.

Issues abound. Some lawyers have questioned whether it is legal even to view a document on the Web, since one’s browser inevitably makes a local copy without explicit authorization. Of course, it is widely accepted that you can view Web documents—after all, that’s what they’re there for. Next comes the question whether you can save Web documents for personal use. Or link to them. Or distribute them to others. Note that, behind the scenes, Web documents are routinely copied and saved. For example, to economize on network traffic and to accelerate delivery, Web cache mechanisms save copies of documents locally and deliver them to other users.

The way that computers in general, and the Web in particular, operate raises the question whether the notion of a “copy” is perhaps no longer the appropriate foundation for copyright law in the digital
age. Legitimate copies of digital information are made so routinely that restrictions on the act of copying no longer serve to regulate and control use on behalf of copyright owners. Because computers make many internal copies when they are used to access information, the fact that a copy has been made says little about the legitimacy of the behavior. In the digital world, copying is so bound up with the way computers work that controlling it provides unexpectedly broad powers, far beyond those intended by copyright law.

All these points have an immediate and practical impact on digital libraries. Digital libraries are organized collections of information. The Web is full of unorganized information. Downloading parts of Web content in order to organize information into focused collections and to make the material more useful to others is a prime application area for digital libraries.

Search engines, one of the most widely used services on the Internet, are a good example. They use software “robots” to continually download huge portions of the Web and create indexes to the content. Although a service provider may retain documents on their own computers, searchers are presented with a summary and are directed to the original source documents rather than to local copies. Search engines are commercial operations, but their services are not sold directly to users; instead, their revenue is derived from advertising—in effect, a tax on the user’s attention. Although search engines are widely accepted and used, their legal status is unclear.

Web sites can safeguard against indiscriminate downloading. A generally accepted robot exclusion protocol allows individual Web sites to prevent their content from being downloaded and indexed. Although this protocol is entirely voluntary, widely used search engines certainly comply with it. But the onus of responsibility has been shifted. Previously, to use someone else’s information legitimately, one had to request explicit permission from the information provider. Now, search engines automatically assume permission unless the provider has set up an exclusion mechanism. This is a key development with wide ramifications. And some Web sites threaten dire consequences for computers that violate the robot exclusion protocol—for example, denial-of-service attacks on the violating computer. This is law enforcement on the wild Web frontier.

Different copyright issues are raised by projects that are archiving the entire World Wide Web. The rationale for creating such an archive is to offer services like supplying documents that are no longer available or providing a “copy of record” for publicly available documents—in effect, supplying the raw material for historical studies. However, creating an archive raises many interesting issues involving privacy and copyright, issues that are not easily resolved.

What if a college student created a Web page that had pictures of her then-current boyfriend? What if she later wanted to “tear them up,” so to speak, yet they lived on in the archive? Should she have the right to remove them? In contrast, should a public figure—a U.S. senator, for instance—be able to erase data posted from his or her college years? Does collecting information made available to the public violate the “fair use” provisions of copyright law?

Most digital libraries aim to provide more comprehensive searching and browsing services than search engines do. Like archives, they probably want to store documents locally, to ensure their continued availability. Documents are more likely to be seen as part of the library, rather than as products of their original Web site. Digital libraries are more likely to modify documents as an aid to the user, least invasively by highlighting search terms or adding metadata, more invasively by re-
presenting them in a standard format, most invasively by producing computer-generated summaries or extracting keywords and key phrases automatically.

Those responsible for such libraries need to consider carefully the ethical and legal issues involved. It is important to respect robot exclusion protocols. It is important to provide mechanisms whereby authors can withdraw their works from the library. It is helpful if explicit permission can be sought to include material. If information is automatically derived from, or added to, the source documents, it is necessary to be sensitive to possible misrepresentation.

The world is changing. Digital libraries are pushing the boundaries of what is possible by organizing anthologies of material. And they are pushing the boundaries of society’s norms for distribution of intellectual property. Those who run large-scale Internet information services tell interesting “war stories” of people’s differing expectations of what it is reasonable for their services to do. For example, search engine operators frequently receive calls from computer users who have noticed that some of their documents are indexed when they think they shouldn’t be. Sometimes users feel their documents couldn’t possibly have been captured legitimately because there are no links to them. Most search engines have a facility for locating any documents that link to a specified one and can easily locate the omitted link. On other occasions, people have put confidential documents into a directory that is open to the Web, perhaps just momentarily while they change the directory permissions, only to have them grabbed by a search engine and made available for all the world to see.

Search technology makes information readily available that may previously have been public in principle, but impossible to find in practice. When a major search engine took over the archives of a corpus of Internet discussion groups on a wide range of topics, it received many pleas from contributors to retract indiscreet postings because, now that postings were easily available for anyone to find, they were causing their authors embarrassment.

**Illegal and harmful material**

Some material is illegal and harmful and clearly inappropriate for public presentation (because such material is distasteful, we will not give examples). For example, a 1999 UNESCO Expert Meeting on Paedophilia on the Internet noted,

> Violence and pornography have invaded the Internet. Photos and videos of children and young teenagers engaged in sexual acts and various forms of paedophilia are readily available. Reports of children being kidnapped, beaten, raped and murdered abound.... The Internet has in many cases replaced the media of paedophiliac magazines, films and videos. It is a practical, cheap, convenient and untraceable means for conducting business as well as for trafficking in paedophilia and child pornography. The Internet has also become the principal medium for dialogue about paedophilia and its perpetuation.

UNESCO has taken the lead on breaking the silence on this topic and is engaged in a number of initiatives to provide safety nets for children online.

Whether information is considered harmful often depends on the cultural, religious, and social context in which it is circulated. Standards vary both within and among nations. However, the international nature of the Internet means that it is no longer possible to police the transfer of information, and sustaining local legal and cultural standards is a huge challenge facing society today. The
challenge includes the dilemma of balancing freedom of expression against citizens’ rights to be protected from illegal or harmful material.

In 2000, a well-publicized example of different views on Internet access concerned online sales of Nazi memorabilia on U.S. Web sites accessed using the Yahoo Internet portal. A judge in Paris ruled that the sites are barred under French law and ordered them to be blocked. However, U.S. Web sites are governed by U.S. laws, and an American judge ruled that the First Amendment protects content generated in the United States by American companies from being regulated by authorities in countries that have laws more restrictive of freedom of expression. Suit and countersuit followed, and the matter was not settled for six years, when a U.S. court decided that Yahoo was liable for a fine levied in France.

Another challenging example is online gambling. The relevant laws are restrictive (or at best muddy) in countries like the United States, China, and Italy. Some international gambling sites claim to comply with local laws by checking the geographical origin of the user (a difficult and unreliable procedure that is easily circumvented) and by refusing to offer their services in countries where gambling is illegal.

**Cultural sensitivity**

Most digital libraries are produced by people from Western backgrounds, yet the majority of the world’s population live in countries that have very different cultures. Some digital libraries are specifically aimed at people in different parts of the world: collections for developing countries, for example, or collections aimed at preserving and promoting indigenous cultures. It is clearly necessary for digital library developers to consider how their creations will affect other people.

First, as we already mentioned, language is the vehicle of thought, communication, and cultural identity, and so digital library users should be able to work in whatever language suits them. But the need for cultural sensitivity goes even deeper. Particular labels can have strong, unexpected connotations: certain models of cars have failed to sell in countries where the model’s name had a serious negative association. Icons also have cultural implications: dogs, for example, are offensive in Arabic cultures and will have negative effects if they are offered as user interface icons. Furthermore, different cultures have different color preferences, and particular colors have different associations.

In Polynesian cultures the concept of *tapu*, usually translated as “sacred,” has rich and complex connotations that are difficult for Westerners to appreciate. Many objects have different degrees of *tapu*, and Polynesians find it rude and offensive if these objects are used inappropriately, in the same way that many Westerners find blasphemy rude and offensive. An example that can affect digital library design is that representations of people—including pictures—are *tapu*, and in Polynesia it is generally inappropriate for them to be on public display.

**1.6 Planning a Digital Library**

Are you going to build a digital library? We hope so. And as with any complex system, it is wise to undergo a planning stage before immersing yourself in the detail. Table 1.1 shows some of the ques-
The world of digital libraries

You may not be able to answer them all immediately, but you will have to answer them all eventually. When you have read this book you will be able to approach such questions with a solid understanding of the spectrum of possible answers and the implications of each one. Perhaps even more importantly, you will be able to reassess your answers when technology changes. This is what it means to understand the field of digital libraries.

The questions are divided into three categories: users, materials, and technology. Despite this neat structure, however, the issues are complex and resist any simple categorization. For example, choosing a suitable format for presenting material to users requires an understanding of:

- the background of the users and the technology they use for access;
- the original format of the material and whether it needs to be converted before presentation;

Table 1.1: Pertinent questions

<table>
<thead>
<tr>
<th>Users</th>
<th>1. Who are the intended users?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. Where are they?</td>
</tr>
<tr>
<td></td>
<td>3. What computer experience do they have?</td>
</tr>
<tr>
<td></td>
<td>4. What languages do they know?</td>
</tr>
<tr>
<td></td>
<td>5. Will they need help in accessing the library?</td>
</tr>
<tr>
<td></td>
<td>6. Why do they want to access this material?</td>
</tr>
<tr>
<td></td>
<td>7. What technology (e.g., Web browser) will they use?</td>
</tr>
<tr>
<td></td>
<td>8. To what extent should the library record usage?</td>
</tr>
<tr>
<td></td>
<td>9. Can users contribute to the digital library?</td>
</tr>
<tr>
<td></td>
<td>10. How will you evaluate the project’s success?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Material</th>
<th>1. What will be in the digital library?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. What format is the material in now?</td>
</tr>
<tr>
<td></td>
<td>3. What format does it need to be in for presentation?</td>
</tr>
<tr>
<td></td>
<td>4. Will users need it to be in multiple formats?</td>
</tr>
<tr>
<td></td>
<td>5. Do the formats require conversion of the material?</td>
</tr>
<tr>
<td></td>
<td>6. How will you resource the conversion?</td>
</tr>
<tr>
<td></td>
<td>7. Are there copyright or other restrictions?</td>
</tr>
<tr>
<td></td>
<td>8. Will the library be public or restricted to specific clients?</td>
</tr>
<tr>
<td></td>
<td>9. Will you add value (i.e., metadata) to the material?</td>
</tr>
<tr>
<td></td>
<td>10. If so, how will you resource this activity?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Technology</th>
<th>1. What computers will host the digital library?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. Who maintains them?</td>
</tr>
<tr>
<td></td>
<td>3. What software will be used?</td>
</tr>
<tr>
<td></td>
<td>4. Do you have resources to purchase/license/maintain it?</td>
</tr>
<tr>
<td></td>
<td>5. How will the material be converted to the delivery format?</td>
</tr>
<tr>
<td></td>
<td>6. How will you control access?</td>
</tr>
<tr>
<td></td>
<td>7. How will you interoperate with other libraries?</td>
</tr>
<tr>
<td></td>
<td>8. Can the material be exported from the digital library software?</td>
</tr>
<tr>
<td></td>
<td>9. What would this cost?</td>
</tr>
<tr>
<td></td>
<td>10. If you add value, can these additions be exported?</td>
</tr>
</tbody>
</table>
Chapter 1

- technological options for conversion;
- the possibilities for information loss during conversion;
- resourcing requirements for conversion.

The issues are strongly interconnected, there are many trade-offs, and each collection has its own idiosyncrasies. Moreover, computer-based systems change extremely rapidly—in so-called Internet-time—and particular ways of resolving issues in Table 1.1 will likely soon be out of date.

This book proceeds in roughly the same way as Table 1.1. Chapter 2 focuses on the people in digital libraries. The most important people are the users (customers, if you like) and the questions about users in Table 1.1 are much harder to answer for digital libraries than for traditional ones. Chapter 3 concerns the user experience: What might the final digital library system look like? We examine different kinds of documents and how they can be presented, multimedia as well as text, and document surrogates as well as full documents. We also consider how users access the library, and how they experience the fundamental operations of searching and browsing.

The following three chapters concern the second block of Table 1.1: the material in digital libraries. Chapters 4 and 5 discuss the form of the source documents, the raw ingredients for the library. A major question in practice is whether material will be obtained from physical documents, such as books and journals, and Chapter 4 includes an extensive section on scanning and optical character recognition (the technology for producing electronic text from scanned pages). This can be a costly process, particularly if high accuracy is needed, which makes questions about resourcing the conversion particularly pertinent and concerns about the cost trade-offs particularly acute. In many digital collections, electronic conversion is by far the major cost. Another potential expense is adding value to the contents by providing metadata, including enhanced descriptions of each item. Metadata effectively binds the raw ingredients into a coherent form: Chapter 6 discusses its shape and form, and explores the role it plays bringing order and organization to a library.

The third block of Table 1.1 is technology. The nitty-gritty details of implementation are given in Part II of this book. However, Chapter 7 introduces a wider set of technological issues, including the protocols and services that aid interoperability. The last three questions in Table 1.1 concern risk management in an unpredictable technological environment. What if circumstances change? Through no fault of your own, software and implementation strategies you choose today may turn out badly tomorrow. If circumstances change and you wish to migrate the library to another system, you will want to preserve the investments you have made—another aspect of interoperability. Chapter 8 looks at the global challenges of internationalization, which are both human and technological—recall the example of the alphabet used when searching for Sophocles. Chapter 9, which concludes Part I, presents visions of digital libraries past, present, and future, and highlights the difficulty of long-term preservation.

Resource considerations plague almost every endeavor in life, and libraries are no exception. Several of the questions in Table 1.1 ask about the resources required for different parts of the process. Our own experience is that the decision-makers who control resources are often tremendously impressed if they see a prototype of the digital library. Somehow, when it’s their own material or material that
they have a stake in, even a small, rough-hewn demo can prove inspirational—and attract funding. It’s usually very worthwhile to obtain a small sample of the material and build it into a small prototype that you can show to people in a position to argue for resources to support a major effort. Think big—but, in the first instance, act small. The important thing is to act.

1.7 Implementing a Digital Library: The Greenstone Software

Part I of this book is general, drawing examples from the panoply of digital library innovation and implementation around the globe. But this is a practical book about *how to build* a digital library. Rather than attempting a high-level survey of the many different software products that are available, Part II focuses on a particular system for building digital libraries: the Greenstone Digital Library software from the New Zealand Digital Library Project at the University of Waikato. Almost all of what is presented in Part I can be implemented using the Greenstone toolkit.

Table 1.2 summarizes key features of the Greenstone software. A *library* can include multiple information collections, each individually designed to meet the needs of users and to take advantage of whatever organizational information is available (or can be provided). A *collection* comprises several (typically several thousand, or several million) documents. *Documents* are the fundamental unit from which collections are built; they may have internal structure, alternative representations, and associated files. Documents may contain text, images, sound, or video; collections may include a mixture of document types. *Metadata* is bibliographic information associated with individual documents.

Each collection provides a uniform interface through which its documents can be accessed, an interface that is designed by the librarian responsible for the collection. Prior to presentation, collections undergo a *building* process that, once established by the librarian, is completely automatic. This creates all the structures that are used to access the collection. Searching is based on indexes of full text and metadata. Browsing is based on structures built from metadata. When new material appears, it can be fully incorporated into the collection by rebuilding.

To address the exceptionally broad requirements of digital libraries, the Greenstone system is public and extensible. It is issued as free open-source software under the GNU General Public License. It is widely used throughout the developed and developing worlds. Further details, and many examples, can be obtained from [www.nzdl.org](http://www.nzdl.org). Part II of the book describes how to use the software to build digital library collections, customize them to particular requirements, and use Greenstone’s many interoperability features.

1.8 Notes and Sources

To avoid breaking up the flow of the main text, all references and explanatory notes are collected in a section at the end of each chapter. This first “Notes and sources” section describes information sources, papers, books, and other resources relevant to the material covered in Chapter 1.

We learned about the Kataayi Cooperative from Emmanuel Kateregga-Ndawulu, the chairman. If you would like to learn more about this fascinating initiative, a web search for *Kataayi* will turn up some
Table 1.2: The Greenstone digital library software

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reader’s interface</td>
<td>Readers access collections using a Web browser</td>
</tr>
<tr>
<td>Multiplatform server</td>
<td>Digital library server runs on any version of Windows, Macintosh (OS X), Linux, or other Unix versions</td>
</tr>
<tr>
<td>Librarian’s interface</td>
<td>Librarians design and build collections, and add documents and metadata, using an interactive interface</td>
</tr>
<tr>
<td>Documents</td>
<td>Ingests all standard document types, e.g., HTML, Word, RTF, PDF, PostScript, PowerPoint, Excel, e-mail</td>
</tr>
<tr>
<td>Multimedia</td>
<td>Ingests multimedia objects: images (e.g., GIF, PNG, JPEG, TIFF), audio (e.g., MP3, Ogg Vorbis, WAV), video (e.g., MPEG, Quicktime, RealMedia)</td>
</tr>
<tr>
<td>Metadata</td>
<td>Ingests all standard metadata types, e.g., XML, MARC, OAI, METS, CSV, CDS/ISIS, ProCite, BibTex</td>
</tr>
<tr>
<td>Metadata sets</td>
<td>Defaults to Dublin Core but can use any metadata set; incorporates metadata set editor to create ones</td>
</tr>
<tr>
<td>Extensibility</td>
<td>“Plug-ins” can be written to accommodate new document and metadata types</td>
</tr>
<tr>
<td>Searching</td>
<td>Full text is searchable by default; librarians can add indexes of (any combination of) metadata elements</td>
</tr>
<tr>
<td>Browsing</td>
<td>Librarians can add linearly or hierarchically browsable lists of any metadata elements</td>
</tr>
<tr>
<td>Access structures</td>
<td>All access structures are created automatically; no links are inserted by hand</td>
</tr>
<tr>
<td>Indexers</td>
<td>Incorporates a choice of full-text indexing software: MG, MGPP and Lucene</td>
</tr>
<tr>
<td>Incremental operation</td>
<td>Can operate incrementally when documents are added to or deleted from collections</td>
</tr>
<tr>
<td>Scheduled rebuilding</td>
<td>Includes end-user facilities for scheduling regular rebuilding of collections</td>
</tr>
<tr>
<td>Interoperability</td>
<td>Serves and harvests OAI-PMH; exports and ingests to/from METS and DSpace; harvests over Z39.50 and SRU; ingests MediaWikis</td>
</tr>
<tr>
<td>Multilingual content</td>
<td>Unicode is used internally and converted on the fly to a coding standard supported by the user’s Web browser.</td>
</tr>
<tr>
<td>Multilingual interface</td>
<td>Reader’s interface is available in 50 languages; librarian interface and documentation are available in several languages</td>
</tr>
<tr>
<td>User authentication</td>
<td>Built-in access control mechanism for collections and individual documents to authorized users</td>
</tr>
<tr>
<td>Logging</td>
<td>(Optionally) records all user actions</td>
</tr>
<tr>
<td>Dynamic operation</td>
<td>Collections can be updated and new ones brought online at any time, without disturbing users</td>
</tr>
<tr>
<td>Publishing collections</td>
<td>Collections can be published on removable media such as CD-ROM/DVD/USB flash drive</td>
</tr>
<tr>
<td>Distributed librarians</td>
<td>Authenticated users can modify collections on a central server using a remote version of the Librarian interface</td>
</tr>
<tr>
<td>Large scale</td>
<td>Existing collections have 10 M documents, 20 GB of raw text, 2 B words, 50 M unique terms, 50 GB metadata</td>
</tr>
<tr>
<td>Space efficient</td>
<td>Uses compression to reduce the size of indexes and text</td>
</tr>
<tr>
<td>License</td>
<td>Free open-source software issued under the GNU General Public License</td>
</tr>
<tr>
<td>Support</td>
<td>Mailing lists in several languages; support centers in India and Africa; several commercial support organizations exist</td>
</tr>
</tbody>
</table>
interesting information. Jon Miller kindly provided the photographs in Figure 1.1 (and Figure 9.3 in Chapter 9). The Humanity Development Library is produced by Michel Loots of the HumanInfo Project in Antwerp, Belgium, using the Greenstone software, and widely distributed in the developing world. The development of the physics archives is described by Paul Ginsparg, its originator, in a paper called “Winners and losers in the global research village” (Ginsparg, 1997); he is responsible for the memorable “sliced dead trees” metaphor. (Note incidentally that adsorbed is a physics term, not a misprint; it means the assimilation of dissolved matter by the surface of a solid.) Lloyd Smith at New Mexico Highlands University conceived of the Zia Pueblo project and came up with the vision described here; the Zia Pueblo kindly supplied the photograph in Figure 1.2. The digital music library is ongoing work in the Department of Computer Science at Waikato.

Our definition of digital library in Section 1.1 is abstracted from 10 definitions of the term digital library culled from the literature by Ed Fox (on the Web at http://ei.cs.vt.edu/~dlib/def.htm). It was the computer pioneer Maurice Wilkes who said that books would be hailed as a great advance if they were invented today. The “data … information … knowledge … wisdom” sequence and characterization is due to Harold Thimbleby at Swansea University.

A good source for the development of libraries is Thompson (1997), who formulates some principles of librarianship, including the one quoted, that it is a librarian’s duty to increase the stock of his library. Thompson is the source of some of the material with which Section 1.2 begins, including the metaphor about snapping the links of the chained book—in fact, he formulates open access as another principle: libraries are for all. The imaginative architectural developments that have occurred in physical libraries at the close of the 20th century are documented, and beautifully illustrated, by Wu (1999). Gore (1976b) recounts the fascinating history of the Alexandrian Library; he edited the book entitled Farewell to Alexandria (Gore, 1976a). The information on Trinity College, Dublin, was kindly supplied by David Abrahamson; that on the Library of Congress was retrieved from the Internet. Much of the other historical information is from an excellent and thought-provoking paper by Gaines that is well worth reading (Gaines, 1993). Thomas Mann (1993), a reference librarian at the Library of Congress, has produced a wonderful source of information on libraries and librarianship, full of practical assistance on how to use conventional library resources to find things.

H. G. Wells’s “world brain” idea was published in 1938 (Wells, 1938). Vannevar Bush’s vision was described in the year the United Nations was founded (Bush, 1945)—although certainly no connection between virtual libraries and the plight of developing countries was made in those days. Licklider’s vision dates from 1960 (Licklider, 1960), while the U.S. Presidential Office weighed in early in 1993 (Clinton and Gore, 1993).

The 19th-century librarian who “makes time for his fellow mortals” is Bowker (1883), quoted by Crawford and Gorman (1995). The modern term “metadata” is an impressive-sounding moniker, but the catchphrase “data about data” is glib but not very enlightening. In some sense, all data is about data: Where does one stop and the other begin? We return to this discussion at the beginning of Chapter 6. Meanwhile we continue to use the term freely, always in the context of digital library collections in which it is clear that the metadata is information about a particular resource.
More information on the promise of digital libraries in developing countries can be found in Witten et al. (2001). The figures on the Nairobi and Brazzaville universities are from the United Nations Human Development Report (United Nations, 1999), as is some of the information on Internet penetration in developing countries. Arunachalam (1998) tells how the Internet is “failing the developing world.” Statistics on the number of computers available in developing countries can be found in the World Bank’s (2000) World Development Indicators. Information on mobile phone penetration can be found in an International Telecommunication Union report (1999). There is much information on the “digital divide,” the widening knowledge gap between rich and poor: read the United Nations’ 1997 statement on Universal Access to Basic Communication and Information Services, the International Telecommunication Union’s 1998 World Telecommunication Development Report, or the World Bank’s 1998/99 World Development Report.

The Greenstone software is produced by the New Zealand Digital Library Project. The toki (adze) shown in Figure 1.11 was a gift from the Māori people in recognition of the digital library’s contributions to indigenous language preservation; it resides in the project laboratory at the University of Waikato. In Māori culture there are several kinds of toki, with different purposes. This one is a ceremonial adze, toki pou tangata, a symbol of chieftainship. The rau (blade) is sharp, hard, and made of pounamu or greenstone—hence the Greenstone software, at the cutting edge of digital library technology. There are three figures carved into the toki. The forward-looking one looks out to where the rau is pointing to ensure that the toki is appropriately targeted. The backward-looking one at the top is a sentinel that guards where the rau can’t see. There is a third head at the bottom of the handle that makes sure that the chief’s decisions—to which the toki lends authority—are properly grounded in reality. The name of this taonga, or art treasure, is Toki Pou Hinengaro, which translates roughly as “the adze that shapes the excellence of thought.” Haramai te toki, haumi e, hui e, tāiki e.

Turning now to copyright, Samuelson and Davis (2000) provide an excellent and thought-provoking overview of copyright and related issues in the information age, which is a synopsis of a larger report published by the National Academy of Sciences Press (Committee on Intellectual Property Rights, 2000). Section 1.5 draws heavily on this material. An earlier paper by Samuelson (1998) discusses specific digital library issues raised by copyright and intellectual property law, from a U.S. perspective. The Association for Computing Machinery has published a collection of papers on the effect of emerging technologies on intellectual property issues (White, 1999). We learned a lot from Lessig’s fine book Free Culture (Lessig, 2004), which has strongly influenced our perspective; the quotation is from p. 227. Lessig is also the originator of the Creative Commons, whose licenses can be found on the web—as can the GNU GFDL license.

There’s plenty of information on copyright on the Web. For example, http://scholar.lib.vt.edu/ copyright is a useful Web site developed by staff at Virginia Tech to share what they learned about policies and common practices that relate to copyright. It includes interpretations of U.S. copyright law, links to the text of the law, sample letters to request permission to use someone else’s work, links to publishers’ e-mail addresses, advice for authors about negotiating to retain some rights, as well as current library policies. Georgia Harper at the University of Texas at Austin has created an
excellent *Crash Course in Copyright* (www.utsystem.edu/ogc/intellectualproperty/cprtindx.htm) that is delightfully presented and well worth reading. The information about the duration of copyright protection is from Lolly Gasaway, director of the Law Library and professor of law at the University of North Carolina, through his Web site at www.unc.edu/~unclng/public-d.htm. The core reference for copyleft is at the GNU Project: http://www.gnu.org/copyleft.

The source of the quotation about the hypothetical college student is Kahle (1997), a *Scientific American* article describing an Internet archiving project that we return to in Chapters 5 and 9. Information on the UNESCO initiative to attack paedophilia on the Internet can be found in Chapter 8 of its *World Communication and Information Report 1999–2000*. Elke Duncker is a marvelous source of information on cultural sensitivity and user interfaces: some of her experiences with user interface issues in different cultures are described in Duncker (2002).

Standard library automation systems are described by Cooper (1996). The first book on digital libraries is *Practical Digital Libraries* by Lesk (1997), a pioneering work that gives a fascinating early

---

**Figure 1.11:** Māori *toki* or ceremonial adze, emblem of the Greenstone project
account of this emerging field. In contrast, Crawford and Gorman (1995) fear that virtual libraries are virtual nonsense that threatens to devastate the cultural mission of libraries. Chen (1998) describes the past, present, and future of digital libraries from his perspective as Program Director of the NSF/DARPA/NASA Digital Libraries Initiative in the United States from 1993 to 1995. Sanders (1999) offers an edited collection of papers that give a librarian’s perspective on many aspects of digital libraries and their use. Although Borgman’s (2000) title focuses on the global information infrastructure, most of her material is relevant to the kind of digital libraries discussed here. Arms (2000) gives an authoritative, comprehensive, and balanced account of digital libraries from many different perspectives. He includes a historical perspective, a survey of the state of the art, and an account of current research.