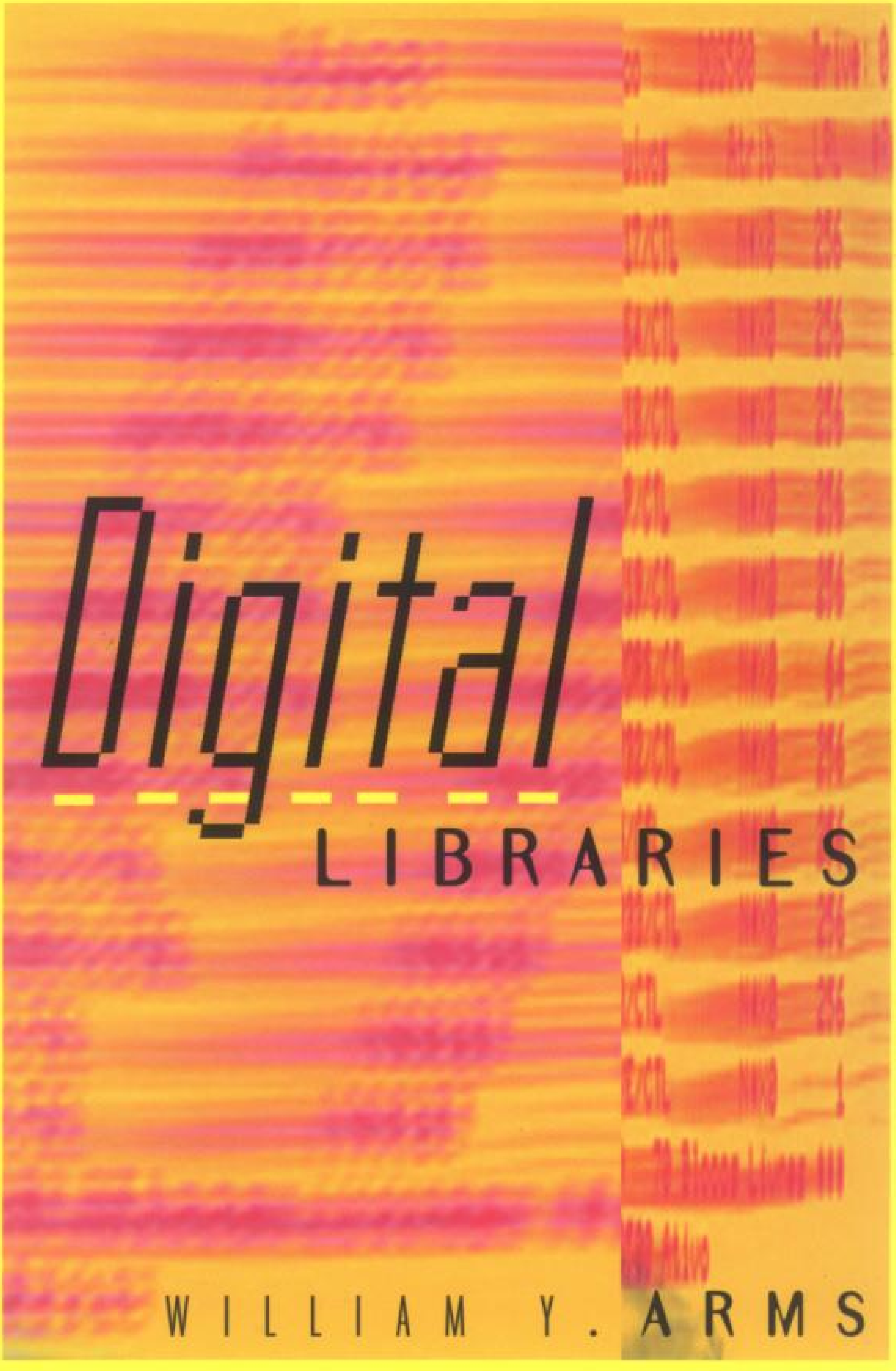


Digital LIBRARIES

WILLIAM Y. ARMS

Second printing, 2001

Set in Sabon by The MIT Press.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Arms, William Y.
Digital libraries / William Y. Arms.
p. cm. — (Digital libraries and electronic publishing)
Includes bibliographical references and index.
ISBN 0-262-01880-8 (alk. paper)
1. Libraries—United States—Special collections—Electronic information resources. 2. Digital libraries—United States.
I. Title. II. Series.
Z692.C65A76 2000
025'.00285—dc21 99-14773
CIP

Contents

Preface to the 1999 Edition

Four years ago, a group of us created *D-Lib Magazine*, a monthly periodical that has become the voice of digital library research and implementation. We started *D-Lib Magazine* because digital libraries are bringing together people from numerous disciplines who know little about each other. Our aim was to create a magazine that would inform people of the vast array of expertise that is feeding this field. Computer scientists are often unaware of the deep understanding of information that librarians have developed over the years. Librarians and publishers may not know that the Internet pioneers have been managing online information for decades. Both communities are vaguely aware that their fields are subject to external economic and legal forces, but have only limited knowledge of the implications.

To scan *D-Lib Magazine* over the first few years shows how prescient this vision was. Articles range from reports on the latest metadata workshop, to implementation projects at the Library of Congress, user interface principles applied to search and retrieval, a historic review of Z39.50, a digital library of Japanese folk tales, the JSTOR approach to legal issues, or a description of methods for handling alternate character sets. Nobody is an expert in all these areas, yet to be a leader in digital libraries requires some appreciation of all of them.

This book is my attempt to survey the entire field of digital libraries. Computers and networks are of fundamental importance, but they are only the technology. The real story of digital libraries is the interplay between people, organizations, and technology. How are libraries and publishers using this new technology? How are individuals bypassing traditional organizations and building their own libraries? Where is this all leading? The answer to the last question is simple. Nobody knows. I have tried to avoid speculation and to concentrate on describing current activities, trends, and research. Thus the heart of this book is a large number of examples described in panels. Each panel describes some significant aspect of digital libraries, technology, application, or research.

However, I have to admit to personal biases. Some are undoubtedly unconscious, but others are quite deliberate. I am definitely biased towards digital libraries that provide open access to information. As a reader, I am thrilled by the high-quality information which I can access over the Internet that is completely unrestricted; as an author, I publish my research online so that everybody has access to my work. Technically, my bias is towards simplicity. I am a great fan of the web, because of it is so simple. This simplicity is an enormous strength and I hope that we can defend it.

My career has given me first hand exposure to many of the topics described in this book. In selecting examples, I have usually chosen those that I know personally, with an emphasis on work carried out by friends and colleagues. Therefore, the examples reflect my experience, working in universities in the United States. As an Englishman by birth, I am conscious of the quality of work that is being carried out around the world, but my examples tend to be American.

Because I know so many of the people whose work is described in this book, I have been shameless in asking them for help. Amy Friedlander, the founding editor of *D-Lib Magazine*, has been a constant guide. My wife Caroline introduced me to digital libraries when she was a graduate student at M.I.T. in 1966. She is now at the Library of Congress and helped with many sections throughout the book. Individuals who have made comments on the manuscript or checked specific sections include: Robert Allen, Kate Arms, Steve Cousins, Gregory Crane, Jim Davis, Peter Denning, Jack Dongarra, George Furnas, Henry Gladney, Steven Griffin, Kevin Guthrie, Larry Lannom, Ron Larsen, Michael Lesk, Ralph LeVan, Mary Levering, Wendy Lougee, Clifford Lynch, Harry S. Martin III, Eric Miller, Andreas Paepcke, Larry Page, Norman Paskin, Vicky Reich, Scott Stevens, Terrence Smith, Sam Sun, Hal Varian, Howard Wactlar, Donald Waters, Stuart Weibel, and Robert Wilensky.

Acknowledgements

Earlier versions of Chapter 4 were provided to the Fundação Getulio Vargas and the Ticer Summer School. A first sketch of Chapter 8 was presented at the SGML/XML Conference in 1997. The image in Panel 2.6 is copyright CNRI. Panel 6.3 is from the U.S. Copyright Office. Figure 7.2 is based on a figure by Henry Gladney. The list of elements in Panel 10.3 is from the Dublin Core web site.

Chapter 1
Background

An introduction to digital libraries

This is a fascinating period in the history of libraries and publishing. For the first time, it is possible to build large-scale services where collections of information are stored in digital formats and retrieved over networks. The materials are stored on computers. A network connects the computers to personal computers on the users' desks. In a completely digital library, nothing need ever reach paper.

This book provides an overview of this new field. Partly it is about technology, but equally it is about people and organizations. Digital libraries bring together facets of many disciplines, and experts with different backgrounds and different approaches. The book describes the contributions of these various disciplines and how they interact. It discusses the people who create information and the people who use it, their needs, motives, and economic incentives. It analyzes the profound changes that are occurring in publishing and libraries. It describes research into new technology, much of it based on the Internet and the World Wide Web. The topics range from technical aspects of computers and networks, through librarianship and publishing, to economics and law. The constant theme is change, with its social, organizational, and legal implications.

One book can not cover all these topics in depth, and much has been left out or described at an introductory level. Most of the examples come from the United States, with prominence given to universities and the academic community, but the development of digital libraries is world-wide with contributions from many sources. Specialists in big American universities are not the only developers of digital libraries, though they are major contributors. There is a wealth and diversity of innovation in almost every discipline, in countries around the world.

People

An informal definition of a digital library is a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network. A key part of this definition is that the information is managed. A stream of data sent to earth from a satellite is not a library. The same data, when organized systematically, becomes a digital library collection. Most people would not consider a database containing financial records of one company to be a digital library, but would accept a collection of such information from many companies as part of a library. Digital libraries contain diverse collections of information for use by many different users. Digital libraries range in size from tiny to huge. They can use any type of computing equipment and any suitable software. The unifying theme is that information is organized on computers and available over a network, with procedures to select the material in the collections, to organize it, to make it available to users, and to archive it.

In some ways, digital libraries are very different from traditional libraries, yet in others they are remarkably similar. People do not change because new technology is invented. They still create information that has to be organized, stored, and distributed. They still need to find information that others have created, and use it for study, reference, or entertainment. However, the form in which the information is expressed and the methods that are used to manage it are greatly influenced by technology and this creates change. Every year, the quantity and variety of collections available in digital form grows, while the supporting technology continues to improve steadily. Cumulatively, these changes are stimulating fundamental alterations in how people create information and how they use it.

To understand these forces requires an understanding of the people who are developing the libraries. Technology has dictated the pace at which digital libraries have been able to develop, but the manner in which the technology is used depends upon people. Two important communities are the source of much of this innovation. One group is the information professionals. They include librarians, publishers, and a wide range of information providers, such as indexing and abstracting services. The other community contains the computer science researchers and their offspring, the Internet developers. Until recently, these two communities had disappointingly little interaction; even now it is commonplace to find a computer scientist who knows nothing of the basic tools of librarianship, or a librarian whose concepts of information retrieval are years out of date. Over the past few years, however, there has been much more collaboration and understanding.

Partly this is a consequence of digital libraries becoming a recognized field for research, but an even more important factor is greater involvement from the users themselves. Low-cost equipment and simple software have made electronic information directly available to everybody. Authors no longer need the services of a publisher to distribute their works. Readers can have direct access to information without going through an intermediary. Many exciting developments come from academic or professional groups who develop digital libraries for their own needs. Medicine has a long tradition of creative developments; the pioneering legal information systems were developed by lawyers for lawyers; the web was initially developed by physicists, for their own use.

Economics

Technology influences the economic and social aspects of information, and vice versa. The technology of digital libraries is developing fast and so are the financial, organizational, and social frameworks. The various groups that are developing digital libraries bring different social conventions and different attitudes to money. Publishers and libraries have a long tradition of managing physical objects, notably books, but also maps, photographs, sound recordings and other artifacts. They evolved economic and legal frameworks that are based on buying and selling these objects. Their natural instinct is to transfer to digital libraries the concepts that have served them well for physical artifacts. Computer scientists and scientific users, such as physicists, have a different tradition. Their interest in digital information began in the days when computers were very expensive. Only a few well-funded researchers had computers on the first networks. They exchanged information informally and openly with colleagues, without payment. The networks have grown, but the tradition of open information remains.

The economic framework that is developing for digital libraries shows a mixture of these two approaches. Some digital libraries mimic traditional publishing by requiring a form of payment before users may access the collections and use the services. Other digital libraries use a different economic model. Their material is provided with open access to everybody. The costs of creating and distributing the information are borne by the producer, not the user of the information. This book describes many examples of both models and attempts to analyze the balance between them. Almost certainly, both have a long-term future, but the final balance is impossible to forecast.

Why digital libraries?

The fundamental reason for building digital libraries is a belief that they will provide better delivery of information than was possible in the past. Traditional libraries are a fundamental part of society, but they are not perfect. Can we do better?

Enthusiasts for digital libraries point out that computers and networks have already changed the ways in which people communicate with each other. In some disciplines, they argue, a professional or scholar is better served by sitting at a personal computer connected to a communications network than by making a visit to a library. Information that was previously available only to the professional is now directly available to all. From a personal computer, the user is able to consult materials that are stored on computers around the world. Conversely, all but the most diehard enthusiasts recognize that printed documents are so much part of civilization that their dominant role cannot change except gradually. While some important uses of printing may be replaced by electronic information, not everybody considers a large-scale movement to electronic information desirable, even if it is technically, economically, and legally feasible.

Here are some of the potential benefits of digital libraries.

- **The digital library brings the library to the user**

  To use a library requires access. Traditional methods require that the user goes to the library. In a university, the walk to a library takes a few minutes, but not many people are member of universities or have a nearby library. Many engineers or physicians carry out their work with depressingly poor access to the latest information.

  A digital library brings the information to the user's desk, either at work or at home, making it easier to use and hence increasing its usage. With a digital library on the desk top, a user need never visit a library building. The library is wherever there is a personal computer and a network connection.

- **Computer power is used for searching and browsing**

  Computing power can be used to find information. Paper documents are convenient to read, but finding information that is stored on paper can be difficult. Despite the myriad of secondary tools and the skill of reference librarians, using a large library can be a tough challenge. A claim that used to be made for traditional libraries is that they stimulate serendipity, because readers stumble across unexpected items of value. The truth is that libraries are full of useful materials that readers discover only by accident.

  In most aspects, computer systems are already better than manual methods for finding information. They are not as good as everybody would like, but they are good and improving steadily. Computers are particularly useful for reference work that involves repeated leaps from one source of information to another.

- **Information can be shared**

  Libraries and archives contain much information that is unique. Placing digital information on a network makes it available to everybody. Many digital libraries or electronic publications are maintained at a single central site, perhaps with a few duplicate copies strategically placed around the world. This is a vast improvement over expensive physical duplication of little used material, or the inconvenience of unique material that is inaccessible without traveling to the location where it is stored.

- **Information is easier to keep current**

  Much important information needs to be brought up to date continually. Printed materials are awkward to update, since the entire document must be reprinted; all copies of the old version must be tracked down and replaced. Keeping information current is much less of a problem when the definitive version is in digital format and stored on a central computer.

  Many libraries provide online the text of reference works, such as directories or encyclopedias. Whenever revisions are received from the publisher, they are installed on the library's computer. The new versions are available immediately. The Library of Congress has an online collection, called Thomas, that contains the latest drafts of all legislation currently before the U.S. Congress; it changes continually.

- **The information is always available**

  The doors of the digital library never close; a recent study at a British university found that about half the usage of a library's digital collections was at hours when the library buildings were closed. Materials are never checked out to other readers, miss-shelved or stolen; they are never in an off-campus warehouse. The scope of the collections expands beyond the walls of the library. Private papers in an office or the collections of a library on the other side of the world are as easy to use as materials in the local library.

  Digital libraries are not perfect. Computer systems can fail and networks may be slow or unreliable, but, compared with a traditional library, information is much more likely to be available when and where the user wants it.

- **New forms of information become possible**

  Most of what is stored in a conventional library is printed on paper, yet print is not always the best way to record and disseminate information. A database may be the best way to store census data, so that it can be analyzed by computer; satellite data can be rendered in many different ways; a mathematics library can store mathematical expressions, not as ink marks on paper but as computer symbols to be manipulated by programs such as Mathematica or Maple.

  Even when the formats are similar, materials that are created explicitly for the digital world are not the same as materials originally designed for paper or other media. Words that are spoken have a different impact from words that are written, and online textual materials are subtly different from either the

> spoken or printed word. Good authors use words differently when they write for different media and users find new ways to use the information. Materials created for the digital world can have a vitality that is lacking in material that has been mechanically converted to digital formats, just as a feature film never looks quite right when shown on television.

Each of the benefits described above can be seen in existing digital libraries. There is another group of potential benefits, which have not yet been demonstrated, but hold tantalizing prospects. The hope is that digital libraries will develop from static repositories of immutable objects to provide a wide range of services that allow collaboration and exchange of ideas. The technology of digital libraries is closely related to the technology used in fields such as electronic mail and teleconferencing, which have historically had little relationship to libraries. The potential for convergence between these fields is exciting.

The cost of digital libraries

The final potential benefit of digital libraries is cost. This is a topic about which there has been a notable lack of hard data, but some of the underlying facts are clear.

Conventional libraries are expensive. They occupy expensive buildings on prime sites. Big libraries employ hundreds of people - well-educated, though poorly paid. Libraries never have enough money to acquire and process all the materials they desire. Publishing is also expensive. Converting to electronic publishing adds new expenses. In order to recover the costs of developing new products, publishers sometimes even charge more for a digital version than the printed equivalent.

Today's digital libraries are also expensive, initially more expensive. However, digital libraries are made from components that are declining rapidly in price. As the cost of the underlying technology continues to fall, digital libraries become steadily less expensive. In particular, the costs of distribution and storage of digital information declines. The reduction in cost will not be uniform. Some things are already cheaper by computer than by traditional methods. Other costs will not decline at the same rate or may even increase. Overall, however, there is a great opportunity to lower the costs of publishing and libraries.

Lower long-term costs are not necessarily good news for existing libraries and publishers. In the short term, the pressure to support traditional media alongside new digital collections is a heavy burden on budgets. Because people and organizations appreciate the benefits of online access and online publishing, they are prepared to spend an increasing amount of their money on computing, networks, and digital information. Most of this money, however, is going not to traditional libraries, but to new areas: computers and networks, web sites and webmasters.

Publishers face difficulties because the normal pricing model of selling individual items does not fit the cost structure of electronic publishing. Much of the cost of conventional publishing is in the production and distribution of individual copies of books, photographs, video tapes, or other artifacts. Digital information is different. The fixed cost of creating the information and mounting it on a computer may be substantial, but the cost of using it is almost zero. Because the marginal cost is negligible, much of the information on the networks has been made openly available, with no access restrictions. Not everything on the world's networks is freely available, but a great deal is open to everybody, undermining revenue for the publishers.

These pressures are inevitably changing the economic decisions that are made by authors, users, publishers, and libraries. Chapter 6 explores some of these financial considerations; the economics of digital information is a theme that recurs throughout the book.

Panel 1.1
Two Pioneers of Digital Libraries

The vision of the digital library is not new. This is a field in which progress is been achieved by the incremental efforts of numerous people over a long period of time. However, a few authors stand out because their writings have inspired future generations. Two of them are Vannevar Bush and J. C. R. Licklider.

As We May Think

In July 1945, Vannevar Bush, who was then director of the U. S. Office of Scientific Research and Development, published an article in *The Atlantic Monthly*, entitled "As We May Think". This article is an elegantly written exposition of the potential that technology offers the scientist to gather, store, find, and retrieve information. Much of his analysis rings as true today as it did fifty years ago.

Bush commented that, "our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purpose." He discussed recent technological advances and how they might conceivably be applied at some distant time in the future. He provided an outline of one possible technical approach, which he called Memmex. An interesting historical footnote is that the Memmex design used photography to store information. For many years, microfilm was the technology perceived as the most suitable for storing information cheaply.

Bush is often cited as the first person to articulate the new vision of a library, but that is incorrect. His article built on earlier work, much of it carried out in Germany before World War II. The importance of his article lies in its wonderful exposition of the inter-relationship between information and scientific research, and in the latent potential of technology.

The original article was presumably read only by those few people who happened to see that month's edition of the magazine. Now *The Atlantic Monthly* has placed a copy of the paper on its web site for the world to see. Everybody interested in libraries or scientific information should read it.

Libraries of the Future

In the 1960s, J. C. R. Licklider was one of several people at the Massachusetts Institute of Technology who studied how digital computing could transform libraries. As with Bush, Licklider's principal interest was the literature of science, but with the emergence of modern computing, he could see many of the trends that have subsequently occurred.

In his book, *The Library of the Future*, Licklider described the research and development needed to build a truly usable digital library. When he wrote, time-shared computing was still in the research laboratory, and computer memory cost a dollar a byte, but he made a bold attempt to predict what a digital library might be like thirty years later, in 1994. His predictions proved remarkably accurate in their overall vision, though naturally he did not foretell every change that has happened in thirty years. In general, he under-estimated how much would be achieved by brute force methods, using huge amounts of cheap computer power, and over-estimated how much progress could be made from artificial intelligence and improvements in computer methods of natural language processing.

Licklider's book is hard to find and less well-known than it should be. It is one of the few important documents about digital libraries that is not available on the Internet.

Technical developments

The first serious attempts to store library information on computers date from the late 1960s. These early attempts faced serious technical barriers, including the high cost of computers, terse user interfaces, and the lack of networks. Because storage was expensive, the first applications were in areas where financial benefits could be gained from storing comparatively small volumes of data online. An early success was the work of the Library of Congress in developing a format for Machine-Readable Cataloguing (MARC) in the late 1960s. The MARC format was used by the Online Computer Library Center (OCLC) to share catalog records among many libraries. This resulted in large savings in costs for libraries.

Early information services, such as shared cataloguing, legal information systems, and the National Library of Medicine's Medline service, used the technology that existed when they were developed. Small quantities of information were mounted on a large central computer. Users sat at a dedicated terminal, connected by a low-speed communications link, which was either a telephone line or a special purpose network. These systems required a trained user who would accept a cryptic user interface in return for faster searching than could be carried out manually and access to information that was not available locally.

Such systems were no threat to the printed document. All that could be displayed was unformatted text, usually in a fixed spaced font, without diagrams, mathematics, or the graphic quality that is essential for easy reading. When these weaknesses were added to the inherent defects of early computer screens - poor contrast and low resolution - it is hardly surprising that most people were convinced that users would never willingly read from a screen.

The past thirty years have steadily eroded these technical barriers. During the early 1990s, a series of technical developments took place that removed the last fundamental barriers to building digital libraries. Some of this technology is still rough and ready, but low-cost computing has stimulated an explosion of online information services. Four technical areas stand out as being particularly important to digital libraries.

- **Electronic storage is becoming cheaper than paper**

  Large libraries are painfully expensive for even the richest organizations. Buildings are about a quarter of the total cost of most libraries. Behind the collections of many great libraries are huge, elderly buildings, with poor environmental control. Even when money is available, space for expansion is often hard to find in the center of a busy city or on a university campus.

  The costs of constructing new buildings and maintaining old ones to store printed books and other artifacts will only increase with time, but electronic storage costs decrease by at least 30 percent per annum. In 1987, we began work on a digital library at Carnegie Mellon University, known as the Mercury library. The collections were stored on computers, each with ten gigabytes of disk storage. In 1987, the list price of these computers was about $120,000. In 1997, a much more powerful computer with the same storage cost about $4,000. In ten years, the price was reduced by about 97 percent. Moreover, there is every reason to believe that by 2007 the equipment will be reduced in price by another 97 percent.

Ten years ago, the cost of storing documents on CD-ROM was already less than the cost of books in libraries. Today, storing most forms of information on computers is much cheaper than storing artifacts in a library. Ten years ago, equipment costs were a major barrier to digital libraries. Today, they are much lower, though still noticeable, particularly for storing large objects such as digitized videos, extensive collections of images, or high-fidelity sound recordings. In ten years time, equipment that is too expensive to buy today will be so cheap that the price will rarely be a factor in decision making.

- **Personal computer displays are becoming more pleasant to use**

Storage cost is not the only factor. Otherwise libraries would have standardized on microfilm years ago. Until recently, few people were happy to read from a computer. The quality of the representation of documents on the screen was too poor. The usual procedure was to print a paper copy. Recently, however, major advances have been made in the quality of computer displays, in the fonts which are displayed on them, and in the software that is used to manipulate and render information. People are beginning to read directly from computer screens, particularly materials that were designed for computer display, such as web pages. The best computers displays are still quite expensive, but every year they get cheaper and better. It will be a long time before computers match the convenience of books for general reading, but the high-resolution displays to be seen in research laboratories are very impressive indeed.

Most users of digital libraries have a mixed style of working, with only part of the materials that they use in digital form. Users still print materials from the digital library and read the printed version, but every year more people are reading more materials directly from the screen.

- **High-speed networks are becoming widespread**

The growth of the Internet over the past few years has been phenomenal. Telecommunications companies compete to provide local and long distance Internet service across the United States; international links reach almost every country in the world; every sizable company has its internal network; universities have built campus networks; individuals can purchase low-cost, dial-up services for their homes.

The coverage is not universal. Even in the U.S. there are many gaps and some countries are not yet connected at all, but in many countries of the world it is easier to receive information over the Internet than to acquire printed books and journals by orthodox methods.

- **Computers have become portable**

  Although digital libraries are based around networks, their utility has been greatly enhanced by the development of portable, laptop computers. By attaching a laptop computer to a network connection, a user combines the digital library resources of the Internet with the personal work that is stored on the laptop. When the user disconnects the laptop, copies of selected library materials can be retained for personal use.

  During the past few years, laptop computers have increased in power, while the quality of their screens has improved immeasurably. Although batteries remain a problem, laptops are no heavier than a large book, and the cost continues to decline steadily.

Access to digital libraries

Traditional libraries usually require that the user be a member of an organization that maintains expensive physical collections. In the United States, universities and some other organizations have excellent libraries, but most people do not belong to such an organization. In theory, much of the Library of Congress is open to anybody over the age of eighteen, and a few cities have excellent public libraries, but in practice, most people are restricted to the small collections held by their local public library. Even scientists often have poor library facilities. Doctors in large medical centers have excellent libraries, but those in remote locations typically have nothing. One of the motives that led the Institute of Electrical and Electronics Engineers (IEEE) to its early interest in electronic publishing was the fact that most engineers do not have access to an engineering library.

Users of digital libraries need a computer attached to the Internet. In the United States, many organizations provide every member of staff with a computer. Some have done so for many years. Across the nation, there are programs to bring computers to schools and to install them in pubic libraries. For individuals who must provide their own computing, adequate access to the Internet requires less than $2,000 worth of equipment, perhaps $20 per month for a dial-up connection, and a modicum of skill. Increase the costs a little and very attractive services can be obtained, with a powerful computer and a dedicated, higher speed connection. These are small investments for a prosperous professional, but can be a barrier for others. In 1998 it was estimated that 95 percent of people in the United States live in areas where there is reasonable access to the Internet. This percentage is growing rapidly.

Outside the United States, the situation varies. In most countries of the world, library services are worse than in the United States. For example, universities in Mexico report that reliable delivery of scholarly journals is impossible, even when funds are available. Some nations are well-supplied with computers and networks, but in most places equipment costs are higher than in the United States, people are less wealthy, monopolies keep communications costs high, and the support infrastructure is lacking. Digital libraries do bring information to many people who lack traditional libraries, but the Internet is far from being conveniently accessible world-wide.

A factor that must be considered in planning digital libraries is that the quality of the technology available to users varies greatly. A favored few have the latest personal computers on their desks, high-speed connections to the Internet, and the most recent release of software; they are supported by skilled staff who can configure and tune the equipment, solve problems, and keep the software up to date. Most people, however, have to make do with less. Their equipment may be old, their software out of date, their Internet connection troublesome, and their technical support from staff who are under-trained and over-worked. One of the great challenges in developing digital libraries is to build systems that take advantage of modern technology, yet perform adequately in less perfect situations.

Basic concepts and terminology

Terminology often proves to be a barrier in discussing digital libraries. The people who build digital libraries come from many disciplines and bring the terminology of those disciplines with them. Some words have such strong social, professional, legal, or technical connotations that they obstruct discussion between people of varying backgrounds. Simple words mean different things to different people. For example, the words "copy" and "publish" have different meanings to computing professionals, publishers, and lawyers. Common English usage is not the same as professional usage, the versions of English around the world have subtle variations of meaning, and discussions of digital libraries are not restricted to the English language.

Some words cause such misunderstandings that it is tempting to ban them from any discussion of digital libraries. In addition to "copy" and "publish", the list includes "document", "object", and "work". At the very least, such words must be used carefully and their exact meaning made clear whenever they are used. This book attempts to be precise when precision is needed. For example, in certain contexts the distinction must be made between "photograph" (an image on paper), and "digitized photograph" (a set of bits in a computer). Most of the time, however, such precision is mere pedantry. Where the context is clear, the book uses terms informally. Where the majority of the practitioners in the field use a word in certain way, their usage is followed.

Collections

Digital libraries hold any information that can be encoded as sequences of bits. Sometimes these are digitized versions of conventional media, such as text, images, music, sound recordings, specifications and designs, and many, many more. As digital libraries expand, the contents are less often the digital equivalents of physical items and more often items that have no equivalent, such as data from scientific instruments, computer programs, video games, and databases.

- **Data and metadata**

  The information stored in a digital library can be divided into data and metadata. Data is a general term to describe information that is encoded in digital form. Whether the word "data" is singular or plural is a source of contention. This book treats the word as a singular collective noun, following the custom in computing.

  Metadata is data about other data. Many people dislike the word "metadata", but it is widely used. Common categories of metadata include **descriptive metadata**, such as bibliographic information, **structural metadata** about formats and structures, and **administrative metadata**, which includes, rights, permissions, and other information that is used to manage access. One item of metadata is the **identifier**, which identifies an item to the outside world.

The distinction between data and metadata often depends upon the context. Catalog records or abstracts are usually considered to be metadata, because they describe other data, but in an online catalog or a database of abstracts they are the data.

- **Items in a digital library**

No generic term has yet been established for the items that are stored in a digital library. In this book, several terms are used. The most general is material, which is anything that might be stored in a library. The word item is essentially synonymous. Neither word implies anything about the content, structure, or the user's view of the information. The word can be used to describe physical objects or information in digital formats. The term **digital material** is used when needed for emphasis. A more precise term is **digital object**. This is used to describe an item as stored in a digital library, typically consisting of data, associated metadata, and an identifier.

Some people call every item in a digital library a **document**. This book reserves the term for a digitized text, or for a digital object whose data is the digital equivalent of a physical document.

- **Library objects**

The term library object is useful for the user's view of what is stored in a library. Consider an article in an online periodical. The reader thinks of it as a single entity, a library object, but the article is probably stored on a computer as several separate objects. They contain pages of digitized text, graphics, perhaps even computer programs, or linked items stored on remote computers. From the user's viewpoint, this is one library object made up of several digital objects.

This example shows that library objects have internal structure. They usually have both data and associated metadata. Structural metadata is used to describe the formats and the relationship of the parts. This is a topic of Chapter 12.

- **Presentations, disseminations, and the stored form of a digital object**

The form in which information is stored in a digital library may be very different from the form in which it is used. A simulator used to train airplane pilots might be stored as several computer programs, data structures, digitized images, and other data. This is called the **stored form** of the object.

The user is provided with a series of images, synthesized sound, and control sequences. Some people use the term **presentation** for what is presented to the user and in many contexts this is appropriate terminology. A more general term is **dissemination**, which emphasizes that the transformation from the stored form to the user requires the execution of some computer program.

When digital information is received by a user's computer, it must be converted into the form that is provided to the user, typically by displaying on the computer screen, possibly augmented by a sound track or other presentation. This conversion is called **rendering**.

- **Works and content**

  Finding terminology to describe content is especially complicated. Part of the problem is that the English language is very flexible. Words have varying meanings depending upon the context. Consider, the example, "the song *Simple Gifts*". Depending on the context, that phrase could refer to the song as a work with words and music, the score of the song, a performance of somebody singing it, a recording of the performance, an edition of music on compact disk, a specific compact disc, the act of playing the music from the recording, the performance encoded in a digital library, and various other aspects of the song. Such distinctions are important to the music industry, because they determine who receives money that is paid for a musical performance or recording.

  Several digital library researchers have attempted to define a general hierarchy of terms that can be applied to all works and library objects. This is a bold and useful objective, but fraught with difficulties. The problem is that library materials have so much variety that a classification may match some types of material well but fail to describe others adequately.

  Despite these problems, the words work and content are useful words. Most people use the word **content** loosely, and this book does the same. The word is used in any context when the emphasis is on library materials, not as bits and bytes to be processed by a computer but as information that is of interest to a user. To misquote a famous judge, we can not define "content", but we know it when we see it.

  While the word content is used as a loosely defined, general term, the word work is used more specifically. The term "literary work" is carefully defined in U. S. copyright law as the abstract content, the sequence of words or music independent of any particular stored representation, presentation, or performance. This book usually uses the word "work" roughly with this meaning, though not always with legal precision.

People

A variety of words are used to describe the people who are associated with digital libraries. One group of people are the **creators** of information in the library. Creators include authors, composers, photographers, map makers, designers, and anybody else who creates intellectual works. Some are professionals; some are amateurs. Some work individually, others in teams. They have many different reasons for creating information.

Another group are the **users** of the digital library. Depending on the context, users may be described by different terms. In libraries, they are often called "readers" or "patrons"; at other times they may be called the "audience", or the "customers". A characteristic of digital libraries is that creators and users are sometimes the same people. In academia, scholars and researchers use libraries as resources for their research, and publish their findings in forms that become part of digital library collections.

The final group of people is a broad one that includes everybody whose role is to support the creators and the users. They can be called **information managers**. The group includes computer specialists, librarians, publishers, editors, and many others.

The World Wide Web has created a new profession of webmaster. Frequently a publisher will represent a creator, or a library will act on behalf of users, but publishers should not be confused with creators, or librarians with users. A single individual may be creator, user, and information manager.

Computers and networks

Digital libraries consists of many computers united by a communications network. The dominant network is the **Internet**, which is discussed Chapter 2. The emergence of the Internet as a flexible, low-cost, world-wide network has been one of the key factors that has led to the growth of digital libraries.

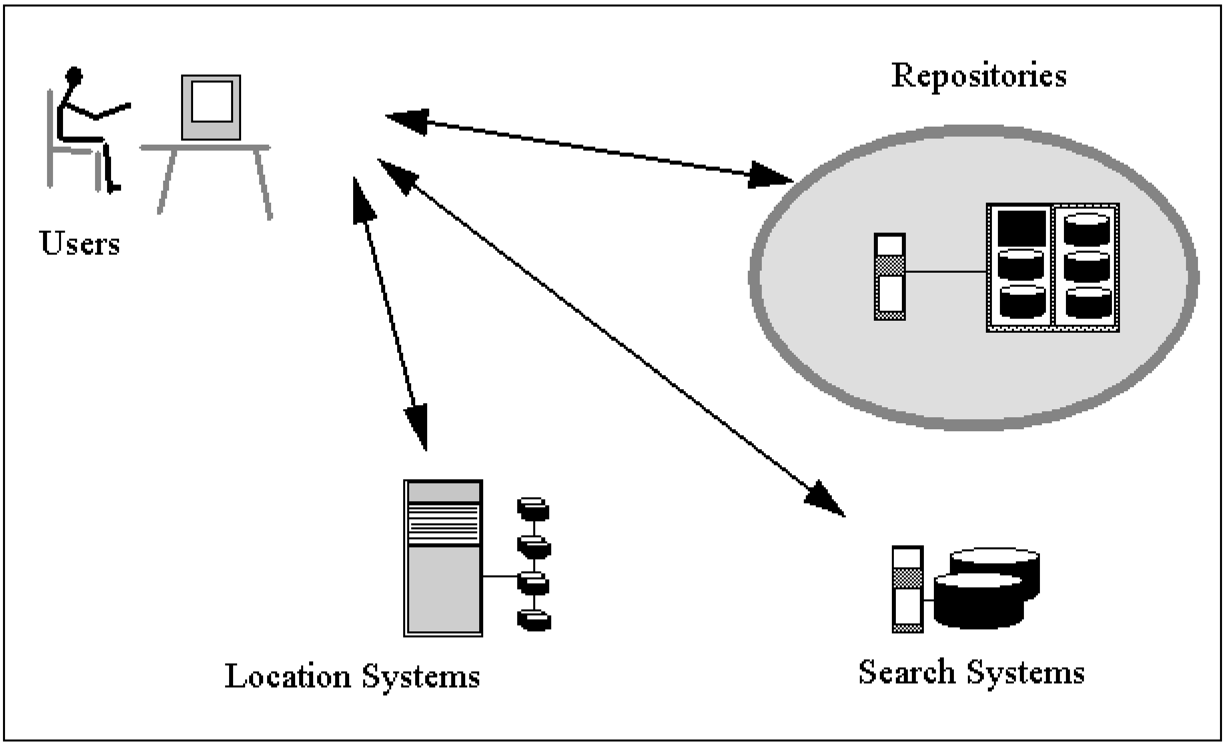


Figure 1.1. Computers in digital libraries

Figure 1.1 shows some of the computers that are used in digital libraries. The computers have three main function: to help users interact with the library, to store collections of materials, and to provide services.

- In the terminology of computing, anybody who interacts with a computer is called a **user** or **computer user**. This is a broad term that covers creators, library users, information professionals, and anybody else who accesses the computer. To access a digital library, users normally use personal computers. These computers are given the general name **clients**. Sometimes, clients may interact with a digital library without no human user involved, such as the robots that automatically index library collections, and sensors that gather data, such as information about the weather, and supply it to digital libraries.
- The next major group of computers in digital libraries are **repositories** which store collections of information and provide access to them. An **archive** is a repository that is organized for long-term preservation of materials.

- The figure shows two typical **services** which are provided by digital libraries: location systems and search systems. **Search systems** provide catalogs, indexes, and other services to help users find information. **Location systems** are used to identify and locate information.

- In some circumstances there may be other computers that sit between the clients and computers that store information. These are not shown in the figure. **Mirrors** and **caches** store duplicate copies of information, for faster performance and reliability. The distinction between them is that mirrors replicate large sets of information, while caches store recently used information only. **Proxies** and **gateways** provide bridges between different types of computer system. They are particularly useful in reconciling systems that have conflicting technical specifications.

The generic term **server** is used to describe any computer other than the user's personal computer. A single server may provide several of the functions listed above, perhaps acting as a repository, search system, and location system. Conversely, individual functions can be distributed across many servers. For example, the domain name system, which is a locator system for computers on the Internet, is a single, integrated service that runs on thousands of separate servers.

In computing terminology, a **distributed system** is a group of computers that work as a team to provide services to users. Digital libraries are some of the most complex and ambitious distributed systems ever built. The personal computers that users have on their desks have to exchange messages with the server computers; these computers are of every known type, managed by thousands of different organizations, running software that ranges from state-of-the art to antiquated. The term **interoperability** refers to the task of building coherent services for users, when the individual components are technically different and managed by different organizations. Some people argue that all technical problems in digital libraries are aspects of this one problem, interoperability. This is probably an overstatement, but it is certainly true that interoperability is a fundamental challenge in all aspects of digital libraries.

The challenge of change

If digital technology is so splendid, what is stopping every library immediately becoming entirely digital? Part of the answer is that the technology of digital libraries is still immature, but the challenge is much more than technology. An equal challenge is the ability of individuals and organizations to devise ways that use technology effectively, to absorb the inevitable changes, and to create the required social frameworks. The world of information is like a huge machine with many participants each contributing their experience, expertise, and resources. To make fundamental changes in the system requires inter-related shifts in the economic, social and legal relationships amongst these parties. These topics are studied in Chapters 5 and 6, but the underlying theme of social change runs throughout the book.

Digital libraries depend on people and can not be introduced faster than people and organizations can adapt. This applies equally to the creators, users, and the professionals who support them. The relationships amongst these groups are changing. With digital libraries, readers are more likely to go directly to information, without visiting a library building or having any contact with a professional intermediary. Authors carry out more of the preparation of a manuscript. Professionals need new skills and new training to support these new relationships. Some of these skills are absorbed through experience, while others can be taught. Since librarians have a career path based around schools of librarianship, these schools are adapting their curriculum, but it will be many years before the changes work through the system. The traditions of hundreds of years go deep.

The general wisdom is that, except in a few specialized areas, digital libraries and conventional collections are going to coexist for the foreseeable future. Institutional libraries will maintain large collections of traditional materials in parallel with their digital services, while publishers will continue to have large markets for their existing products. This does not imply that the organizations need not change, as new services extend the old. The full deployment of digital libraries will require extensive reallocation of money, with funds moving from the areas where savings are made to the areas that incur increased cost. Within an institution, such reallocations are painful to achieve, though they will eventually take place, but some of the changes are on a larger scale.

When a new and old technology compete, the new technology is never an exact match. Typically, the new has some features that are not in the old, but lacks some basic characteristics of the old. Therefore the old and new usually exist along side. However, the spectacular and continuing decline in the cost of computing with the corresponding increase in capabilities sometimes leads to complete substitution. Word processors were such an improvement that they supplanted typewriters in barely ten years. Card catalogs in libraries are on the same track. In 1980, only a handful of libraries could afford an online catalog. Twenty years later, a card catalog is becoming a historic curiosity in American libraries. In some specialized areas, digital libraries may completely replace conventional library materials.

Since established organizations have difficulties changing rapidly, many exciting developments in digital libraries have been introduced by new organizations. New organizations can begin afresh, but older organizations are faced with the problems of maintaining old services while introducing the new. The likely effect of digital libraries will be a massive transfer of money from traditional suppliers of information to new information entrepreneurs and to the computing industry. Naturally, existing organizations will try hard to discourage any change in which their importance diminishes, but the economic relationships between the various parties are already changing. Some important organizations will undoubtedly shrink in size or even go out of business. Predicting these changes is made particularly difficult by uncertainties about the finances of digital libraries and electronic publishing, and by the need for the legal system to adapt. Eventually, the pressures of the marketplace will establish a new order. At some stage, the market will have settled down sufficiently for the legal rules to be clarified. Until then, economic and legal uncertainties are annoying, though they have not proved to be serious barriers to progress.

Overall, there appear to be no barriers to digital libraries and electronic publishing. Technical, economic, social, and legal challenges abound, but they are being overcome steadily. We can not be sure exactly what form digital libraries will take, but it is clear that they are here to stay.

Chapter 2
The Internet and the World Wide Web

The Internet

The Internet and the World Wide Web are two of the principal building blocks that are used in the development of digital libraries. This chapter describes their main features. It also discusses the communities of engineers and computer scientists who created and continue to develop them. In contrast, the next chapter focuses on librarians, publishers, and other information professionals. Each chapter includes a discussion of how the community functions, its sense of purpose, the social conventions, priorities, and some of the major achievements.

As its name suggest, the Internet is not a single homogeneous network. It is an interconnected group of independently managed networks. Each network supports the technical standards needed for inter-connection - the TCP/IP family of protocols and a common method for identifying computers - but in many ways the separate networks are very different. The various sections of the Internet use almost every kind of communications channel that can transmit data. They range from fast and reliable to slow and erratic. They are privately owned or operated as public utilities. They are paid for in different ways. The Internet is sometimes called an information highway. A better comparison would be the international transportation system, with everything from airlines to dirt tracks.

Historically, these networks originated in two ways. One line of development was the local area networks that were created to link computers and terminals within a department or an organization. Many of the original concepts came from Xerox's Palo Alto Research Center. In the United States, universities were pioneers in expanding small local networks into campus-wide networks. The second source of network developments were the national networks, known as wide area networks. The best known of these was the ARPAnet, which, by the mid-1980s, linked about 150 computer science research organizations. In addition, there were other major networks, which have been almost forgotten, such as Bitnet in the United States. Their users are now served by the Internet.

Since the early networks used differing technical approaches, linking them together was hard. During the late 1980s the universities and the research communities converged on TCP/IP, the protocols of the ARPAnet, to create the Internet that we know today. A key event was the 1986 decision by the National Science Foundation to build a high-speed network backbone for the United States and to support the development of regional networks connected to it. For this backbone, the foundation decided to build upon the ARPAnet's technical achievements and thus set the standards for the Internet. Meanwhile, campus networks were using the same standards. We built one of the first high-speed networks at Carnegie Mellon, completing the installation in 1986. It accommodated several methods of data transmission, and supported competing network protocols from Digital, Apple, and IBM, but the unifying theme was the use of TCP/IP.

Future historians can argue over the combination of financial, organizational, and technical factors that led to the acceptance of the ARPAnet technical standards. Several companies made major contributions to the development and expansion of the

Internet, but the leadership came from two U.S. government organizations: DARPA and the National Science Foundation. Chapter 4 describes the contribution that these organizations continue to make to digital libraries research.

An introduction to Internet technology

The technology of the Internet is the subject of whole books. Most of the details are unimportant to users, but a basic understanding of the technology is useful when designing and using digital libraries. Panel 2.1 is a brief introduction to TCP/IP, the underlying protocols that define the Internet.

Panel 2.1
An introduction to TCP/IP

The two basic protocols that hold the Internet together are known as TCP/IP. The initials TCP and IP are joined together so often that it is easy to forget that they are two separate protocols.

IP

IP, the Internet Protocol, joins together the separate network segments that constitute the Internet. Every computer on the Internet has a unique address, known as an IP address. The address consists of four numbers, each in the range 0 to 255, such as 132.151.3.90. Within a computer these are stored as four bytes. When printed, the convention is to separate them with periods as in this example. IP, the Internet Protocol, enables any computer on the Internet to dispatch a message to any other, using the IP address. The various parts of the Internet are connected by specialized computers, known as "routers". As their name implies, routers use the IP address to route each message on the next stage of the journey to its destination.

Messages on the Internet are transmitted as short packets, typically a few hundred bytes in length. A router simply receives a packet from one segment of the network and dispatches it on its way. An IP router has no way of knowing whether the packet ever reaches its ultimate destination.

TCP

Users of the network are rarely interested in individual packets or network segments. They need reliable delivery of complete messages from one computer to another. This is the function of TCP, the Transport Control Protocol. On the sending computer, an application program passes a message to the local TCP software. TCP takes the message, divides it into packets, labels each with the destination IP address and a sequence number, and sends them out on the network. At the receiving computer, each packet is acknowledged when received. The packets are reassembled into a single message and handed over to an application program.

This protocol should be invisible to the user of a digital library, but the responsiveness of the network is greatly influenced by the protocol and this often affects the performance that users see. Not all packets arrive successfully. A router that is overloaded may simply ignore some packets. This is called, "dropping a packet." If this happens, the sending computer never receives an acknowledgment. Eventually it gets tired of waiting and sends the packet again. This is known as a "time-out" and may be perceived by the user as an annoying delay.

Other protocols

TCP guarantees error-free delivery of messages, but it does not guarantee that they will be delivered punctually. Sometimes, punctuality is more important than complete accuracy. Suppose one computer is transmitting a stream of audio that another is playing immediately on arrival. If an occasional packet fails to arrive on time, the human ear would much prefer to lose tiny sections of the sound track rather than wait for a missing packet to be retransmitted, which would be horribly jerky. Since TCP is unsuitable for such applications, they use an alternate protocol, named UDP, which also runs over IP. With UDP, the sending computer sends out a sequence of packets, hoping that they will arrive. The protocol does its best, but makes no guarantee that any packets ever arrive.

Panel 2.1 introduced the Internet addresses, known as IP addresses. Another way to identify a computer on the Internet is to give it a name such as tulip.mercury.cmu.edu. Names of this form are known as **domain names** and the system that relates domain names to IP addresses is known as the **domain name system** or DNS. Domain names and their role in digital libraries are studied in Chapter 12.

Computers that supports the TCP/IP protocols usually provide a standard set of basic applications. These applications are known as the TCP/IP suite. Some of the most commonly used are listed in Panel 2.2.

Panel 2.2
The TCP/IP suite

The TCP/IP suite is a group of computer programs, based on TCP/IP, that are provided by most modern computers. They include the following.

Terminal emulation. Telnet is a program that allows a personal computer to emulate an old-fashioned computer terminal that has no processing power of its own but relies on a remote computer for processing. Since it provides a lowest common denominator of user interface, telnet is frequently used for system administration.

File transfer. The basic protocol for moving files from one computer to another across the Internet is FTP (the file transfer protocol). Since FTP was designed to make use of TCP it is an effective way to move large files across the Internet.

Electronic mail. Internet mail uses a protocol known as Simple Mail Transport Protocol (SMTP). This is the protocol that turned electronic mail from a collection of local services to a single, world-wide service. It provides a basic mechanism for delivering mail. In recent years, a series of extensions have been made to allow messages to include wider character sets, permit multi-media mail, and support the attachment of files to mail messages.

The Internet community

The technology of the Internet is important, but the details of the technology may be less important than the community of people who developed it and continue to enhance it. The Internet pioneered the concept of open standards. In 1997, Vinton Cerf and Robert Kahn received the National Medal of Technology for their contributions to the Internet. The citation praised their work on the TCP/IP family of protocols, but it also noted that they, "pioneered not just a technology, but also an economical and efficient way to transfer that technology. They steadfastly maintained that their internetworking protocols would be freely available to anyone. TCP/IP was deliberately designed to be vendor-independent to support networking across all lines of computers and all forms of transmission."

The Internet tradition emphasizes collaboration and, even now, the continuing development of the Internet remains firmly in the hands of engineers. Some people seem unable to accept that the U.S. government is capable of anything worthwhile, but the creation of the Internet was led by government agencies, often against strong resistance by companies who now profit from its success. Recently, attempts have been made to rewrite the history of the Internet to advance vested interest, and for individuals to claim responsibility for achievements that many shared. The contrast is striking between the coherence of the Internet, led by far-sighted government officials, and the mess of incompatible standards in areas left to commercial competition, such as mobile telephones.

An important characteristic of the Internet is that the engineers and computer scientists who develop and operate it are heavy users of their own technology. They communicate by e-mail, dismissing conventional mail as "snail mail." When they write a paper they compose it at their own computer. If it is a web page, they insert the mark-up tags themselves, rather than use a formatting program. Senior computer scientists may spend more time preparing public presentations than writing computer programs, but programming is the basic skill that everybody is expected to have.

Panel 2.3
NetNews or Usenet

The NetNews bulletin boards, also known as Usenet, are an important and revealing examples of the Internet community's approach to the open distribution of information. Thousands of bulletin boards, called news groups, are organized in a series of hierarchies. The highest level groupings include "comp", for computer-related information, "rec" for recreational information, the notorious "alt" group, and many more. Thus "rec.arts.theatre.musicals" is a bulletin board for discussing musicals. (The British spelling of "theatre" suggests the origin of this news group.)

The NetNews system is so decentralized that nobody has a comprehensive list of all the news groups. An individual who wishes to post a message to a group sends it to the local news host. This passes it to its neighbors, who pass it to their neighbors and so on.

If we consider a digital library to contain managed information, NetNews is the exact opposite. It is totally unmanaged information. There are essentially no restrictions on who can post or what they can post. At its worst the system distributes libel, hate, pornography, and simply wrong information, but many news groups work remarkably well. For example, people around the world who use the Python programming language have a news group, "comp.lang.python", where they exchange technical information, pose queries, and communicate with the developer.

Scientific publishing on the Internet

The publishing of serious academic materials on the Internet goes back many years. Panels 2.4 and 2.5 describe two important examples, the Internet RFC series and the Physics E-Print Archives at the Los Alamos National Laboratory. Both are poorly named. The letters "RFC" once stood for "Request for Comment", but the RFC series is now the definitive technical series for the Internet. It includes a variety of technical information and the formal Internet standards. The Los Alamos service is not an archive in the usual sense. Its primary function is as a pre-print server, a site where researchers can publish research as soon as it is complete, without the delays of conventional journal publishing.

Whatever the merits of their names, these two services are of fundamental importance for publishing research in their respective fields. They are important also because they demonstrate that the best uses of digital libraries may be new ways of doing things. One of the articles of faith within scholarly publishing is that quality can be achieved only by peer review, the process by which every article is read by other specialists before publication. The process by which Internet drafts become RFCs is an intense form of peer review, but it takes place after a draft of the paper has been officially posted. The Los Alamos service has no review process. Yet both have proved to be highly effective methods of scientific communication.

Economically they are also interesting. Both services are completely open to the user. They are professionally run with substantial budgets, but no charges are made for authors who provide information to the service or to readers who access the information. Chapter 6 looks more deeply at the economic models behind the services and the implications for other scientific publishing.

The services were both well-established before the emergence of the web. The web has been so successful that many people forget that there are other effective ways to distribute information on the Internet. Both the Los Alamos archives and the RFC series now use web methods, but they were originally built on electronic mail and file transfer.

Panel 2.4
The Internet Engineering Task Force and the RFC series

The Internet Engineering Task Force

The Internet Engineering Task Force (IETF) is the body that coordinates technical aspects of the Internet. Its methods of working are unique, yet it has proved extraordinarily able to get large numbers of people, many from competing companies, to work together. The first unusual feature is that the IETF is open to everybody. Anybody can go to the regular meetings, join the working groups, and vote.

The basic principle of operation is "rough consensus and working code". Anybody who wishes to propose a new protocol or other technical advance is encouraged to provide a technical paper, called an Internet Draft, and a reference implementation of the concept. The reference implementation should be software that is openly available to all. At the working group meetings, the Internet Drafts are discussed. If there is a consensus to go ahead, then they can join the RFC standards track. No draft standard can become a formal standard until there are implementations of the specification, usually computer programs, openly available for everybody to use.

The IETF began in the United States, but now acts internationally. Every year, one meeting is outside the U.S.. The participants, including the working group leaders, come from around the world. The IETF was originally funded by U.S. government grants, but it is now self-sufficient. The costs are covered by meeting fees.

The IETF as a community

The processes of the IETF are open to everybody who wishes to contribute. Unlike some other standards bodies, whose working drafts are hard to obtain and whose final standards are complex and expensive, all Internet Drafts and RFCs are openly available. Because of the emphasis on working software, when there is rivalry between two technical approaches, the first to be demonstrated with software that actually works has a high chance of acceptance. As a result, the core Internet standards are remarkably simple.

In recent years, IETF meetings have grown to more than two thousand people, but, because they divide into working groups interested in specific topics, the basic feeling of intimacy remains. Almost everybody is a practicing engineer or computer scientists. The managers stay at home. The formal meetings are short and informal; the informal meetings are long and intense. Many an important specification came from a late evening session at the IETF, with people from competing organizations working together.

Because of its rapid growth, the Internet is in continually danger of breaking down technically. The IETF is the fundamental reason that it shows so much resilience. If a single company controlled the Internet, the technology would be as good as the company's senior engineering staff. Because the IETF looks after the Internet technology, whenever challenges are anticipated, the world's best engineers combine to solve them.

Internet Drafts

Internet Drafts are a remarkable series of technical publications. In science and engineering, most information goes out of date rapidly, but journals sit on library shelves for ever. Internet Drafts are the opposite. Each begins with a fixed statement that includes the words:

"Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as work in progress."

The IETF posts online every Internet Draft that is submitted, and notifies interested people through mailing lists. Then the review begins. Individuals post their comments on the relevant mailing list. Comments range from detailed suggestions to biting criticism. By the time that the working group comes to discuss a proposal, it has been subjected to public review by the experts in the field.

The RFC series

The RFC series are the official publications of the IETF. These few thousand publications form a series that goes back almost thirty years. They are the heart of the documentation of the Internet. The best known RFCs are the standards track. They include the formal specification of each version of the IP protocol, Internet mail, components of the World Wide Web, and many more. Other types of RFC include informational RFCs, which publish technical information relevant to the Internet.

Discussions of scientific publishing rarely mention the RFC series, yet it is hard to find another set of scientific or engineering publications that are so heavily reviewed before publication, or are so widely read by the experts in the field. RFCs have never been published on paper. Originally they were available over the Internet by FTP, more recently by the web, but they still have a basic text-only format, so that everybody can read them whatever computing system they use.

Panel 2.5
The Los Alamos E-Print Archives

The Physics E-print Archives are an intriguing example of practicing scientists taking advantage of the Internet technology to create a new form of scientific communication. They use the Internet to extend the custom of circulating pre-prints of research papers. The first archive was established in1991, by Paul Ginsparg of Los Alamos National Laboratory, to serve the needs of a group of high-energy physicists. Subsequent archives were created for other branches of physics, mathematics, and related disciplines. In 1996, Ginsparg reported, "These archives now serve over 35,000 users worldwide from over 70 countries, and process more than 70,000 electronic transactions per day. In some fields of physics, they have already supplanted traditional research journals as conveyers of both topical and archival research information."

The operation of the archives

The primary function of the archives is for scientists to present research results. Frequently, but not always, these are papers that will be published in traditional journals later. Papers for the archives are prepared in the usual manner. Many physicists use the TeX format, which has particularly good support for mathematics, but other formats, such as PostScript, or HMTL are also used. Graphs and data can be embedded in the text or provided as separate files.

The actual submission of a paper to an archive can be by electronic mail, file transfer using the FTP protocol, or via a web form. In each case, the author provides the paper, a short abstract, and a standard set of indexing metadata. Processing by the archive is entirely automatic. The archives provide a search service based on electronic mail, a web-based search system, and a notification service to subscribers, which also uses electronic mail. The search options include searching one or many of the archives. Readers can search by author and title, or search the full text of the abstracts.

As this brief description shows, the technology of the archives is straightforward. They use the standard formats, protocols, and networking tools that researchers know and understand. The user interfaces have been designed to minimize the effort required to maintain the archive. Authors and readers are expected to assist by installing appropriate software on their own computers, and by following the specified procedures.

The economic model

This is an open access system. The cost of maintaining the service is funded through annual grants from the National Science Foundation and the Department of Energy. Authors retain copyright in their papers. Indeed, about the only mention of copyright in the archive instructions is buried in a page of disclaimers and acknowledgments.

As Ginsparg writes, "Many of the lessons learned from these systems should carry over to other fields of scholarly publication, i.e. those wherein authors are writing not for direct financial remuneration in the form of royalties, but rather primarily to communicate information (for the advancement of knowledge, with attendant benefits to their careers and professional reputations)."

The World Wide Web

The **World Wide Web**, or "the web" as it is colloquially called, has been one of the great successes in the history of computing. It ranks with the development of word processors and spread sheets as a definitive application of computing. The web and its associated technology have been crucial to the rapid growth of digital libraries. This section gives a basic overview of the web and its underlying technology. More details of specific aspects are spread throughout the book.

The web is a linked collection of information on many computers on the Internet around the world. These computers are called web servers. Some of web servers and the information on them are maintained by individuals, some by small groups, perhaps at universities and research centers, others are large corporate information services. Some sites are consciously organized as digital libraries. Other excellent sites are managed by people who would not consider themselves librarians or publishers. Some web servers have substantial collections of high-quality information. Others are used for short term or private purposes, are informally managed, or are used for purposes such as marketing that are outside the scope of libraries.

The web technology was developed about 1990 by Tim Berners-Lee and colleagues at CERN, the European research center for high-energy physics in Switzerland. It was made popular by the creation of a user interface, known as Mosaic, which was developed by Marc Andreessen and others at the University of Illinois, Urbana-Champaign. Mosaic was released in 1993. Within a few years, numerous commercial versions of Mosaic followed. The most widely used are the Netscape Navigator and Microsoft's Internet Explorer. These user interfaces are called **web browsers**, or simply **browsers**.

The basic reason for the success of the web can be summarized succinctly. It provides a convenient way to distribute information over the Internet. Individuals can publish information and users can access that information by themselves, with no training and no help from outsiders. A small amount of computer knowledge is needed to establish a web site. Next to none is needed to use a browser to access the information.

The introduction of the web was a grass roots phenomenon. Not only is the technology simple, the manner in which it was released to the public avoided almost all barriers to its use. Carnegie Mellon's experience was typical. Individuals copied the web software over the Internet onto private computers. They loaded information that interested them and made it accessible to others. Their computers were already connected to the campus network and hence to the Internet. Within six months of the first release of Mosaic, three individuals had established major collections of academic information, covering statistics, English, and environmental studies. Since the Internet covers the world, huge numbers of people had immediate access to this information. Subsequently, the university adopted the web officially, with a carefully designed home page and information about the university, but only after individuals had shown the way.

One reason that individuals have been able to experiment with the web is that web software has always been available at no charge over the Internet. CERN and the University of Illinois set the tradition with open distribution of their software for web servers and user interfaces. The most widely used web servers today is a no-cost version of the Illinois web server, known as Apache. The availability of software that is distributed openly over the Internet provides a great stimulus in gaining acceptance for new technology. Most technical people enjoy experimentation but dislike bureaucracy. The joy goes out of an experiment if they need to issue a purchase order or ask a manager to sign a software license.

Another reason for the instant success of the web was that the technology provides gateways to information that was not created specifically for the web. The browsers are designed around the web protocol called HTTP, but the browsers also support other Internet protocols, such as file transfer (FTP), Net News, and electronic mail. Support for Gopher and WAIS, two other protocols which are now almost obsolete, allowed earlier collections of information to coexist with the first web sites. Another mechanism, the Common Gateway Interface (CGI), allows browsers to bridge the gap between the web and any other system for storing online information. In this way, large amounts of information were available as soon as Mosaic became available.

From the first release of Mosaic, the leading browsers have been available for the most common types of computer - the various versions of Windows, Macintosh and Unix - and browsers are now provided for all standard computers. The administrator of a web site can be confident that users around the world will see the information provided on the site in roughly the same format, whatever computers they have.

The technology of the web

Technically, the web is based on four simple techniques. They are: the Hyper-Text Mark-up Language (HTML), the Hyper-Text Transport Protocol (HTTP), MIME data types, and Uniform Resource Locators (URLs). Each of these concepts is introduced below and discussed further in later chapters. Each has importance that goes beyond the web into the general field of interoperability of digital libraries.

HTML

HTML is a language for describing the structure and appearance of text documents. Panel 2.6 shows a simple HTML file and how a typical browser might display, or render, it.

Panel 2.6
An HTML example

An HTML file

Here is a simple text file in HTML format as it would be stored in a computer. It shows the use of tags to define the structure and format of a document.

```
<html>
<head>
<title>D-Lib</title>
</head>

<body>
<h1>D-Lib Magazine</h1>
<img src = "logo.gif">

<p>Since the first issue appeared in July 1995,
<a href = "http://www.dlib.org/dlib.html">D-Lib Magazine</a> has
appeared monthly as a compendium of research, news, and progress in
```

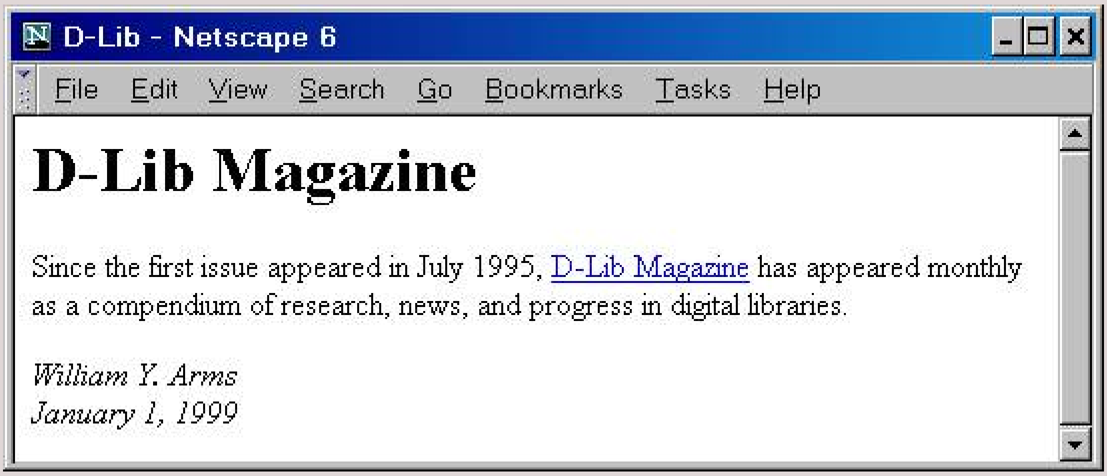
```
digital libraries.</p>

<p><i>William Y. Arms
<br>January 1, 1999</i></p>

</body>
</html>
```

The document displayed by a web browser

When displayed by a browser, this document might be rendered as follows. The exact format depends upon the specific browser, computer, and choice of options.



D-Lib Magazine

Since the first issue appeared in July 1995, D-Lib Magazine has appeared monthly as a compendium of research, news, and progress in digital libraries.

William Y. Arms
January 1, 1999

As this example shows, the HTML file contains both the text to be rendered and codes, known as tags, that describe the format or structure. The HMTL tags can always be recognized by the angle brackets (< and >). Most HTML tags are in pairs with a "/" indicating the end of a pair. Thus <title> and </title> enclose some text that is interpreted as a title. Some of the HTML tags show format; thus <i> and </i> enclose text to be rendered in italic, and <br> shows a line break. Other tags show structure: <p> and </p> delimit a paragraph, and <h1> and </h1> bracket a level one heading. Structural tags do not specify the format, which is left to the browser.

For example, many browsers show the beginning of a paragraph by inserting a blank line, but this is a stylistic convention determined by the browser. This example also shows two features that are special to HMTL and have been vital to the success of the web. The first special feature is the ease of including color image in web pages. The tag:

```
<img src = "logo.gif">
```

is an instruction to insert an image that is stored in a separate file. The abbreviation "img" stands for "image" and "src" for "source". The string that follows is the name of the file in which the image is stored. The introduction of this simple command by Mosaic brought color images to the Internet. Before the web, Internet applications were drab. Common applications used unformatted text with no images. The web was the first, widely used system to combine formatted text and color images. Suddenly the Internet came alive.

The second and even more important feature is the use of hyperlinks. Web pages do not stand alone. They can link to other pages anywhere on the Internet. In this example, there is one hyperlink, the tag:

<a href = "http://www.dlib.org/dlib.html">

This tag is followed by a string of text terminated by </a>. When displayed by as browser, as in the panel, the text string is highlighted; usually it is printed in blue and underlined. The convention is simple. If something is underlined in blue, the user can click on it and the hyperlink will be executed. This convention is easy for the both the user and the creator of the web page. In this example, the link is to an HTML page on another computer, the home page of *D-Lib Magazine*.

The basic concepts of HTML can be described in a short tutorial and absorbed quickly, but even simple tagging can create an attractive document. In the early days of the web, everybody was self-taught. A useful characteristic of HTML is that its syntax is forgiving of small mistakes. Other computing languages have strict syntax. Omit a semi-colon in a computer program and the program fails or gives the wrong result. With HTML, if the mark-up is more or less right, most browsers will usually accept it. The simplicity has the added benefit that computers programs to interpret HTML and display web pages are straightforward to write.

Uniform Resource Locator (URL)

The second key component of the web is the Uniform Resource Locator, known as a URL. URLs are ugly to look at but are flexible. They provide a simple addressing mechanism that allows the web to link information on computers all over the world. A simple URL is contained in the HTML example in Panel 2.6:

http://www.dlib.org/dlib.html

This URL has three parts:

http is the name of a protocol
www.dlib.org is the domain name of a computer
dlib.html is a file on that computer

Thus an informal interpretation of this URL is, "Using the HTTP protocol, connect to the computer with the Internet address www.dlib.org, and access the file dlib.html."

HTTP

In computing, a protocol is a set of rules that are used to send messages between computer systems. A typical protocol includes description of the formats to be used, the various messages, the sequences in which they should be sent, appropriate responses, error conditions, and so on. HTTP is the protocol that is used to send messages between web browsers and web servers.

The basic message type in HTTP is get. For example, clicking on the hyperlink with the URL:

http://www.dlib.org/dlib.html

specifies an HTTP get command. An informal description of this command is:

- Open a connection between the browser and the web server that has the domain name "www.dlib.org".
- Copy the file "dlib.html" from the web server to the browser.
- Close the connection.

MIME types

A file of data in a computer is simply a set of bits, but, to be useful the bits need to be interpreted. Thus, in the previous example, in order to display the file "dlib.html" correctly, the browser must know that it is in the HTML format. The interpretation depends upon the data type of the file. Common data types are "html" for a file of text that is marked-up in HMTL format, and "jpeg" for a file that represents an image encoded in the jpeg format.

In the web, and in a wide variety of Internet applications, the data type is specified by a scheme called **MIME**. (The official name is Internet Media Types.) MIME was originally developed to describe information sent by electronic mail. It uses a two part encoding, a generic part and a specific part. Thus text/ascii is the MIME type for text encoded in ASCII, image/jpeg is the type for an image in the jpeg format, and text/html is text marked-up with HMTL tags. As described in Chapter 12, there is a standard set of MIME types that are used by numerous computer programs, and additional data types can be described using experimental tags.

The importance of MIME types in the web is that the data transmitted by an HTTP get command has a MIME type associated with it. Thus the file "dlib.html" has the MIME type text/html. When the browser receives a file of this type, it knows that the appropriate way to handle this file is to render it as HTML text and display it in the screen.

Many computer systems use file names as a crude method of recording data types. Thus some Windows programs use file names that end in ".htm" for file of HMTL data and Unix computers use ".html" for the same purpose. MIME types are a more flexible and systematic method to record and transmit typed data.

Information on the web

The description of the web so far has been technical. These simple components can be used to create many applications, only some of which can be considered digital libraries. For example, many companies have a web site to describe the organization, with information about products and services. Purchases, such as airline tickets, can be made from web servers. Associations provide information to their members. Private individuals have their own web sites, or personal home pages. Research results may be reported first on the web. Some web sites clearly meet the definition of a digital library, as managed collections of diverse information. Others do not. Many web sites meet the definition of junk.

In aggregate, however, these sites are all important for the development of digital libraries. They are responsible for technical developments beyond the simple building blocks described above, for experimentation, for the establishments of conventions for organizing materials, for the increasingly high quality of graphical design, and for the large number of skilled creators, users, and webmasters. The size of the web has stimulated numerous companies to develop products, many of which are now being used to build digital libraries. Less fortunately, the success of all these web sites frequently overloads sections of the Internet, and has generated social and legal concerns about abusive behavior on the Internet.

Panel 2.7
The World Wide Web Consortium

No central organization controls the web, but there is a need for agreement on the basic protocols, formats, and practices, so that the independent computer systems can interoperate. In 1994, recognizing this need, the Massachusetts Institute of Technology created the World Wide Web Consortium (W3C) and hired Tim Berners-Lee, the creator of the web, as its director. Subsequently, MIT added international partners at the Institut National de Recherche en Informatique et en Automatique in France and the Keio University Shonan Fujisawa Campus in Japan. W3C is funded by member organizations who include most of the larger companies who develop web browsers, servers, and related products.

W3C is a neutral forum in which organizations work together on common specifications for the web. It works through a series of conferences, workshops, and design processes. It provides a collection of information about the web for developers and users, especially specifications about the web with sample code that helps to promote standards. In some areas, W3C works closely with the Internet Engineering Task Force to promulgate standards for basic web technology, such as HTTP, HTML, and URLs.

By acting as a neutral body in a field that is now dominated by fiercely competing companies, W3C's power depends upon its ability to influence. One of its greatest successes was the rapid development of an industry standard for rating content, known as PICS. This was a response to political worries in the United States about pornography and other undesirable content being accessible by minors. More recently it has been active in the development of the XML mark-up language.

Companies such as Microsoft and Netscape sometimes believe that they gain by supplying products that have non-standard features, but these features are a barrier to the homogeneity of the web. W3C deserves much of the credit for the reasonably coherent way that the web technology continues to be developed.

Conventions

The first web sites were created by individuals, using whatever arrangement of the information they considered most appropriate. Soon, conventions began to emerge about how to organize materials. Hyperlinks permit an indefinite variety of arrangements of information on web sites. Users, however, navigate most effectively through familiar structures. Therefore these conventions are of great importance in the design of digital libraries that are build upon the web. The conventions never went through any standardization body, but their widespread adoption adds a coherence to the web that is not inherent in the basic technology.

- **Web sites.** The term "web site" has already been used several times. A **web site** is a collection of information that the user perceives to be a single unit. Often, a web site corresponds to a single web server, but a large site may be physically held on several servers, and one server may be host to many web sites.

  The convention rapidly emerged for organizations to give their web site a domain name that begins "www". Thus "www.ibm.com" is the IBM web site"; "www.cornell.edu" is Cornell University; "www.elsevier.nl" is the Dutch publisher, Elsevier.

- **Home page.** A **home page** is the introductory page to a collection of web information. Almost every web site has a home page. If the address in a URL does not specify a file name, the server conventionally supplies a page called "index.html". Thus the URL:

  http://www.loc.gov/

  is interpreted as http://www.loc.gov/index.html. It is the home page of the Library of Congress. Every designer has slightly different ideas about how to arrange a home page but, just as the title page of a book follows standard conventions, home pages usually provide an overview of the web site. Typically, this combines an introduction to the site, a list of contents, and some help in finding information.

  The term "home page" is also applied to small sets of information within a web site. Thus it is common for the information relevant to a specific department, project, or service to have its own home page. Some individuals have their own home pages.

- **Buttons.** Most web pages have buttons to help in navigation. The buttons provide hyperlinks to other parts of the web site. These have standard names, such as "home", "next", and "previous". Thus, users are able to navigate confidently through unfamiliar sites.

- **Hierarchical organization.** As seen by the user, many web sites are organized as hierarchies. From the home page, links lead to a few major sections. These lead to more specific information and so on. A common design is to provide buttons on each page that allow the user to go back to the next higher level of the hierarchy or to move horizontal to the next page at the same level. Users find a simple hierarchy an easy structure to navigate through without losing a sense of direction.

The web as a digital library

Some people talk about the web technology as though it were an inferior stop-gap until proper digital libraries are created. One reason for this attitude is that members of other professions having difficulty in accepting that definitive work in digital libraries was carried out by physicists at a laboratory in Switzerland, rather than by well-known librarians or computer scientists. But the web is not a detour to follow until the real digital libraries come along. It is a giant step to build on.

People who are unfamiliar with the online collections also make derogatory statements about the information on the web. The two most common complaints are that the information is of poor quality and that it is impossible to find it. Both complaints have some validity, but are far from the full truth. There is an enormous amount of material on the web; much of the content is indeed of little value, but many of the web servers are maintained conscientiously, with information of the highest quality. Finding information on the web can be difficult, but tools and services exist that enable a user, with a little ingenuity, to discover most of the information that is out there.

Today's web, however, is a beginning, not the end. The simplifying assumptions behind the technology are brilliant, but these same simplifications are also limitations. The web of today provides a base to build the digital libraries of tomorrow. This requires better collections, better services, and better underlying technology. Much of the current research in digital libraries can be seen as extending the basic building blocks of the web. We can expect that, twenty five years from now, digital libraries will be very different; it will be hard to recall the early days of the web. The names "Internet" and "web" may be history or may be applied to systems that are unrecognizable as descendants of the originals. Digital libraries will absorb materials and technology from many places. For the next few years, however, we can expect to see the Internet and the web as the basis on which the libraries of the future are being built. Just as the crude software on early personal computers has developed into modern operating systems, the web can become the foundation for many generations of digital libraries.

Chapter 3
Libraries and publishers

Computing in libraries

Faced with the management of huge collections over long periods of time, libraries have a proud tradition as early adopters of new technology, from microfilm, through online information services and CD-ROMs, to digital libraries. (The second typewriter imported into Brazil was for the National Library of Brazil to type library catalog cards.) The Internet and the web are the latest examples of technology that has been embraced by libraries and is being adapted to their needs, but far from the first.

Mainstream commercial products do not always fit the needs of libraries. The computing industry aims its product developments at the large markets provided by commercial businesses, science, medicine, and the military. In the United States, much of the basic research behind these products has been funded by the Department of Defense or companies such as IBM. Until recently, these large markets had little interest in managing the kinds of information that is found in digital libraries. Consequently, the products sold by the computing industry did not address important needs in information management. Therefore, libraries are accustomed to taking core technology from other fields and tailoring it to their needs. This chapter introduces several such areas, most of which are described more fully in later chapters.

Resource sharing and online catalogs

In libraries, the word "networking" has two meanings. Before modern computer networks, libraries had a wide variety of arrangements to share resources. These include inter-library lending, reciprocal reading privileges between institutions, and the exchange of photocopies. The term given to these arrangements was "networking" or "resource sharing". Resource sharing services, such as the photocopying and document delivery provided by the British Library, enhance the availability of information around the world. However, nothing in resource sharing has had an impact to rival computer networking.

Shared cataloguing was mentioned in Chapter 1 as an early use of computers to support libraries. Almost every library has a catalog with records of the materials in the collections. The catalog has several uses. It helps users find material in the library, it provides bibliographic information about the items, and it is an important tool in managing the collections.

Cataloguing is a area in which librarians use precise terminology, some of which may be unfamiliar to outsiders. Whereas a non-specialist may use the term **catalog** as a generic term, in a library it has a specific meaning as a collection of bibliographic records created according to strict rules. Rather than the everyday word "book", librarians use the more exact term **monograph**. Over the years, the information that is included in a monograph catalog record has been codified into cataloguing rules. English speaking countries use the Anglo-American Cataloguing Rules (AACR). The current version is known as AACR2.

The task of cataloguing each monograph is time consuming and requires considerable expertise. To save costs, libraries share catalog records. Major libraries that catalog large numbers of monographs, such as the Library of Congress and the major university libraries, make their catalog records available to others, free of charge. Rather than create duplicate records, most libraries look for an existing catalog record and create their own records only when they can not find another to copy.

Since the late 1960s, this sharing of catalog records has been computer based. Technically, the fundamental tool for distributing catalog records is the MARC format (an abbreviation for Machine-Readable Cataloging). MARC was developed by Henriette Avram and colleagues at the Library of Congress, initially as a format to distribute monograph catalog records on magnetic tape. To be precise, it is necessary to distinguish between AACR2, which provides the cataloguing rules, and MARC, which is a format for representing the resulting catalog records. In practice, the term "MARC cataloguing" is often used in a general sense to cover both the catalog records and the computer format in which they are stored. MARC cataloguing has been expanded far beyond monographs and is now used for most categories of library materials, including serials, archives, manuscripts, and many more. Panel 3.1 gives an example of a MARC record.

Panel 3.1. A MARC record
A MARC record

Consider a monograph, for which the conventional bibliographic citation is:

> Caroline R. Arms, editor, Campus strategies for libraries and electronic information. Bedford, MA: Digital Press, 1990.

A search of the catalog at the Library of Congress, using one of the standard terminal-based interfaces, displays the catalog in a form that shows the information in the underlying MARC record format.

```
&00189-16879 r93
&050 Z675.U5C16 1990
&082 027.7/0973 20
&245 Campus strategies for libraries and electronic information/Caroline
Arms, editor.
&260 {Bedford, Mass.} : Digital Press, c1990.
&300 xi, 404 p. : ill. ; 24 cm.
&440 EDUCOM strategies series on information technology
&504 Includes bibliographical references (p. {373}-381).
&020 ISBN 1-55558-036-X : $34.95
&650 Academic libraries--United States--Automation.
&650 Libraries and electronic publishing--United States.
&650 Library information networks--United States.
&650 Information technology--United States.
&700 Arms, Caroline R. (Caroline Ruth)
&040 DLC DLC DLC &043 n-us---
&955 CIP ver. br02 to SL 02-26-90
&985 APIF/MIG
```

The information is divided into fields, each with a three-digit code. For example, the 440 field is the title of a monograph series, and the 650 fields are Library of Congress subject headings. Complex rules specify to the cataloguer which fields should be used and how relationships between elements should be interpreted.

The actual coding is more complex than shown. The full MARC format consists of a pre-defined set of fields each identified by a tag. Within each field, subfields are permitted. Fields are identified by three-digit numeric tags and subfields by single letters. To get a glimpse of how information is encoded in this format, consider the 260 field, which begins "&260". In an actual MARC record, this is encoded as:

```
&2600#abc#{Bedford, Mass.} :#Digital Press,#c1990.%
```

This has information about publication, divided into three subfields. The string "abc" indicates that there are three subfields. The first with tag "a" is the place of publication, the next with tag "b" is the publisher, and the third with tag "c" is the date.

The development of MARC led to two important types of computer-based system. The first was **shared cataloguing**; the pioneer was OCLC, created by Fred Kilgour in 1967. OCLC has a large computer system which has grown to more 35 million catalog records in MARC format, including records received from the Library of Congress. When an OCLC member library acquires a book that it wishes to catalog, it begins by searching the OCLC database. If it finds a MARC record, it downloads the record to its own computer system and records the holding in the OCLC database. In the past it could also have ordered a printed catalog card. This is called "copy cataloguing". If the OCLC database does not contain the record, the library is encouraged to create a record and contribute it to OCLC. With copy cataloguing, each item is catalogued once and the intellectual effort shared among all libraries. MARC cataloguing and OCLC's success in sharing catalog records have been emulated by similar services around the world.

The availability of MARC records stimulated a second development. Individual libraries were are to create **online catalogs** of their holdings. In most cases, the bulk of the records were obtained from copy cataloguing. Today, almost every substantial library in the United States has its online catalog. Library jargon calls such a catalog an "OPAC" for "online public access catalog". Many libraries have gone to great efforts to convert their old card catalogs to MARC format, so that the online catalog is the record of their entire holdings, rather than having an online catalog for recent acquisitions, but traditional card catalogs for older materials, The retrospective conversion of Harvard University's card catalog to MARC format has recently been completed. Five million cards were converted at a cost approaching $15 million.

A full discussion of MARC cataloguing and online public access catalogs is outside the scope of this book. MARC was an innovative format at a time when most computer systems represented text as fixed length fields with capital letters only. It remains a vital format for libraries, but it is showing its age. Speculation on the future of MARC is complicated by the enormous investment that libraries have made in it. Whatever its future, MARC was a pioneering achievement in the history of both computing and libraries. It is a key format that must be accommodated by digital libraries.

Linking online catalogs and Z39.50

During the 1980s, universities libraries began to connect their online catalogs to networks. As an early example, by 1984 there was a comprehensive campus network at Dartmouth College. Since the computer that held the library catalog was connected to the network, anybody with a terminal or personal computer on-campus could search the catalog. Subsequently, when the campus network was connected to the Internet, the catalog became available to the whole world. People outside the university could search the catalog and discover what items were held in the libraries at Dartmouth. Members of the university could use their own computers to search the catalogs of other universities. This sharing of library catalogs was one of the first, large-scale examples of cooperative sharing of information over the Internet.

In the late 1970s, several bibliographic utilities, including the Library of Congress, the Research Libraries Group, and the Washington Libraries Information Network, began a project known as the Linked Systems Project, which developed the protocol now known by the name of Z39.50. This protocol allows one computer to search for information on another. It is primarily used for searching records in MARC format, but the protocol is flexible and is not restricted to MARC. Technically, Z39.50 specifies rules that allow one computer to search a database on another and retrieve the records that are found by the search. Z39.50 and its role in fostering interoperability among digital libraries are discussed in Chapter 11. It is one of the few protocols to be widely used for interoperation among diverse computer systems.

Abstracts and indexes

Library catalogs are the primary source of information about monographs, but they are less useful for journals. Catalogs provide a single, brief record for an entire run of a journal. This is of little value to somebody who wants to discover individual articles in academic journals. Abstracting and indexing services developed to help researchers to find such information. Typical services are Medline for the biomedical literature, Chemical Abstracts for chemistry, and Inspec for the physical sciences including computing. The services differ in many details, but the basic structures are similar. Professionals, who are knowledgeable about a subject area, read each article from a large number of journals and assign index terms or write abstracts. Sometimes services use index terms that are drawn from a carefully controlled vocabulary, such as the MeSH headings that the National Library of Medicine uses for its Medline service. Others services are less strict. Some generate all their own abstracts. Others, such as Inspec, will use an abstract supplied by the publisher.

Most of these services began as printed volumes that were sold to libraries, but computer searching of these indexes goes back to the days of batch processing and magnetic tape. Today, almost all searching is by computer. Some indexing services run computer systems on which users can search for a fee; others license their data to third parties who provide online services. Many large libraries license the data and mount it on their own computers. In addition, much of the data is available on CD-ROM.

Once the catalog was online, libraries began to mount other data, such as abstracts of articles, indexes, and reference works. These sources of information can be stored in a central computer and the retrieved records displayed on terminals or personal computers. Reference works consisting of short entries are particularly suited for this form of distribution, since users move rapidly from one entry to another and will accept a display that has text characters with simple formatting. Quick retrieval and flexible searching are more important than the aesthetics of the output on the computer screen.

As a typical example, here are some of the many information sources that the library at Carnegie Mellon University provided online during 1997/8.

Carnegie Mellon library catalog
Carnegie Mellon journal list
Bibliographic records of architectural pictures and drawings
Who's who at CMU
American Heritage Dictionary
Periodical Abstracts
ABI/Inform (business periodicals)
Inspec (physics, electronics, and computer science)
Research directory (Carnegie Mellon University)

Several of these online collections provide local information, such as Who's who at CMU, which is the university directory of faculty, students, and staff. Libraries do not provide their patrons only with formally published or academic materials. Public libraries, in particular, are a broad source of information, from tax forms to bus timetables. Full-text indexes and web browsers allow traditional and non-traditional library materials to be combined in a single system with a single user interface. This approach has become so standard that it is hard to realize that only a few years ago merging information from such diverse sources was rare.

Mounting large amounts of information online and keeping it current is expensive. Although hardware costs fall continually, they are still noticeable, but the big costs are in licensing the data and the people who handle both the business aspects and the large data files. To reduce these costs, libraries have formed consortia where one set of online data serves many libraries. The MELVYL system, which serves the campuses of the University of California, was one of the first. It is described in Chapter 5.

Information retrieval

Information retrieval is a central topic for libraries. A user, perhaps a scientist, doctor, or lawyer, is interested in information on some topic and wishes to find the objects in a collection that cover the topic. This requires specialized software. During the mid-1980s, libraries began to install computers with software that allowed full-text searching of large collections. Usually, the MARC records of a library's holdings were the first data to be loaded onto this computer, followed by standard references works. Full-text searching meant that a user could search using any words that appeared in the record and did not need to be knowledgeable about the structures of the records or the rules used to create them.

Research in this field is at least thirty years old, but the basic approach has changed little. A user expresses a request as a **query**. This may be a single word, such as "cauliflower", a phrase, such as "digital libraries", or a longer query, such as, "In what year did Darwin travel on the Beagle?" The task of information retrieval is to find objects in the collection that match the query. Since a computer does not have the time to go through the entire collection for each search, looking at every object separately, the computer must have an index of some sort that allows information retrieval by looking up entries in indexes.

As computers have grown more powerful and the price of storage declined, methods of information retrieval have moved from carefully controlled searching of short records, such as catalog records or those used by abstracting and indexing services, to searching the full text of every word in large collections. In the early days, expensive computer storage and processing power, stimulated the development of compact methods of storage and efficient algorithms. More recently, web search programs have intensified research into methods for searching large amounts of information which are distributed across many computers. Information retrieval is a topic of Chapters 10 and 11.

Representations of text and SGML

Libraries and publishers share an interest in using computers to represent the full richness of textual materials. Textual documents are more than simple sequences of letters. They can contain special symbols, such as mathematics or musical notation, characters from any language in the world, embedded figures and table, various fonts, and structural elements such as headings, footnotes, and indexes. A desirable way to store a document in a computer is to encode these features and store them with the text, figures, tables, and other content. Such an encoding is called a **mark-up language**. For several years, organizations with a serious interest in text have been developing a mark-up scheme known as **SGML**. (The name is an abbreviation for Standard Generalized Markup Language.) HTML, the format for text that is used by the web, is a simple derivative of SGML.

Since the representation of a document in SGML is independent of how it will be used, the same text, defined by its SGML mark-up, can be displayed in many forms and formats: paper, CD-ROM, online text, hypertext, and so on. This makes SGML attractive for publishers who may wish to produce several versions of the same underlying work. A pioneer application in using SGML in this way was the new Oxford English Dictionary. SGML has also been heavily used by scholars in the humanities who find in SGML a method to encode the structure of text that is independent of any specific computer system or method of display. SGML is one of the topics in Chapter 9.

Digital libraries of scientific journals

The early experiments

During the late 1980s several publishers and libraries became interested in building online collections of scientific journals. The technical barriers that had made such projects impossible earlier were disappearing, though still present to some extent. The cost of online storage was coming down, personal computers and networks were being deployed, and good database software was available. The major obstacles to building digital libraries were that academic literature was on paper, not in electronic formats, and that institutions were organized around physical media, not computer networks.

One of the first attempts to create a campus digital library was the Mercury Electronic Library, a project that we undertook at Carnegie Mellon University between 1987 and 1993. Mercury was able to build upon the advanced computing infrastructure at Carnegie Mellon, which included a high-performance network, a fine computer science department, and the tradition of innovation by the university libraries. A slightly later effort was the CORE project at Cornell University to mount images of chemistry journals. Both projects worked with scientific publishers to scan journals and establish collections of online page images. Whereas Mercury set out to build a production system, CORE also emphasized research into user interfaces and other aspects of the system by chemists. The two projects are described in Panel 3.2.

Panel 3.2. Mercury and CORE

Mercury

Mercury was a five year to build a prototype digital library at Carnegie Mellon University. It began in 1988 and went live in 1991 with a dozen textual databases and a small number of page images of journal articles in computer science. It provides a good example of the state of the art before the arrival of the web.

One of the objectives was to mount page images of journal articles, using materials licensed from publishers. Four publishers were identified as publishing sixteen of the twenty computer science journals most heavily used on campus. They were ACM, IEEE, Elsevier, and Pergammon. During the project, Pergammon was taken over by Elsevier. None of the publishers had machine-readable versions of their journals, but they gave permission to convert printed materials for use in the library. Thus, an important part of the work was the conversion, storage, and delivery of page images over the campus network.

The fundamental paradigm of Mercury was searching a text database, to identify the information to be displayed. An early version of Z39.50 was chosen as the protocol to send queries between the clients and the server computers on which the indexes were stored. Mercury introduced the concept of a reference server, which keeps information about the information stored on the servers, the fields that could be searched, the indexes, and access restrictions. To display bit-mapped images, the project developed a new algorithm to take page images stored in compressed format, transmit them across the network, decompress them and display them, with an overall response time of one to two seconds per page.

Since Mercury was attached to the Internet and most materials were licensed from publishers, security was important. Carnegie Mellon already had a mature set of network services, known as Andrew. Mercury was able to use standard Andrew services for authentication and printing. Electronic mail was used to dispatch information to other computers.

CORE

CORE was a joint project by Bellcore, Cornell University, OCLC, and the American Chemical Society that ran from 1991 to 1995. The project converted about 400,000 pages, representing four years of articles from twenty journals published by the American Chemical Society.

The project used a number of ideas that have since become popular in conversion projects. CORE included two versions of every article, a scanned image and a text version marked up in SGML. The scanned images ensured that when a page was displayed or printed it had the same design and layout as the original paper version. The SGML text was used to build a full-text index for information retrieval and for rapid display on computer screens. Two scanned images were stored for each page, one for printing and the other for screen display. The printing version was black and white, 300 dots per inch; the display version was 100 dots per inch, grayscale.

CORE was one of the first projects to articulate several issues that have since been confirmed in numerous studies. The technical problems of representing, storing, and storing complex scientific materials are substantial, particularly if they were originally designed to be printed and digitization was an afterthought. CORE also highlighted the problems of scale. The CORE collections, which were only twenty journals, occupied some 80 gigabytes of storage. All the early digitization projects found that the operational problems of managing large collections were greater than predicted.

Despite such annoyances, the user interface studies showed considerable promise. Chemists liked the collections. Although they may prefer paper for reading an article in detail, they found the online versions easier to search and the screen displays of articles more than adequate. CORE was one of the first studies to emphasize the importance of browsing in addition to searching, an experience which the web has amply demonstrated.

Although both the Mercury and CORE projects converted existing journal articles from print to bit-mapped images, conversion was not seen as the long-term future of scientific libraries. It simply reflected the fact that none of the journal publishers were in a position to provide other formats. Printers had used computer typesetting for many years, but their systems were organized entirely to produce printed materials. The printers' files were in a wide variety of formats. Frequently, proof corrections were held separately and not merged with the master files, so that they were not usable in a digital library without enormous effort.

Mercury and CORE were followed by a number of other projects that explored the use of scanned images of journal articles. One of the best known was Elsevier Science Publishing's Tulip project. For three years, Elsevier provided a group of universities, which included Carnegie Mellon and Cornell, with images from forty three journals in material sciences. Each university, individually mounted these images on their own computers and made them available locally.

Projects such as Mercury, CORE, and Tulip were not long-term production systems. Each had rough edges technically and suffered from the small size of the collection provided to researchers, but they were followed by systems that have become central parts of scientific journal publishing. They demonstrated that the potential benefits of a digital library could indeed be achieved in practice. The next generation of developments in electronic publishing were able to take advantage of much cheaper computer storage allowing large collections of images to be held online. The emergence of the web and the widespread availability of web browsers went a long way towards solving the problem of user interface development. Web browsers are not ideal for a digital library, though they are a good start, but they have the great advantage that they exist for all standard computers and operating systems. No longer does every project have to develop its own user interface programs for every type of computer. Scientific journal publishers

Until the mid-1990s, established publishers of scientific journals hesitated about online publishing. While commercial publishing on CD-ROM had developed into an important part of the industry, few journals were available online. Publishers could not make a business case for online publishing and feared that, if materials were available online, sales of the same materials in print would suffer. By about 1995, it became clear that broad-reaching changes were occurring in how people were using information. Libraries and individuals were expanding their uses of online information services and other forms of electronic information much more rapidly than their use of printed materials. Print publications were competing for a declining portion of essentially static library budgets.

Panel 3.3. HighWire Press

HighWire Press is a venture of the Stanford University Libraries. It has brought some of the best scientific and medical journals online by building partnerships with the scientific and professional societies that publish the journals. Its success can be credited to attention to the interests of the senior researchers, a focus on the most important journals, and technical excellence.

HighWire Press began in 1995 with an online version of the Journal of Biological Chemistry, taking advantage of several members of the editorial board being Stanford faculty. This is a huge journal, the second largest in the world. Since each weekly edition is about 800 pages, nobody reads it from cover to cover. Therefore the HighWire Press interface treats articles as separate documents. It provides two ways to find them, by browsing the contents, issue by issues, or by searching. The search options include authors, words in title, or full-text searching of abstract or the entire article. The search screen and the display of the articles are designed to emphasize that this is indeed the Journal of Biological Chemistry not a publication of HighWire Press. Great efforts have been made to display Greek letters, mathematical symbols, and other special characters.

In three years, HighWire Press went from an experiment to a major operation with almost one hundred journals online, including the most prominent of all, Science, published by the American Society for the Advancement of Science (AAAS). Despite its resources, the AAAS had a problem that is faced by every society that publishes only a few journals. The society feared that an in-house effort to bring Science online might over-extend their staff and lower the overall quality of their work. A partnership with HighWire Press enabled them to share the development costs with other society journals and to collaborate with specialists at Stanford University. The AAAS has been delighted by the number of people who visit their site every week, most of whom were not regular readers of Science.

Chapter 6 explains that journal publishing is in the grip of a price spiral with flat or declining circulation. In the past, publishers have used new technology to reduce production costs, thus mitigating some of the problems of declining circulation, but these savings are becoming exhausted. They realize that, if they want to grow their business or even sustain its current level, they need to have products for the expanding online market. Electronic information is seen as a promising growth market.

Most of the major publishers of scientific journals have moved rapidly to electronic publishing of their journals. The approach taken by the Association for Computing Machinery (ACM) is described in Panel 3.4, but, with only minor differences, the same strategy is being followed by the large commercial publishers, including Elsevier, John Wiley, and Academic Press, by societies such as the American Chemical Society, and by university presses, including M.I.T. Press.

Panel 3.4
The Association for Computing Machinery's Digital Library

The Association for Computing Machinery (ACM) is a professional society that publishes seventeen research journals in computer science. In addition, its thirty eight special interest groups run a wide variety of conferences, many of which publish proceedings. The ACM members are practicing computer scientists, including many of the people who built the Internet and the web. These members were some of the first people to become accustomed to communicating online and they expected their society to be a leader in the movement to online journal publication.

Traditionally, the definitive version of a journal has been a printed volume. In 1993, the ACM decided that its future production process would use a computer system that creates a database of journal articles, conference proceedings, magazines and newsletters, all marked up in SGML. Subsequently, ACM also decided to convert large numbers of its older journals to build a digital library covering its publications from 1985.

One use of the SGML files is a source for printed publications. However, the plan was much more progressive. The ACM planned for the day when members would retrieve articles directly from the online database, sometimes reading them on the screen of a computer, sometimes downloading them to a local printer. Libraries would be able to license parts of the database or take out a general subscription for their patrons.

The collection came online during 1997. It uses a web interface that offers readers the opportunity to browse through the contents pages of the journals, and to search by author and keyword. When an article has been identified, a subscriber can read the full text of the article. Other readers pay a fee for access to the full text, but can read abstracts without payment.

Business issues

ACM was faced with the dilemma of paying for this digital library. The society needed revenue to cover the substantial costs, but did not want to restrain authors and readers unduly. The business arrangements fall into two parts, the relationships with authors and with readers. Initially, both are experimental.

In 1994, ACM published an interim copyright policy, which describes the relationship between the society as publisher and the authors. It attempts to balances the interest of the authors against the needs for the association to generate revenue from its publications. It was a sign of the times, that ACM first published the new policy on its web server. One of the key features of this policy is the explicit acknowledgment that many of the journal articles are first distributed via the web.

To generate revenue, ACM charges for access to the full text of articles. Members of the ACM can subscribe to journals or libraries can subscribe on behalf of their users. The electronic versions of journals are priced about 20 percent below the prices of printed versions. Alternatively, individuals can pay for single articles. The price structure aims to encourage subscribers to sign up for the full set of publications, not just individual journals.

Electronic journals

The term **electronic journal** is commonly used to describe a publication that maintains many of the characteristics of printed journals, but is produced and distributed online. Rather confusingly, the same term is used for a journal that is purely digital, in that it exists only online, and for the digital version of a journal that is primarily a print publication, e.g., the ACM journals described in Panel 3.4.

Many established publishers have introduced a small number of purely online periodicals and there have been numerous efforts by other groups. Some of these online periodical set out to mimic the processes and procedures of traditional journals. Perhaps the most ambitions publication using this approach was the *On-line Journal of Current Clinical Trials* which was developed by the American Association for the Advancement of Science (AAAS) in conjunction with OCLC. Unlike other publications for which the electronic publication was secondary to a printed publication, this new journal was planned as a high-quality, refereed journal for which the definitive version would be the electronic version. Since the publisher had complete control over the journal, it was possible to design it and store it in a form that was tailored to electronic delivery and display, but the journal was never accepted by researchers or physicians. It came out in 1992 and never achieved its goals, because it failed to attract the numbers of good papers that were planned for. Such is the fate of many pioneers.

More recent electronic periodicals retain some characteristics of traditional journals but experiment with formats or services that take advantage of online publishing. In 1995, we created *D-Lib Magazine* at CNRI as an online magazine with articles and news about digital libraries research and implementation. The design of *D-Lib Magazine* illustrates a combination of ideas drawn from conventional publishing and the Internet community. Conventional journals appear in issues, each containing several articles. Some electronic journals publish each article as soon as it has is ready, but *D-Lib Magazine* publishes a monthly issue with strong emphasis on punctuality and rapid publication. The graphic design is deliberately flexible, while the constraints of print force strict design standards, but online publications can allow authors to be creative in their use of technology.

Research libraries and conversion projects

One of the fundamental tasks of research libraries is to save today's materials to be the long-term memory for tomorrow. The great libraries have wonderful collections that form the raw material of history and of the humanities. These collections consist primarily of printed material or physical artifacts. The development of digital libraries has created great enthusiasm for converting some of these collections to digital forms. Older materials are often in poor physical condition. Making a digital copy preserves the content and provides the library with a version that it can make available to the whole world. This section looks at two of these major conversion efforts.

Many of these digital library projects convert existing paper documents into bit-mapped images. Printed documents are scanned one page at a time. The scanned images are essentially a picture of the page. The page is covered by an imaginary grid. In early experiments this was often at 300 dots per inch, and the page was recorded as an array of black and white dots. More recently, higher resolutions and full color scanning have become common. Bit-mapped images of this kind are crisp enough that they can be displayed on large computer screens or printed on paper with good legibility. Since this process generates a huge number of dots for each page, various methods are used to compress the images which reduces the number of bits to be stored and the size of files to be transmitted across the network, but even the simplest images are 50,000 bytes per page.

Panel 3.5. American Memory and the National Digital Library Program

Background

The Library of Congress, which is the world's biggest library, has magnificent special collections of unique or unpublished materials. Among the Library's treasures are the papers of twenty three presidents. Rare books, pamphlets, and papers provide valuable material for the study of historical events, periods, and movements. Millions of photographs, prints, maps, musical scores, sound recordings, and moving images in various formats reflect trends and represent people and places. Until recently, anybody who wanted to use these materials had to come to the library buildings on Capitol Hill in Washington, DC.

American Memory was a pilot program that, from 1989 to 1994, reproduced selected collections for national dissemination in computerized form. Collections were selected for their value for the study of American history and culture, and to explore the problems of working with materials of various types, such as prints, negatives, early motion pictures, recorded sound, and textual documents. Initially, American Memory used a combination of digitized representations on CD-ROM and analog forms on videodisk, but, in June 1994, three collections of photographs were made available on the web.

The National Digital Library Program (NDLP) builds on the success of American Memory. Its objective is to convert millions of items to digital form and make them available over the Internet. The focus is on Americana, materials that are important in American history. Typical themes are Walt Whitman's notebooks, or documents from the Continental Congress and the Constitutional Convention.

Some of the collections that are being converted are coherent archives, such as the papers or the photograph collection of a particular individual or organization. Some are collections of items in a special original form, such as daguerreotypes or paper prints of early films. Others are thematic compilations by curators or scholars, either from within an archival collection or selected across the library's resources.

Access to the collections

American Memory discovered the enthusiasm of school teachers for access to these primary source materials and the new program emphasizes education as its primary, but certainly not its only audience. The most comprehensive method of access to the collections is by searching an index. However, many of the archival collections being converted do not have catalog records for each individual item. The collections have finding aids. These are structured documents that describe the collection, and groups of items within it, without providing a description for each individual item.

Hence, access to material in American Memory is a combination of methods, including searching bibliographic records for individual items where such records exist, browsing subject terms, searching full text, and, in the future, searching the finding aids.

Technical considerations

This is an important project technically, because of its scale and visibility. Conscious of the long-term problems of maintaining large collections, the library has placed great emphasis on how it organizes the items within its collections.

The program takes seriously the process of converting these older materials to digital formats, selecting the most appropriate format to represent the content, and putting great emphasis on quality control. Textual material is usually converted twice: to a scanned page image, and an SGML marked-up text. Several images are made from each photograph, ranging from a small thumbnail to a high-resolution image for archival purposes.

Many of the materials selected for conversion are free from copyright or other restrictions on distribution, but others have restrictions. In addition to copyright, other reasons for restrictions include conditions required by donors of the original materials to the library. Characteristic of older materials, especially unpublished items, is that it is frequently impossible to discover all the restrictions that might conceivably apply, and prohibitively expensive to make an exhaustive search for every single item. Therefore the library's legal staff has to develop policies and procedures that balance the value to the nation of making materials available, against the risk of inadvertently infringing some right.

Outreach

A final, but important, aspect of American Memory is that people look to the Library of Congress for leadership. The expertise that the library has developed and its partnerships with leading technical groups permit it to help the entire library community move forward. Thus the library is an active member of several collaborations, has overseen an important grant program, and is becoming a sponsor of digital library research.

Since the driving force in electronic libraries and information services has come from the scientific and technical fields, quite basic needs of other disciplines, such as character sets beyond English, have often been ignored. The humanities have been in danger of being left behind, but a new generation of humanities scholars is embracing computing. Fortunately, they have friends. Panel 3.6 describes JSTOR, a project of the Andrew W. Mellon Foundation which is both saving costs for libraries and bringing important journal literature to a wider audience than would ever be possible without digital libraries.

Panel 3.6. JSTOR

JSTOR is a project that was initiated by the Andrew W. Mellon Foundation to provide academic libraries with back runs of important journals. It combines both academic and economic objectives. The academic objective is to build a reliable archive of important scholarly journals and provide widespread access to them. The economic objective is to save costs to libraries by eliminating the need for every library to store and preserve the same materials.

The JSTOR collections are developed by fields, such as economics, history, and philosophy. The first phase is expected to have about one hundred journals from some fifteen fields. For each journal, the collection consists of a complete run, usually from the first issue until about five years before the current date.

The economic and organizational model

In August 1995, JSTOR was established as an independent not-for-profit organization with a goal to become self-sustaining financially. It aims to do so by charging fees for access to the database to libraries around the world. These fees are set to be less than the comparable costs to the libraries of storing paper copies of the journals.

The organization has three offices. Its administrative, legal and financial activities are managed and coordinated from the main office in New York, as are relationships with publishers and libraries. In addition, staff in offices at the University of Michigan and Princeton University maintain two synchronized copies of the database, maintain and develop JSTOR's technical infrastructure, provide support to users, and oversee the conversion process from paper to computer formats. Actual scanning and keying services are provided by outside vendors. JSTOR has recently established a third database mirror site at the University of Manchester, which supplies access to higher education institutions in the United Kingdom.

JSTOR has straightforward licenses with publishers and with subscribing institutions. By emphasizing back-runs, JSTOR strives not to compete with publishers, whose principal revenues come from current issues. Access to the collections has initially been provided only to academic libraries who subscribe to the entire collection. They pay a fee based on their size. In the best Internet tradition, the fee schedule and the license are available online for anybody to read.

The technical approach

At the heart of the JSTOR collection are scanned images of every page. These are scanned at a high resolution, 600 bits per inch, with particular emphasis on quality control. Unlike some other projects, only one version of each image is stored. Other versions, such as low-resolution thumbnails, are computed when required, but not stored. Optical character recognition with intensive proof reading is used to convert the text. This text is used only for indexing. In addition, a table of contents file is created for each article. This includes bibliographic citation information with keywords and abstracts if available.

These two examples, American Memory and JSTOR, are further examples of a theme that has run throughout this chapter. Libraries, by their nature, are conservative organizations. Collections and their catalogs are developed over decades or centuries. New services are introduced cautiously, because they are expected to last for a long time. However, in technical fields, libraries have frequently been adventurous. MARC, OCLC, the Linked Systems Project, Mercury, CORE, and the recent conversion projects may not have invented the technology they used, but the deployment as large-scale, practical systems was pioneering.

Chapter 2 discussed the community that has grown up around the Internet and the web. Many members of this community discovered online information very recently, and act as though digital libraries began in 1993 with the release of Mosaic. As discussed in this chapter, libraries and publishers developed many of the concepts that are the basis for digital libraries, years before the web. Combining the two communities of expertise provides a powerful basis for digital libraries in the future.

Chapter 4
Innovation and research

The research process

Innovation is a theme that runs throughout digital libraries and electronic publishing. Some of this innovation comes from the marketplace's demand for information, some from systematic research in universities or corporations. This chapter looks at the process of research and provides an overview of areas of current research. Most of the topics that are introduced in this chapter are discussed in greater detail later in the book.

Innovation in libraries and publishing

Until recently, innovation by libraries and publishers was far from systematic. Most had no budget for research and development. The National Library of Medicine, which has supported digital libraries since the 1960s, and OCLC's Office of Research were notable exceptions; both carry out research in the true sense of exploring new ideas without preconceptions of how the ideas will eventually be deployed, but they have been most unusual. Other libraries and publishers have an approach to innovation that is motivated by more practical considerations, how to provide enhanced services and products in the near term. The rate of innovation is controlled by the state of technology and the availability of funds, but also by an interpretation of the current demand for services.

Although they are well-organized to bring new ideas to the market, publishers carry out little research. Advanced projects are seen as business development and the large publishers have the resources to undertake substantial investments in new ventures. Chapter 3 looked at the movement to make scientific journals available online, illustrated by the Association for Computing Machinery's digital library. This can not be called research, but it is a complex task to change the entire publishing process, so that journals are created in computer formats, such as SGML, which can be read online or used as the source of printed journals.

Libraries tend to be more innovative than publishers, though they often appear poorly organized in their approach to change. Typically, they spend almost their entire budget on current activities. Innovation is often treated as an extra, not as the key to the future with its own resources. Large libraries have large budgets. Many employ several hundred people. Yet the budget systems are so inflexible that research and innovation are grossly under-staffed.

Thus, at the Library of Congress, the National Digital Library Program is a vital part of the library's vision and may be the most important library project in the United States. The program is a showcase of how the library can expand beyond its traditional role, by managing its collections differently, handling new types of material, and providing wider access to its collections. Yet the Library of Congress provides little financial support to the project. The program raises most of its funds from private foundations and other gifts, and is staffed by people on short term contracts. The Library of Congress does support technical areas, such as Z39.50 and MARC. However, the library's most visionary project has no long-term funding.

Despite such apparent conservatism, libraries are moving ahead on a broad front, particularly in universities. The structure of university libraries inhibits radical change, but university librarians know that computing is fundamental to the future of scholarly communication. Computer scientists can be seen as the hares of digital libraries, leaping ahead, trying out new fields, then jumping somewhere else. The large libraries are the tortoises. They move more slowly, but at each step they lay a foundation that can be built on for the next. The steps often take the form of focused projects, sometimes in partnership with other organizations, often with grants from foundations. The individual projects may not have the visibility of the big government-funded research initiatives, but the collective impact may well have an equal importance in the long term.

Many notable projects have their origins in university libraries. Some are converting materials to digital formats; others are working with publishers to make materials accessible online. Several projects that were mentioned in Chapter 3 are of this type, including HighWire Press at Stanford University which is putting scientific journals online, the collaboration of university libraries and Elsevier Science in the Tulip project to explore digitized version of scientific journals, and the contribution of the University of Michigan and Princeton University to the JSTOR project to convert historic back-runs of important journals. Another partnership, between Rutgers University and Princeton, created the Center for Electronic Texts in the Humanities. Each of these projects came out of the university libraries.

Most projects are made possible by funds from foundations, industry, or the government. For example, our Mercury project at Carnegie Mellon received major grants from the Pew Charitable Trusts, Digital Equipment Corporation, and DARPA, with smaller, but welcome, support from other donors. Several private foundations have been strong supporters of digital libraries for the humanities, notably the Andrew W. Mellon Foundation, and the J. Paul Getty Trust, which specializes in art and art history information. Finally, although its budget is small compared with the NSF or DARPA, the National Endowment for the Humanities devotes a significant proportion of its funds to digital libraries.

Panel 4.1
The Coalition for Networked Information

A focal point for innovation amongst libraries and publishers in the United States is the Coalition for Networked Information, known as CNI. Most of the people at the twice yearly meetings are from university libraries or computing centers, but the meetings also attract key individuals from publishers, computer companies, national libraries, and the government. In recent years, many people from outside the United States have attended.

The coalition is a partnership of the Association of Research Libraries and Educause. It was founded in March 1990 with a mission to help realize the promise of high-performance networks and computers for the advancement of scholarship and the enrichment of intellectual productivity. More than two hundred institutions are member of the Coalition Task Force. They include higher education institutions, publishers, network service providers, computer hardware, software, and systems companies, library networks and organizations, and public and state libraries.

In 1991, several years before web browsers were introduced, CNI set an example, by being one of the first organizations to create a well-managed information service on the Internet. The following list is taken from its web site. It is a sample of the coalition's projects over the years since its founding. Some of these projects had many participants and lasted for several years, while others were a single event, such as a conference. All of them made their impact by bringing together people from various fields to work side-by-side on shared problems.

- Access to networked government information via the Internet.
- Federal information for higher education.
- Evaluation of how networks have affected academic institutions.
- Authentication, authorization, and access management.
- Cost centers and measures in the networked information value chain.
- Consortium for university printing and information distribution.
- Capture and storage of electronic theses and dissertations.
- Electronic billboards on the digital superhighway.
- 51 reasons to invest in the national information infrastructure.
- The Government Printing Office wide information network data on-line act.
- Humanities and arts on the information highway.
- Institution-wide information policies.
- CNI/OCLC metadata workshop.
- National initiative for networked cultural heritage.
- Networked information discovery and retrieval.
- Creating new learning communities via the network.
- Technology, scholarship and the humanities.
- Rights for electronic access to and dissemination of information.
- Regional conferences.
- Scholarship from California on the net.
- Teaching and learning via the network.
- Protecting intellectual property in the networked multimedia environment.
- The transformation of the public library.
- Planning retreat for library and information technology professionals.
- Z39.50 resources.

This list illustrates CNI's broad ranging interests, which emphasize practical applications of digital libraries, collections, relationships between libraries and publishers, and policy issues of access to intellectual property. What the list does not show is the human side of CNI, established by its founding director, the late Paul Evan Peters, and continued by his successor, Clifford Lynch. CNI meetings are noted for the warmth with which people from different fields meet together. People whose professional interests may sometimes appear in conflict have learned to respect each other and come to work together. Progress over the past few years has been so rapid that it is easy to forget the vacuum that CNI filled by bringing people together to discuss their mutual interests in networked information.

Computer science research

Federal funding of research

Computer scientists take research seriously. They are accustomed to long-term projects, where the end result is not products or services, but new concepts or a deeper understanding of the field. In the United States, much of the money for research comes from federal agencies. The world's biggest sponsor of computer science research is DARPA, the Defense Advanced Research Projects Agency. (DARPA keeps changing its name. It was originally ARPA, changed to DARPA, back to ARPA, and back again to DARPA.) The next largest is the National Science Foundation (NSF). DARPA is a branch of the Department of Defense. Although ultimately its mission is to support the military, DARPA has always taken a broad view and encourages fundamental research in almost every aspect of computer science. It is particularly noted for its emphasis on research projects that build large, experimental systems. The NSF has a general responsibility for promoting science and engineering in the United States. Its programs invest over $3.3 billion per year in almost 20,000 research and education projects. Thus it supports research both in computer science and in applications of computing to almost every scientific discipline.

Many engineering and computer companies have large budgets for computing research and development, often as much as ten percent of total operations. Most industrial research is short term, aimed at developing new products, but fundamental advances in computing have come from industrial laboratories, such as Xerox PARC, Bell Laboratories, and IBM. More recently, Microsoft has established an impressive research team.

Much of the underlying technology that makes digital libraries feasible was created by people whose primary interests were in other fields. The Internet, without which digital libraries would be very different, was originally developed by DARPA (then known as ARPA) and the NSF. The web was developed at CERN, a European physics laboratory with substantial NSF funding. The first web browser, Mosaic, was developed at the NSF supercomputing center at the University of Illinois. Several areas of computer science research are important to information management, usually with roots going back far beyond the current interest in digital libraries; they include networking, distributed computer systems, multimedia, natural language processing, databases, information retrieval, and human-computer interactions. In addition to funding specific research, the federal agencies assist the development of research communities and often coordinate deployment of new technology. The NSF supported the early years of the Internet Engineering Task Force (IETF). The World Wide Web Consortium based at MIT is primarily funded by industrial associates, and also has received money from DARPA.

Digital libraries were not an explicit subject of federal research until the 1990s. In 1992, DARPA funded a project, coordinated by the Corporation for National Research Initiatives (CNRI), involving five universities: Carnegie Mellon, Cornell, M.I.T., Stanford, and the University of California at Berkeley. The project had the innocent name of the Computer Science Technical Reports project (CSTR), but its true impact was to encourage these strong computer science departments to develop research programs in digital libraries.

The initiative that really established digital libraries as a distinct field of research came in 1994, when the NSF, DARPA, and the National Aeronautic and Space Agency (NASA) created the **Digital Libraries Initiative**. The research program of the Digital Libraries Initiative is summarized in Panel 4.2. Research carried out under this program is mentioned throughout this book.

Panel 4.2. The Digital Libraries Initiative

In 1994, the computer science divisions of NSF, DARPA, and NASA provided funds for six research projects in digital libraries. They were four year projects. The total government funding was $24 million, but numerous external partners provided further resources which, in aggregate, exceeded the government funding. Each project was expected to implement a digital library testbed and carry out associated research.

The impact of these projects indicates that the money has been well-spent. An exceptionally talented group of researchers has been drawn to library research, with wisdom and insights that go far beyond the purely technical. This new wave of computer scientists who are carrying out research in digital libraries brings new experiences to the table, new thinking, and fresh ideas to make creative use of technology.

Here is a list of the six projects and highlights of some of the research. Many of these research topics are described in later chapters.

The University of California at Berkeley built a large collection of documents about the California environment. They included maps, pictures, and government reports scanned into the computer. Notable research included: multivalent documents which are a conceptual method for expressing documents as layers of information, Cheshire II a search system that combines the strengths of SGML formats with the information in MARC records, and research into image recognition so that features such as dams or animals can be recognized in pictures.

The University of California at Santa Barbara concentrated on maps and other geospatial information. Their collection is called the Alexandria Digital Library. Research topics included: metadata for geospatial information, user interfaces for overlapping maps, wavelets for compressing and transmitting images, and novel methods for analyzing how people use libraries.

Carnegie Mellon University built a library of segments of video, called Informedia. The research emphasized automatic processing for information discovery and display. It included: multi-modal searching in which information gleaned from many sources is combined, speech recognition, image recognition, and video skimming to provide a brief summary of a longer video segment.

The University of Illinois worked with scientific journal publishers to build a federated library of journals for science and engineering. Much of the effort concentrated on manipulation of documents in SGML. This project also used supercomputing to study the problems of semantic information in very large collections of documents.

The University of Michigan has built upon the digital library collections being developed by the university libraries. The project has investigated applications in education, and experimented with economic models and an agent-based approach to interoperability.

Stanford University concentrated on computer science literature. At the center of the project was the InfoBus, a method to combine digital library services from many sources to provide a coherent set of services. Other research topics included novel ways to address user interfaces, and modeling of the economic processes in digital libraries.

The Digital Libraries Initiative focused international attention on digital libraries research. Beyond the specific work that it funded, the program gave a shape to an emerging discipline. Research in digital libraries was not new, but previously it had been fragmented. Even the name "digital libraries" was uncertain. The Digital Libraries Initiative highlighted the area as a challenging and rewarding field of research. It led to the growth of conferences, publications, and the appointment of people in academic departments who describe their research interests as digital libraries. This establishment of a new field is important because it creates the confidence that is needed to commit to long-term research.

Another impact of the Digital Libraries Initiative has been to clarify the distinction between research and the implementation of digital libraries. When the initiative was announced, some people thought that the federal government was providing a new source of money to build digital libraries, but these are true research projects. Some of the work has already migrated into practical applications; some anticipates hardware and software developments; some is truly experimental.

The emergence of digital libraries as a research discipline does have dangers. There is a danger that researchers will concentrate on fascinating theoretical problems in computer science, economics, sociology, or law, forgetting that this is a practical field where research should be justified by its utility. The field is fortunate that the federal funding agencies are aware of these dangers and appear determined to keep a focus on the real needs.

Agencies, such as the NSF, DARPA, and NASA, are granted budgets by Congress to carry out specific objectives and each agency operates within well-defined boundaries. None of the three agencies has libraries as a primary mission; the money for the Digital Libraries Initiative came from existing budgets that Congress had voted for computer science research. As a result, the first phase of research emphasized the computer science aspects of the field, but the staff of the agencies know that digital libraries are more than a branch of computer science. When in 1998, they created a second phase of the initiative, they sought partners whose mission is to support a broader range of activities. The new partners include the National Library of Medicine, the National Endowment for the Humanities, the Library of Congress, and the NSF's Division of Undergraduate Education. This book was written before these new grants were announced, but everybody expects to see funding for a broad range of projects that reflect the missions of all these agencies.

To shape the agenda, the funding agencies have sponsored a series of workshops. Some of these workshops have been on specific research topics, such as managing access to information. Others have had a broader objective, to develop a unified view of the field and identify key research topics. The small amounts of federal funding used for coordination of digital libraries research have been crucial in shaping the field. The key word is "coordination", not standardization. Although the output sometimes includes standards, the fundamental role played by these efforts is to maximize the impact of research, and build on it.

Areas of research

The central part of this chapter is a quick overview of the principal areas of research in digital libraries. As digital libraries have established themselves as a field for serious research, certain problems have emerged as central research topics and a body of people now work on them.

Object models

One important research topic is to understand the objects that are in digital libraries. Digital libraries store and disseminate any information that can be represented in digital form. As a result, the research problems in representing and manipulating information are varied and subtle.

The issue is an important one. What users see as a single work may be represented in a computer as an assembly of files and data structures in many formats. The relationship between these components and the user's view of the object is sometimes called an **object model**. For example, in a digital library, a single image may be stored several times, as a high-quality archival image, a medium resolution version for normal use, and a small thumbnail that gives an impression of the image but omits many details. Externally, this image is referenced by a single bibliographic identifier but to the computer it is a group of distinct files. A journal article stored on a web server appears to a user as a single continuous text with a few graphics. Internally it is stored as several text files, several images, and perhaps some executable programs. Many versions of the same object might exist. Digital libraries often have private versions of materials that are being prepared for release to the public. After release, new versions may be required to correct errors, the materials may be reorganized or moved to different computers, or new formats may be added as technology advances.

The ability of user interfaces and other computer programs to present the work to a user depends upon being able to understand how these various components relate to form a single library object. Structural metadata is used to describe the relationships. Mark-up languages are one method to represent structure in text. For example, in an HTML page, the <img> tag is structural metadata that indicates the location of an image. The simplicity of HTML makes it a poor tool for many tasks, whereas the power of SGML is overly complex for most purposes. A new hybrid, XML, has recently emerged that may succeed in bridging the gap between HTML and the full generality of SGML.

Much of the early work on structural metadata has been carried out with libraries of digitized pictures, music, video clips, and other objects converted from physical media. Maps provide an interesting field of their own. Looking beyond conventional library materials, content created in digital form is not constrained by linearity of print documents. There is research interest in real-time information, such as that obtained from remote sensors, in mobile agents that travel on networks, and on other categories of digital objects that have no physical counterpart. Each of these types has its problems, how to capture the information, how to store it, how to describe it, how to find the information that it contains, and how to deliver it. These are tough questions individually and even tougher in combination. The problem is to devise object models that will support library materials that combine many formats, and will enable independent digital libraries to interoperate.

User interfaces and human-computer interaction

Improving how users interact with information on computers is clearly a worthwhile subject. This is such a complex topic that it might be thought of as an art, rather than a field where progress is made through systematic research. Fortunately, such pessimism has proved unfounded. The development of web browsers is an example of creative research in areas such as visualization of complex sets of information, layering of the information contained in documents, and automatic skimming to extract a summary or to generate links.

To a user at a personal computer, digital libraries are just part of the working environment. Some user interface research looks at the whole environment, which is likely to include electronic mail, word processing, and applications specific to the individual's field of work. In addition, the environment will probably include a wide range of information that is not in digital form, such as books, papers, video tapes, maps, or photographs. The ability of users to interact with digital objects, through annotations, to manipulate them, and to add them to their own personal collections is proving to be a fertile area of research.

Information discovery

Finding and retrieving information is a central aspects of libraries. Searching for specific information in large collections of text, the field known as information retrieval, has long been of interest to computer scientists. Browsing has received less research effort despite its importance. Digital libraries bring these two areas together in the general problem of information discovery, how to find information. Enormous amounts of research is being carried out in this area and only a few topics are mentioned here.

- **Descriptive metadata** Most of the best systems for information discovery use cataloguing or indexing metadata that has been produced by somebody with expert knowledge. This includes the data in library catalogs, and abstracting and indexing services, such as Medline or Inspec. Unfortunately, human indexing is slow and expensive. Different approaches are required for the huge volumes of fast-changing material to be expected in digital libraries. One approach is for the creator to provide small amounts of descriptive metadata for each digital object. Some of this metadata will be generated automatically; some by trained professionals; some by less experienced people. The metadata can then be fed into an automatic indexing program.

- **Automatic indexing.** The array of information on the networks is too vast and changes too frequently for it all to be catalogued by skilled cataloguers. Research in automatic indexing uses computer programs to scan digital objects, extract indexing information and build searchable indexes. Web search programs, such as Altavista, Lycos, and Infoseek are the products of such research, much of which was carried out long before digital libraries became an established field.

- **Natural language processing.** Searching of text is greatly enhanced if the search program understands some of the structure of language. Relevant research in computational linguistics includes automatic parsing to identify grammatical constructs, morphology to associate variants of the same word, lexicons, and thesauruses. Some research goes even further, attempting to bring knowledge of the subject matter to bear on information retrieval.

- **Non-textual material.** Most methods of information discovery use text, but researchers are slowly making progress in searching for specific content in other formats. Speech recognition is just beginning to be usable for indexing radio programs and the audio track of videos. Image recognition, the automatic extraction of features from pictures, is an active area of research, but not yet ready for deployment.

Collection management and preservation

Collection management is a research topic that is just beginning to receive the attention than it deserves. Over the years, traditional libraries have developed methods that allow relatively small teams of people to manage vast collections of material, but early digital libraries have often been highly labor intensive. In the excitement of creating digital collections, the needs of organizing and preserving the materials over long periods of time were neglected. These topics are now being recognized as difficult and vitally important.

- **Organization of collections.** The organization of large collections of online materials is complex. Many of the issues are the same whether the materials are an electronic journal, a large web site, a software library, an online map collection, or a large information service. They include how to load information in varying formats, and how to organize it for storage and retrieval. For access around the world, several copies are needed, using various techniques of replication. The problems are amplified by the fact that digital information changes. In the early days of printing, proof corrections were made continually, so that every copy of a book might be slightly different. Online information can also change continually. Keeping track of minor variations is never easy and whole collections are reorganized at unpleasantly frequent intervals. Many of the research topics that are important for interoperability between collections are equally important for organizing large collections. In particular, current research on identifiers, metadata, and authentication applies to both collection management and interoperation among collections.

- **Archiving and preservation.** The long-term preservation of digital materials has recently emerged as a key research topic in collection management. Physical materials, such as printed books, have the useful property that they can be neglected for decades but still be readable. Digital materials are the opposite. The media on which data is stored have quite short life expectancies, often frighteningly short. Explicit action must be taken to refresh the data by copying the bits periodically onto new media. Even if the bits are preserved, problems remain. The formats in which information is stored are frequently replaced by new versions. Formats for word processor and image storage that were in common use ten years ago are already obsolete and hard to use. To interpret archived information, future users will need to be able to recognize the formats and display them successfully.

- **Conversion.** Conversion of physical materials into digital formats illustrates the difficulties of collection management. What is the best way to convert huge collections to digital format? What is the trade off between cost and quality? How can today's efforts be useful in the long term?

This area illustrates the differences between small-scale and large-scale efforts, which is an important research topic in its own right. A small project may convert a few thousand items to use as a testbed. The conversion is perceived as a temporary annoyance that is necessary before beginning the real research. A group of students will pass the materials through a digital scanner, check the results for obvious mistakes, and create the metadata required for a specific project. In contrast, libraries and publishers convert millions of items. The staff carrying out the work is unlikely to be as motivated as members of a research team; metadata must be generated without knowing the long-term uses of the information; quality control is paramount. The current state of the art is that a number of organizations have developed effective processes for converting large volumes of material. Often, part of the work is shipped to countries where labor costs are low. However, each of these organizations has its own private method of working. There is duplication of tools and little sharing of experience.

Conversion of text is an especially interesting example. Optical character recognition, which uses a computer to identify the characters and words on a page, has reached a tantalizing level of being almost good enough, but not quite. Several teams have developed considerable expertise in deciding how to incorporate optical character recognition into conversion, but little of this expertise is systematic or shared.

Interoperability

From a computing viewpoint, many of the most difficult problems in digital libraries are aspects of a single challenge, **interoperability**, how to get a wide variety of computing systems to work together. This embraces a range of topics, from syntactic interoperability that provides a superficial uniformity for navigation and access, but relies almost entirely on human intelligence for coherence, to a deeper level of interoperability, where separate computer systems share an understanding of the information itself.

Around the world, many independently managed digital libraries are being created. These libraries have different management policies and different computing systems. Some are modern, state-of-the-art computer systems; others are elderly, long past retirement age. The term **legacy system** is often used to describe old systems, but this term is unnecessarily disparaging. As soon as the commitment is made to build a computer system, or create a new service or product, that commitment is a factor in all future decisions. Thus, every computer system is a legacy system, even before it is fully deployed.

Interoperability and standardization are interrelated. Unfortunately, the formal process of creating international standards is often the opposite of what is required for interoperability in digital libraries. Not only is the official processes of standardization much too slow for the fast moving world of digital libraries. The process encourages standards that are unduly complex and many international standards have never been tested in real life. In practice, the only standards that matter are those that are widely used. Sometimes a de facto standard emerges because a prominent group of researchers uses it; the use of TCP/IP for the embryonic Internet is an example. Some standards become accepted because the leaders of the community decide to follow certain conventions; the MARC format for catalog records is an example. Sometimes, generally accepted standards are created from a formal standards process; MPEG, the compression format used for video, is a good example. Other de facto standards are proprietary products from prominent corporations; Adobe's Portable Document Format (PDF) is a recent example. TCP/IP and MARC are typical of the standards that were initially created by the communities that use them, then became official standards that have been enhanced through a formal process.

The following list gives an idea of the many aspects of interoperability:

- **User interfaces.** A user will typically use many different digital library collections. Interoperability aims at presenting the materials from those collection in a coherent manner, though it is not necessary to hide all the differences. A collection of maps is not the same as a music collection, but the user should be able to move smoothly between them, search across them, and be protected from idiosyncrasies of computer systems or peculiarities of how the collections are managed.

- **Naming and identification.** Some means is needed to identify the materials in a digital library. The Internet provides a numeric identifier for every computer, an IP address, and the domain name system that identifies every computer on the Internet. The web's Uniform Resource Locator (URL) extends these names to individual files. However, neither domain names nor URLs are fully satisfactory. Library materials need identifiers that identify the material, not the location where an instance of the material is stored at a given moment of time. Location independent identifiers are sometimes called Uniform Resource Names (URNs).

- **Formats.** Materials in every known digital format are stored in digital libraries. The web has created de facto standards for a few formats, notably HTML for simple text, and GIF and JPEG for images. Beyond these basic formats, there is little agreement. Text provides a particular challenge for interoperability. During the 1980s, ASCII emerged as the standard character set for computers, but has few characters beyond those used in English. Currently, Unicode appears to be emerging as an extended character set that supports a very wide range of scripts, but is not yet supported by many computer systems. Although SGML has been widely advocated, and is used in some digital library systems, it is so complex and has so much flexibility, that full interoperability is hard to achieve.

- **Metadata.** Metadata plays an important role in many aspects of digital libraries, but is especially important for interoperability. As discussed earlier, metadata is often divided into three categories: descriptive metadata is used for bibliographic purposes and for searching and retrieval; structural metadata relates different objects and parts of objects to each other; administrative metadata is used to manage collections, including access controls. For interoperability, some of this metadata must be exchanged between computers. This requires agreement on the names given to the metadata fields, the format used to encode them, and at least some agreement on semantics. As a trivial example of the importance of semantics, there is little value in having a metadata field called "date" if one collection uses the field for the date when an object was created and another uses it for the date when it was added to the collection.

- **Distributed searching.** Users often wish to find information that is scattered across many independent collections. Each may be organized in a coherent way, but the descriptive metadata will vary, as will the capabilities provided for searching. The distributed search problem is how to find information by searching across collections. The traditional approach is to insist that all collections agree on a standard set of metadata and support the same search protocols. Increasingly, digital library researchers are recognizing that this is a dream world. It must be possible to search sensibly across collections despite differing organization of their materials.

- **Network protocols.** To move information from one computer to another, requires interoperability at the network level. The almost universal adoption of the Internet family of protocols has largely solved this problem, but there are gaps. For example, the Internet protocols are not good at delivering continuous streams of data, such audio or video materials, which must arrive in a steady stream at predictable time intervals.

- **Retrieval protocols.** One of the fundamental operations of digital libraries is for a computer to send a message to another to retrieve certain items. This message must be transmitted in some protocol. The protocol can be simple, such as HTTP, or much more complex. Ideally, the protocol would support secure authentication of both computers, high-level queries to discover what resources each provides, a variety of search and retrieval capabilities, methods to store and modify intermediate results, and interfaces to many formats and procedures. The most ambitious attempt to achieve these goals is the Z39.50 protocol, but Z39.50 is in danger of collapsing under its own complexity, while still not meeting all the needs.

- **Authentication and security.** Several of the biggest problems in interoperability among digital libraries involve authentication. Various categories of authentication are needed. The first is authentication of users. Who is the person using the library? Since few methods of authentication have been widely adopted, digital libraries are often forced to provide every user with a ID and password. The next category is authentication of computers. Systems that handle valuable information, especially financial transactions or confidential information, need to know which computer they are connecting to. A crude approach is to rely on the Internet IP address of each computer, but this is open to abuse. The final need is authentication of library materials. People need to be confident that they have received the authentic version of an item, not one that has been modified, either accidentally or deliberately. For some of these needs, good methods of authentication exist, but they are not deployed widely enough to permit full interoperability.

- **Semantic interoperability.** Semantic interoperability is a broad term for the general problem that, when computers pass messages, they need to share the same semantic interpretation of the information in the messages. Semantic interoperability deals with the ability of a user to access similar classes of digital objects, distributed across heterogeneous collections, with compensation for site-by-site variations. Full semantic interoperability embraces a family of deep research problems. Some are extraordinarily difficult.

The web provides a base level of interoperability, but the simplicity of the underlying technology that has led to its wide acceptance also brings weaknesses. URLs make poor names for the long term; HTML is restricted in the variety of information that it can represent; MIME, which identifies the type of each item, is good as far as it goes, but library information is far richer than the MIME view of data types; user interfaces are constrained by the simplicity of the HTTP protocol. Developing extensions to the web technology has become big business. Some extensions are driven by genuine needs, but others by competition between the companies involved. A notable success has been the introduction of the Java programming language, which has made a great contribution to user interfaces, overcoming many of the constraints of HTTP.

Paradoxically, the web's success is also a barrier to the next generation of digital libraries. It has become a legacy system. The practical need to support this installed base creates a tension in carrying out research. If researchers wish their work to gain acceptance, they must provide a migration path from the web of today. As an example, the fact that the leading web browsers do not support URNs has been a barrier to using URNs to identify materials within digital libraries.

The secret of interoperability is easy to state but hard to achieve. Researchers need to develop new concepts that offer great improvements, yet are easy to introduce. This requires that new methods have high functionality and low cost of adoption. A high-level of functionality is needed to overcome the inertia of the installed base. Careful design of extensibility in digital library systems allows continued research progress with the least disruption to the installed base.

Scaling

Interoperability and collection management are two examples of problems that grow rapidly as the scale of the library increases. A user may find more difficulty in using a monograph catalog in a very large library, such as the Library of Congress or Harvard University, than in a small college library where there are only a few entries under each main heading. As the size of the web has grown, many people would agree that the indexing programs, such as Infoseek, have become less useful. The programs often respond to simple queries with hundreds of similar hits; valuable results are hard to find among the duplicates and rubbish. Scaling is a difficult topic for research, without building large-scale digital libraries. Currently the focus is on the technical problems, particularly reliability and performance.

Questions of reliability and robustness of service pervade digital libraries. The complexity of large computer systems exceeds the ability to understand fully how all the parts interact. In a sufficiently large system, inevitably, some components are out of service at any given moment. The general approach to this problem is to duplicate data. Mirror sites are often used. Unfortunately, because of the necessary delays in replicating data from one place to another, mirror sites are rarely exact duplicates. What are the implications for the user of the library? What are the consequences for distributed retrieval if some part of the collections can not be searched today, or if a back-up version is used with slightly out-of-date information?

Research in performance is a branch of computer networking research, not a topic peculiar to digital libraries. The Internet now reaches almost every country of the world, but it is far from providing uniformly high performance everywhere, at all times. One basic technique is caching, storing temporary copies of recently used information, either on the user's computer or on a close by server. Caching helps achieve decent performance across the world-wide Internet, but brings some problems. What happens if the temporary copies are out of date? Every aspect of security and control of access is made more complex by the knowledge that information is likely to be stored in insecure caches around the world. Interesting research on performance has been based on the concept of locality. Selected information is replicated and stored at a location that has been chosen because it has good Internet connections. For example, the Networked Computer Science Technical Reference Library (NCSTRL), described in Chapter 11, uses a series of zones. Everything needed to search and identify information is stored within a zone. The only messages sent outside the zone are to retrieve the actual digital objects from their home repositories.

Economic, social, and legal issues

Digital libraries exist within a complex social, economic, and legal framework, and succeed only to the extent that they meet these broader needs. The legal issues are both national and international. They range across several branches of law: copyright, communications, privacy, obscenity, libel, national security, and even taxation. The social context includes authorship, ownership, the act of publication, authenticity, and integrity. These are not easy areas for research.

Some of the most difficult problems are economic. If digital libraries are managed collections of information, skilled professionals are needed to manage the collections. Who pays these people? The conventional wisdom assumes that the users of the collections, or their institutions, will pay subscriptions or make a payment for each use of the collections. Therefore, there is research into payment methods, authentication, and methods to control the use made of collections. Meanwhile, the high quality of many open-access web sites has shown that there are other financial models. Researchers have developed some interesting economic theories, but the real advances in understanding of the economic forces come from the people who are actually creating, managing, or using digital information. Pricing models cease to be academic when the consequence of a mistake is for individuals to lose their jobs or an organization to go out of business.

Access management is a related topic. Libraries and publishers sometimes wish to control access to their materials. This may be to ensure payment, requirements from the copyright owners, conditions laid down by donors, or a response to concerns of privacy, libel, or obscenity. Such methods are sometimes called "rights management", but issues of access are much broader than simply copyright control or the generation of revenue. Some of the methods of access management involve encryption, a highly complex field where issues of technology, law, and public policy have become hopelessly entangled.

A related research topic is evaluation of the impact made by digital libraries. Can the value of digital libraries and research into digital libraries be measured? Unfortunately, despite some noble efforts, it is not clear how much useful information can be acquired. Systematic results are few and far between. The problem is well-known in market research. Standard market research techniques, such as focus groups and surveys, are quite effective in predicting the effect of incremental changes to existing products. The techniques are much less effective in anticipating the impact of fundamental changes. This does not imply that measurements are not needed. It is impossible to develop any large scale system without good management data. How many people use each part of the service? How satisfied are they? What is the unit cost of the services? What is the cost of adding material to the collections? What are the delays? Ensuring that systems provide such data is essential, but this is good computing practice, not a topic for research.

Economic, social and legal issues were left to the end of this survey of digital libraries research, not because they are unimportant but because they are so difficult. In selecting research topics, two criteria are important. The topic must be worthwhile and the research must be feasible. The value of libraries to education, scholarship, and the public good is almost impossible to measure quantitatively. Attempts to measure the early impact of digital libraries are heavily constrained by incomplete collections, rapidly changing technology, and users who are adapting to new opportunities. Measurements of dramatically new computer systems are inevitably a record of history, interesting in hind sight, but of limited utility in planning for the future. Determining the value of digital libraries may continue to be a matter for informed judgment, not research.

Research around the world

This overview of digital library research and innovation is written from an American perspective, but digital libraries are a worldwide phenomenon. The Internet allows researchers from around the world to collaborate on a day to day basis. Researchers from Australia and New Zealand have especially benefited from these improved communications, and are important contributors. The web itself was developed in Switzerland. The British e-lib project provided added stimuli to a variety of library initiatives around the theme of electronic publication and distribution of material in digital forms. Recently, the European Union and the National Science Foundation have sponsored a series of joint planning meetings. A notable international effort has been the series of Dublin Core metadata workshops, which are described in Chapter 10. The stories on digital libraries research published in *D-Lib Magazine* each month illustrate that this is indeed a worldwide field; during the first three years of publication, articles came from authors in more than ten countries. The big, well-funded American projects are important, but they are far from being the whole story.

Chapter 5
People, organizations, and change

People and change

This is the first of four chapters that examine the relationship between people, organizations, and digital libraries. This chapter examines how individuals and organizations are responding to the changes brought by new technology. Chapter 6 looks more specifically at economic and legal issues. It is followed by a chapter on the management of access to digital libraries, and related issues of security. Finally, Chapter 8 examines user interfaces, which are the boundary between people and computers.

The story of digital libraries is one of change. Authors, readers, librarians, publishers and information services are adopting new technology with remarkable speed; this is altering relationships among people. Every organization has some members who wish to use the most advanced systems (even when not appropriate) and others who demand the traditional ways of doing things (even when the new ones are superior). This is sometimes called the "generation gap", but the term is a misnomer, since people of all ages can be receptive to new ideas. The pace of change, also, differs widely amongst organizations and disciplines. Some corporate libraries, such as those of drug companies, already spend more than half their acquisitions budgets on electronic materials and services, while, for the foreseeable future, humanities libraries will be centered around collections of printed material, manuscripts, and other tangible items.

Perhaps the most fundamental change is that computing is altering the behavior of the creators and users of information. Tools on personal computers allow individuals with a modicum of skill to carry out processes that previously required skilled craftsmen. Word processing and desktop publishing have made office correspondence almost as elegant as professionally designed books. Figures, graphs, and illustrations can be created in full color. Not every creator wishes to learn these techniques, and some privately produced efforts have miserably poor design, but many people create elegant and effective materials with no professional assistance. This has impact on every information professional - publisher, librarian, archivist, indexer and cataloguer, or webmaster.

A few people argue that the new technology removes the need for professionals to manage information. That is naive. Publishers and libraries perform functions that go far beyond the management of physical items. Services such as editing and refereeing, or abstracting and indexing are not tied to any particular technology. Although the web permits users to mount their own information, most people are pleased to have support from a professional webmaster. The overall need for information professionals will continue, perhaps grow, even as their specific practices change with the technology, but the new forms of organizations and the new types of professional that will emerge are open to speculation.

Digital libraries created by the users

Some of the most successful digital libraries were created by researchers or groups of professionals for themselves and their colleagues, with minimal support from publishers or librarians. Chapter 2 described two of these, the Physics E-Print Archives at the Los Alamos National Laboratory and the Internet RFC series. The panels in this section describe three more: the Netlib library of mathematical software, the data archives of the International Consortium for Political Science Research, and the Perseus collections of classical texts. These digital library collections are well-established and heavily used. They employ professionals, but the leadership and most of the staff come from the respective disciplines of physics, computing, applied mathematics, the social sciences, and classics.

Digital libraries that were created by the user communities are particularly interesting because the services have been constructed to meet the needs of the disciplines, without preconceived notions of how collections are conventionally managed. When creators and users develop the systems that they want for their own work, they encounter the normal questions about organizing information, retrieving it, quality control, standards, and services that are the lifeblood of publishing and libraries. Sometimes, they find new and creative answers to these old questions.

Panel 5.1.
Netlib

Netlib is a digital library that provides high-quality mathematical software for researchers. It was founded in 1985 by Jack Dongarra and Eric Grosse, who have continued as editors-in-chief. It is now maintained by a consortium which is based on Bell Laboratories, and the University of Tennessee and Oak Ridge National Laboratory, with mirror sites around the world.

The original focus of Netlib was the exchange of software produced from research in numerical analysis, especially software for supercomputers with vector or parallel architectures. The collections now include other software tools, technical reports and papers, benchmark performance data, and professional information about conferences and meetings. Most of the material in Netlib is freely available to all, but some programs have licensing restrictions, e.g., payment is required for commercial use.

The technical history of Netlib spans a period of rapid development of the Internet. Beginning with an electronic mail service, at various times Netlib has provided an X-Windows interface, anonymous FTP, CD-ROMs, and Gopher services. Currently, it uses web technology. The Netlib team continues to be among the leaders in developing advanced architectures for organizing and storing materials in digital libraries.

The organization of the Netlib collections is highly pragmatic. It assumes that the users are mathematicians and scientists who are familiar with the field and will incorporate the software into their own computer programs. The collections are arranged in a hierarchy, with software grouped by discipline, application, or source. Each collection has its own editor and the editors use their knowledge of the specific field to decide the method of organization. Netlib has developed a form of indexing record that is tailored to its specialized needs and the collections are also classified under the Guide to Available Mathematical Software (GAMS), which is a cross-index provided by the National Institute of Standards and Technology.

Netlib is a success story. Hundreds of thousands of software programs are downloaded from its sites every year, contributing to almost every branch of scientific research.

Panel 5.2
Inter-university Consortium for Political and Social Research

The Inter-university Consortium for Political and Social Research (ICPSR), based at the University of Michigan, has been in continuous operation since 1962. The archive provides social scientists with a digital library to store collections of data that they have gathered, for others to use. Data that was expensive to gather will lose its usefulness unless it is organized, documented, and made available to researchers. The ICPSR collects research data in a broad range of disciplines, including political science, sociology, demography, economics, history, education, gerontology, criminal justice, public health, foreign policy, and law.

About two hundred data sets are added to the archive every year, with several thousand files of data. Some of these data sets are very large. In addition to the data itself, the archive stores documentation about the data, and codebooks, which explain the design of the study, decisions made by the original researchers, how they gathered the data, any adjustments made, and the technical information needed to used it in further research. The collection is organized hierarchically by discipline for easy browsing. Each data set is described by a short record which contains basic cataloguing information and an abstract.

The archive has been in existence through many generations of computer system. Currently it has a web-based user interface, which can be used for browsing or searching the catalog records. Data sets are delivered over the Internet by FTP, with selected data available on CD-ROM.

ICPSR is a not-for-profit consortium. Hundreds of colleges and universities in the United States and around the world are members. Their annual subscriptions provide access to the collections for their faculty and students. People whose organizations do not belong to ICPSR can pay for the use of individual data sets.

Libraries and museums play a special role for users in the humanities, because they provide the raw material on which the humanities are based. Digital libraries can provide much wider access to these materials than could ever be provided by physical collections. A university with a fine library will always have an advantage in humanities teaching and research, but it need not have exclusive use of unique items. The British Library is beginning to digitize and mount on the web treasures, such as the Magna Carta and the manuscript of Beowulf. In the past, access to such documents was restricted to scholars who visited the library or to occasional, expensive facsimile editions. In the future, everybody will be able to see excellent reproductions. Panel 5.3 describes Perseus, a digital library of classical materials organized to be accessible to people who are not specialists in the field.

Panel 5.3.
Perseus

Some of the leading projects in electronic information have been led by established faculty in the humanities, but many have been maverick projects with little institutional support. Sometimes junior faculty members have pursued new ideas against the opposition of senior members in their departments. In the mid-1980s, while a junior faculty member in classics at Harvard University, Gregory Crane began work on the project known as Perseus, which uses hyperlinks to relate sources such as texts and maps with tools such as dictionaries. In particular, Crane aimed to give the general student an appreciation of the poems of the Greek poet Pindar. From this early work has emerged one of the most important digital libraries in the humanities.

The collections now have comprehensive coverage of the classical Greek period and are extending steadily into other periods of Greek history, the Roman period, and beyond. Source materials include texts, in both the original language and in translation, and images of objects, such as vases, and architectural sites. However, perhaps the greatest resource is the effort that has been made in structuring the materials and the database of links between items.

In Perseus, an emphasis on the content has enabled the collections to migrate through several generations of computer system. Classical texts are fairly stable, though new editions may have small changes, and supporting works such as lexicons and atlases have a long life. Therefore, the effort put into acquiring accurate versions of text, marking them up with SGML, and linking them to related works is a long term investment that will outlive any computer system. Perseus has never had more than one programmer on its staff, but relies on the most appropriate computer technology available. It was an early adopter of Apple's Hypercard system, published a high-quality CD-ROM, and quickly moved to the web when it emerged. The only elaborate software that the project has developed are rule-based systems to analyze the morphology of inflected Greek and Latin words.

The long-term impact of Perseus is difficult to predict, but its goals are ambitious. In recent years, academic studies in the humanities have become increasingly esoteric and detached. It is typical of the field that Crane was unable to continue his work at Harvard, because it was not considered serious scholarship; he moved to Tufts University. Perseus may not be Harvard's idea of scholarship but it is certainly not lightweight. The four million words of Greek source texts include most of the commonly cited texts; when there are no suitable images of a vase, Perseus has been know to take a hundred new photographs; the user interface helps the reader with easy access to translations and dictionaries, but has a strong focus on the original materials. Perseus is a treasure trove for the layman and is increasingly being used by researchers as an excellent resource for traditional studies. Hopefully, Perseus's greatest achievement will be to show the general public the fascination of the humanities and to show the humanities scholar that popularity and scholarship can go hand in hand.

The motives of creators and users

Creators

An understanding of how libraries and publishing are changing requires an appreciation of the varied motives that lead people to create materials and others to use them. A common misconception is that people create materials primarily for the fees and royalties that they generate. While many people make their livelihood from the works that they create, other people have different objectives.

Look at the collections in any library; large categories of material were created for reasons where success is measured by the impact on the readers, not by revenue. Charles Darwin's *The Origin of Species* was written to promulgate ideas; so were Tom Paine's *The Rights of Man*, Karl Marx's *Das Kapital*, and St. Paul's *Epistle to the Romans*. Since classical times, books, manuscripts, pictures, musical works and poems were commissioned for personal aggrandizement; many of the world's great buildings, from the pyramids to the Bibliothèque de France, were created because of an individual's wish to be remembered. Photographs, diaries, poems, and letters may be created simply for the private pleasure of the creator, yet subsequently be important library items. Few activities generate so many fine materials as religion; conversely, few activities create as much of dubious merit. The Christian tradition of fine writing, art, music, and architecture is repeated by religions around the world. The web has large amounts of material that are advertising or promotional, some of which will eventually become part of digital libraries.

The act of creation can be incidental to another activity. A judge who delivers a legal opinion is creating material that will become part of digital libraries. So is a museum curator preparing an exhibition catalog, or a drug researcher filing a patent claim. Materials are created by government agencies for public use. They range from navigational charts and weather forecasts, to official statistics, treaties, and trade agreements. Many of the materials in libraries were created to provide a record of some events or decisions. They include law reports, parish records, government records, official histories, and wartime photographs. Preserving an official record is an important functions of archives.

People who convert material to digital formats from other media can also be considered creators; conversion activities range from an individual who transcribes a favorite story and mounts it on the web, to projects that convert millions of items. The actual act of creation can even be carried out by a machine, such as images captured by a satellite circling the earth.

A second misconception is that creators and the organizations they work for have the same motives. Some works are created by teams, others by individuals; a feature film must be a team effort, but nobody would want a poem written by a committee. Hence, while some creators are individuals, such as free-lance writers, photographers, or composers, others belong to an organization and the materials that they create are part of the organization's activities. When somebody is employed by an organization, the employer often directs the act of creation and owns the results. This is called a "work for hire". In this case, the motivations of the individual and the organization may be different. A corporation that makes a feature film could be motivated by profit, but the director might see an opportunity to advocate a political opinion, while the leading actress has artistic goals.

Creators whose immediate motive is not financial usually benefit from the widest possible exposure of their work. This creates a tension with their publishers, whose business model is usually to allow access only after payment. Academic journal are an important category of materials where the author's interests can be in direct conflict with those of the publisher. Journal articles combine a record of the author's research with an opportunity to enhance the writer's reputation. Both objectives benefit from broad dissemination. The tension between creators who want wide distribution and the publishers' need for revenue is one of the themes of Chapter 6.

Users

Library users are as varied as creators in their interests and levels of expertise. Urban public libraries serve a particularly diverse group of users. For some people, a library is a source of recreational reading. For others, it acts as an employment center, providing information about job openings, and bus timetables for commuters. The library might provide Internet connections that people use as a source of medical or legal information. It might have audio-tapes of children's stories, and reference materials for local historians, which are used by casual visitors and by experts.

Individuals are different and a specific individual may have different needs at different times. Even when two users have similar needs, their use of the library might be different. One person uses catalogs and indexes extensively, while another relies more on links and citations. Designers of digital libraries must resist the temptation to assume a uniform, specific pattern of use and create a system specifically for that pattern.

Among the diversity of users, some broad categories can be distinguished, most of which apply to both digital libraries and to conventional libraries. One category is that people use libraries for recreation; in digital libraries this sometimes takes the form of unstructured browsing, colloquially known as "surfing". Another common use of a library is to find an introductory description of some subject: an engineer begins the study of a technical area by reading a survey article; before traveling, a tourist looks for information about the countries to be visited. Sometimes a user wants to know a simple fact. What is the wording of the First Amendment? What is the melting point of lead? Who won yesterday's football game? Some of these facts are provided by reference materials, such as maps, encyclopedias and dictionaries, others lie buried deep within the collections. Occasionally, a user wants comprehensive knowledge of a topic: a medical researcher wishes to know every published paper that has information about the effects of a certain drug; a lawyer wishes to know every precedent that might apply to a current case.

In many of these situations, the user does not need specific sources of information. There will be several library objects that would be satisfactory. For example, in answering a geographic question, there will be many maps and atlases that have the relevant information. Only for comprehensive study of a topic is there less flexibility. These distinctions are important in considering the economics of information (Chapter 6), since alternative sources of information lead to price competition, and in studying information retrieval (Chapter 10), where comprehensive searching has long been given special importance.

The information professions and change

Librarians and change

As digital information augments and sometimes replaces conventional methods, the information professions are changing. Librarians and publishers, in particular, have different traditions, and it is not surprising that their reactions to change differ. To examine how change affects librarians, it is useful to examine four aspects separately: library directors, mid-career librarians, the education of young librarians, and the increasing importance in libraries of specialists from other fields, notably computing.

Library directors are under pressure. To be director of a major library used to be a job for life. It provided pleasant work, prestige, and a good salary. The prestige and salary remain, but the work has changed dramatically. Digital libraries offer long-term potential but short-term headaches. Libraries are being squeezed by rising prices across the board. Conservative users demand that none of their conventional services be diminished, while other users want every digital service immediately. Many directors do not receive the support that they deserve from the people to whom they report, whose understanding of the changes in libraries is often weak. Every year a number of prominent directors decide not to spend their working life being buffeted by administrative confusion and resign to find areas where they have more control over their own destiny.

Mid-career librarians find that digital libraries are both an opportunity and a challenge. There is a serious shortage of senior librarians who are comfortable with modern technology. This means that energetic and creative individuals have opportunities. Conversely, people who are not at ease with technology can find that they get left behind. Adapting to technical change is more than an issue of retraining. Training is important, but it fails if it merely replaces one set of static skills with another. Modern libraries need people who are aware of the changes that are happening around them, inquisitive and open to discover new ideas. Panel 5.4 describes one attempt to educate mid-career librarians, the Ticer summer school run by Tilburg University in the Netherlands.

Panel 5.4
The Ticer Summer School

In 1996, Tilburg University in the Netherlands introduced a summer school to educate senior librarians about digital libraries. The program was an immediate success and has been fully subscribed every year. Students for the two week course come from around the world. The largest numbers come from northern Europe, but, in 1998, there were students from Malaysia, Japan, India, South Korea and many other countries. Most are senior staff in academic or special libraries.

Tilburg University has been a leader in digital library implementation for many years; the course reflects the combination of strategic planning and practical implementation that has marked the university's own efforts. Many of the lecturers at the summer school are members of the libraries or computing services at Tilburg. In addition, a range of visiting experts provide breadth and visibility for the program. The fat book of lecture notes that every student receives is a major resource.

Some other features have led to the success of the Ticer school. The costs have been kept reasonably low, yet a high standard of facilities is provided. A pleasant social program enhances the value of fifty people from around the world living and working together for two weeks. Ticer has close relationship with Elsevier Science, the large Dutch publisher. Elsevier staff from around the world attend as students and senior Elsevier personnel give lectures. Finally, in a country where the weather is unpredictable, Ticer always seems to provide warm summer weather where people can sit outside, relax and work together.

Ticer demonstrates how international the field of digital libraries has become and the privileged position of the English language. The Ticer program is so thoroughly English-language that the publicity materials do not even mention that English is the language of the summer school, yet few students come from the English speaking countries.

Ticer is considering other programs. In 1998, several students expressed frustration that, while they were learning a great deal from the summer school, the directors of their libraries needed to hear the same story. Perhaps Ticer could offer a shortened program for executives.

The education of young librarians revolves around library schools. Librarianship is a profession. In the United States, a masters degree from a library school is a requirement for many library jobs. For years, the curriculum of library schools was rather pedestrian, centered around the basic skills needed by mid-level librarians. In many universities, the library schools was one of the weaker schools academically. Over the past few years, universities have realized that digital libraries provide opportunities for a new type of library school, with a new curriculum and a vigorous program of research. Some library schools are being rebuilt to focus on the modern world. Panel 5.5 describes one of them, at the University of California at Berkeley.

Table 5.5
The School of Information Management and Systems at the University of California at Berkeley

Early in the 1990s, several leading universities questioned the quality of their library schools. Other professional schools, such as law and business, were attracting the nation's brightest students and faculty, contributing outstanding research, and making a hefty profit. Library schools were drifting. The curriculum was more suitable for a trade school than a university. The faculty were underpaid and unproductive. Worst of all, the educational programs were not creating the leaders that the new digital libraries and electronic publishing would require.

Faced with this situation, Columbia University simply closed down its library school. The programs were considered to be less valuable than the buildings that the school occupied. The University of Chicago also closed its library school. The University of California at Berkeley and the University of Michigan went the other way and completely refurbished their schools. Perhaps the most dramatic change was at Berkeley.

The decision to create a new school

Many universities claim that academic decisions are made by the faculty. In most universities, this is pretense, but at Berkeley the academic community has great power. Berkeley's first instinct was to follow Columbia and to close down its library school, but enough people were opposed that the university changed its mind and set up a high-level planning group. The report of this group, Proposal for a School of Information Management and Systems, was released in December 1993. In classic bureaucratic doublespeak, the fundamental recommendation was to "disestablish" the existing school and "reconstitute" a new school from its ashes.

In creating a new academic program, good universities look at two factors. The first is the academic content. Is this an area with deep intellectual content that will attract first-rate faculty, whose scholarship will be a credit to the university? The second is education. Will the program attract able students who will go out and become leaders? The report answered these questions emphatically, but only if certain criteria were met.

As a starting point the report explicitly rejected the traditional library school curriculum and the master of library science degrees. With remarkable frankness, the report stated, "The degree to be awarded by this program ... is not designed to meet American Library Association requirements; rather, it will serve as a model for the development of accreditation criteria for the emerging discipline upon which the School is focused."

An inter-disciplinary program

The report was accepted and the school set out to recruit faculty and students. From the start the emphasis was on inter-disciplinary studies. The planning report is full of references to joint programs with other schools and departments. The program announcement for the new masters program mentioned, "aspects of computer science, cognitive science, psychology and sociology, economics, business, law, library/information studies, and communications." Students were required to have significant computer skills, but no other constraints were placed on their previous academic background.

The appointment of faculty was equally broad. The first dean, Hal Varian, is an economist. Other faculty members include a prominent scholar in copyright law and a former chair of the Berkeley computer scientist department. Many of the appointments were joint appointments, so that the faculty teach and carry out research across traditional departmental lines.

Success

It is to early to claim success for this new school or the similar effort at the University of Michigan, but the first signs are good. The school has met the basic requirements of high-quality faculty and students. The research programs have grown fast, with large external grants for research on interesting topics. Hopefully, a few years from now the first graduates will be emerging as the leaders of the next generation of information professionals.

Digital libraries require experts who do not consider themselves professional librarians, such as computer specialists and lawyers. Fitting such people into the highly structured world of libraries is a problem. Libraries need to bring in new skills, yet libraries are hesitant to recruit talent from outside their ranks, and needlessly restrict their choices by requiring a library degree from candidates for professional positions. Few of the young people who are creating digital libraries see library school on their career path. Compared with other professions, librarianship is notable for how badly the mid-level professionals are paid. Top-class work requires top-class people, and in digital libraries the best people have a high market value. It is troubling to pay a programmer more than a department head, but it is even more troubling to see a good library deteriorate because of a poor computing staff. Part of the success of the Mercury project at Carnegie Mellon was that the technical staff were administratively members of the university's computing center. Their offices were in the library and their allegiance was to the digital library, but they were supervised by technical managers, worked their usual irregular hours, had excellent equipment, and were paid the same salary as other computing professionals. Few libraries are so flexible.

Publishers and change

The changes that are happening in publishing are as dramatic as those in libraries. Since the fifteenth century, when printing was introduced in Europe, publishing has been a slow moving field. The publisher's task has been to create and distribute printed documents. Today, the publishing industry still draws most of its revenue from selling physical products - books, journals, magazines, and more recently videos and compact disks - but many publishers see their future growth in electronic publications.

Publishing is a mixture of cottage industry and big business. Large publishers, such as Time Warner, Reed Elsevier, and the Thomson group, rank as some of the biggest and most profitable corporations in the world, yet most of the 50,000 publishers in the United States publish fewer than ten items per year. Academic publishing has the strange feature that some publishers are commercial organizations, in business to make profits for their shareholders, while others are not-for-profit societies and university presses whose primary function is to support scholarship.

Success in publishing depends upon the editors who select the materials to be published, work with the creators, and oversee each work as it goes through production. Publishing may be a business, but many of the people who come to work in publishing are not looking for a business career. They enter the field because they are interested in the content of the materials that they publish.

Publishers use sub-contractors extensively. A few, such as West Publishing, the big legal publisher, run their own printing presses, but most printing is by specialist printers. Services that support the editor, such as copy-editing, are often carried out by free-lance workers. Books are sold through booksellers, and journals through subscription agents. This use of contractors gives publishers flexibility that they would not have if everything were carried out in-house. When Elsevier moved to producing journals in SGML mark-up, they could withdraw contracts from those printers who were not prepared to change to SGML. Recently, the publishing industry has had a wave of corporate takeovers. This has created a few huge companies with the wealth to support large-scale computing projects. The future will tell whether size is necessarily a benefit in electronic publishing. Web technology means that small companies can move rapidly into new fields, and small companies sometimes have an advantage in developing close relationships between editors and authors. Some observers considered that the decision to be sold by West Publishing, which had been privately held, was driven by fear that electronic publishing might weaken its dominance of the legal market.

Almost every university has a university press, but university publishing is less central to the university than its library. A few, such as Oxford University Press and Chicago University Press, have much in common with commercial publishers of academic materials. They publish large numbers of general interest, reference, and scholarly books. These presses operate on a sound financial footing and give no priority to authors from their own universities. Other university presses have a different role. Most scholarly monographs in the humanities have such narrow appeal that only a few hundred copies are sold. Such books would never be considered by commercial publishers. They are published by university presses, who operate on shoe-string budgets, usually with subsidies from their host universities. As universities have been tight for money during the past few years, these subsidies have been disappearing.

Computer professionals, webmasters, and change

Computing professional are as much part of digital libraries as are librarians and publishers. Success requires cooperation between these professions, but the cultural differences are great. The members of the Internet community are a wonderful resource, but they have an unorthodox style of working. Many have grown up immersed in computing. As in every discipline, some people are so knowledgeable about the details that they see nothing else and have difficulty in explaining them to others. A few people are deliberately obscure. Some technical people appear to define merit as knowledge of the arcane. Fortunately, deliberate obscurity is rare. Most people would like others to know what they do, but technical people often have genuine difficulty in describing technology to non-specialists.

The Internet community has its foundation in the generation of computer scientists who grew up on the Unix operating system. Much of the success of Unix came from a tradition of openness. The Unix, Internet, and web communities share their efforts with each other. They have discovered that this makes everybody more productive. An appreciation of this spirit of openness is fundamental to understanding how computer scientists view the development of digital libraries. This attitude can become a utopian dream, and a few idealists advocate a completely uncontrolled information society, but this is an extreme viewpoint. Most members of the Internet community have a healthy liking for money; entrepreneurship is part of the tradition, with new companies being formed continuously.

Even in the rapidly changing world of computing, the emergence of **webmaster** as a new profession has had few parallels. At the beginning of 1994, the web was hardly known, yet, in summer 1995, over one thousand people attended the first meeting of the Federal Webmasters in Washington, DC. This is an informal group of people whose primary job is to maintain web sites for the U.S. government. The emergence of the name "webmaster" was equally rapid. It appeared overnight and immediately became so firmly entrenched that attempts to find a word that applies equally to men and women have failed. Webmaster must be understood as a word that applies to both sexes, like librarian or publisher.

A webmaster is a publisher, a librarian, a computer administrator, and a designer. The job requires a range of expertise that include the traditional publishing skills of selection of material and editing, with the addition of user interface design and the operation of a high-performance computer system. A web site is the face that an organization presents to the world. The quality of its graphics and the way that it presents the material on the site are as important as any public relations material that the organization issues. At CNRI, the webmaster refers to herself, half-jokingly, as "the Art Department". She is a professional librarian who is highly skilled technically, but she spends much of her time carrying out work that in other contexts would be called graphic design. Her designs are the first impression that a user sees on visiting CNRI's web site.

In some organizations, the webmaster selects and even creates, the materials that appear on a web site. More commonly, the materials are generated by individuals or groups within the organization. The webmaster's task is to edit and format individual items, and to organize them within the web site. Thus the overall shape of the site is created by the webmaster, but not the individual items. For example, CNRI manages the web site of the Internet Engineering Task Force (IETF). The content is firmly controlled by the IETF secretariat, while the webmaster contributed the design of the home page the links with other items on the web site, so that a user who arrives at any part of the site has a natural way to find any of the material.

Webmasters vary in their range of computing skills. The material on a web site can range from simple HTML pages to complex. Some user interface methods, such as the Java programming language, require skilled programmers. Some web sites are very large. They replicate information on many powerful computers, which are distributed across the world. A really popular site, such as Cable Network News, has over one hundred million hits every day. Designing and monitoring these systems and their network connections is a skilled computing job. If the web site handles commercial transactions, the webmaster needs expertise in network security.

Many of the organizations that contribute to digital libraries have computing departments. The webmasters can rely on them for the occasional technical task, such as setting up a web server or a searchable index to the site. In other organizations, the webmaster must be a jack of all trades. Many web sites serve organizations that have no computing staff. A plethora of companies provide web services for such organizations. They design web pages, run server computers, and perform administrative tasks, such as registration of domain names.

New forms of organization

Consortia

Managing large online collections is expensive and labor-intensive. Libraries can save effort by working together in consortia to acquire and mount shared collections. This saves effort for the libraries and also for publishers since they have fewer customers to negotiate with and support. In the United States, consortia have been organized by states, such as Ohio, or by academic groupings. In Europe, where most universities are state run, there are well-established national consortia that provide digital library services for the entire academic community.

Panel 5.6 describes MELVYL which is a good example of a collaborative efforts. MELVYL was established by the University of California, before the days of the web, to provide services that concentrated on sharing catalog and indexing records. When the web emerged and publishers began to supply the full text of journals, the organization and technical structures were in place to acquire these materials and deliver them to a large community of users. This panel also describes the California Digital Library, a new venture built on the foundations of MELVYL.

Panel 5.6
MELVYL

The nine campuses of the University of California often act as though they were nine independent universities. Campuses, such as the University of California, Berkeley and the University of California, Los Angeles (UCLA), rank as major universities in their own right, but organizationally they are parts of a single huge university. They have shared a digital library system, MELVYL, for many years. For much of its life, MELVYL was under the vigorous leadership of Clifford Lynch.

At the center of MELVYL is a computer-based catalog of holdings of all libraries in the nine campuses, the California State Library in Sacramento, and the California Academy of Sciences. This catalog has records of more than 8.5 million monographic titles, representing 13 million items. In addition to book holdings, it includes materials such as maps, films, musical scores, data files, and sound recordings. The periodicals database has about 800,000 unique titles of newspapers, journals, proceedings, etc., including the holdings of other major California libraries. MELVYL also provides access to numerous article abstracting and indexing files, including the entire Medline and Inspec databases. In 1995, MELVYL added bit-mapped images of the publications of the Institute of Electrical and Electronics Engineers (IEEE). The images were linked through the Inspec database. Users who accessed the Inspec database could see the message "Image available" for records with linked IEEE bit-mapped images. The user could then request the server to open a window on the user's workstation to display the bit-mapped images. Use of the books and periodicals files is open to everybody. Use of the article databases is limited to members of the University of California.

MELVYL has consistently been an early adopter of new digital library technology. Much of the development of Z39.50 has been associated with the service. The MELVYL team was also responsible for creating the communications network between the University of California campuses.

The California Digital Library

The success of MELVYL helped the creation in 1998 of an even bolder project, the California Digital Library. This is the University of California's approach to the organizational challenges posed by digital libraries.

Each of the nine campuses has its own library and each recognizes the need to provide digital library services. After a two year planning process, the university decided in 1997 to create a tenth library, the California Digital Library, which will provide digital library services to the entire university. Organizationally this digital library is intended to be equal to each of the others. It director, Richard Lucier, ranks equally with the nine other directors; its initial budget was about $10 million and is expected to rise sharply.

The university could easily have justified central digital library services through arguments of economies of scale, a focus for licensing negotiations, and leveraged purchasing power. For these reasons, the campuses have historically shared some library services, notably MELVYL, which is incorporated in the new digital library. The California Digital Library is much more ambitious, however. The university anticipates that digital libraries will transform scholarly communication. The digital library is explicitly charged with being an active part of this process. It is expected to have a vigorous research program and to work with organizations everywhere to transform the role of libraries in supporting teaching and research. These are bold objectives, but the organization of the library almost guarantees success. At the very least, the University of California will receive excellent digital library services; at the best the California Digital Library will be a catalyst that changes academic life for everybody.

Secondary information providers and aggregators

The term **secondary information** covers a wide range of services that help users find information, including catalogs, indexes, and abstracting services. While many publishers are household names, the suppliers of secondary information are less well known. Some, such as Chemical Abstracts, grew out of professional societies. Others have always been commercial operations, such as ISI, which publishes *Current Contents* and *Science Citation Index*, and Bowker, which publishes *Books in Print*. OCLC has a special niche as a membership organization that grew up around shared cataloguing.

These organizations are simultaneously vulnerable to change and well-placed to expand their services into digital libraries. Their advantages are years of computing experience, good marketing, and large collections of valuable data. Many have strong financial reserves or are subsidiaries of conglomerates with the money to support new ventures. Almost every organization sees its future as integration between secondary information and the primary information. Therefore there are numerous joint projects between publishers and secondary information services.

Aggregators are services that assemble publications from many publishers and provide them as a single package to users, usually through sales to libraries. Some had their roots in the early online information systems, such as Dialog and BRS. These services licensed indexes and other databases, mounted them on central computers with a specialized search interface and sold access. Nowadays, the technical advantages that aggregators bring is comparatively small, but they provide another advantage. A large aggregator might negotiate licenses with five thousand or more publishers. The customer has a single license with the aggregator.

Universities and their libraries

Changes in university libraries

Like most organizations, universities have difficulty in handling change. Universities are large, conservative organizations. The senior members, the tenured faculty, are appointed for life to work in narrowly defined subject areas. Universities are plagued by caste distinctions that inhibit teamwork. The cultural divide between the humanities and the sciences is well-known, but an equally deep divide lies between scholars, librarians, and computing professionals. Faculty treat non-faculty colleagues with disdain, librarians have a jargon of their own, and computing professionals consider technical knowledge the only measure of worth.

The academic department dominated by tenured faculty is a powerful force toward retaining the status quo. To close a department, however moribund, is seen as an act of academic vandalism. When a corporation closes a factory, its stock price usually goes up. When Columbia University closed its library school, nobody's stock went up. There is little obvious incentive and much vocal disincentive to change. Until recently, Oxford University still had more professors of divinity than of mathematics.

Yet, universities are a continual source of innovation. Chapters 2 and 3 included numerous examples where universities developed and deployed new technology, long before the commercial sector. The flow of high-technology companies that fuels the American economy is driven by a small number of research universities, such as Stanford and Berkeley near San Francisco, Harvard and M.I.T. in Boston, and Carnegie Mellon in Pittsburgh.

Innovation in a large organization requires strategic reallocation of resources. New initiatives require new funding. Resources can be found only by making hard choices. The process by which resources are allocated at a research university appears arcane. Moving resources from one area to build up another is fraught with difficulties. The director of the Ashmolean Museum in Oxford once mused that the best strategy for the museum might be to sell part of its collection to provide funds to look after the rest; he doubted whether he would retain his position if he carried out such a strategy. Few deans would advocate cutting faculty numbers to provide resources that will make the remaining faculty more productive. Yet funds are reallocated. Year-by-year the portion of the budget that goes into computers, networks and support staff increases, one percent at a time.

In the long term, it is not clear whether such reallocations will bring more resources to existing library organizations, or whether universities will develop new information services outside their libraries. The signals are mixed. By a strange paradox, good information has never been more important than it is today, yet the university library is declining in importance relative to other information sources. The university library, with its collections of journals and monographs, is only one component in the exchange of academic information. Fifty years ago, faculty and students had few sources of information. Today they have dozens of methods of communication. New technology, from desk-top computing and the Internet, to air travel and video conferences, allows individual scholars to exchange large amounts of information. The technology has become so simple that scholars are able to create and distribute information with less help from professionals and outside the formal structure of libraries.

If libraries are to be the center for new information services they have to reallocate resources internally. Discussions of library budgets usually focus on the rising cost of materials, overcrowding in the buildings, and the cost of computing, but the biggest costs are the staff. Few universities make an honest effort to estimate the costs of their libraries, but a true accounting of a typical research library would show that about twenty five percent of the cost is in acquisitions and fifty percent in staff costs. The other costs are for building and equipment. If libraries are to respond to the opportunities brought by electronic information, while raising salaries to attract excellent staff, there is only one option. They will have to reorganize their staff. This is not simply a question of urging people to work harder or streamlining internal processes; it implies fundamental restructuring.

Buildings for digital libraries

An area of change that is difficult for all libraries, but particularly universities, is how to plan for library buildings. While digital libraries are the focus of research and development around the world, for many libraries the biggest problem is seen as the perennial lack of space. For example, in December 1993, the funding councils for higher education in the United Kingdom released a report on university libraries, known as the Follett Report. In terms of money, the biggest recommendation from the Follett Committee was the need for a major building program. This need was especially acute in Britain, because the numbers of students at universities had grown sharply and much of the space was required to provide study space on campus.

The expense of a new building to house rarely-used paper is hard to justify, but old habits are slow to change. Imposing library buildings are being built around the world, including new national libraries in London and Paris. Many libraries retain a card catalog in elegant oak cabinets to satisfy the demands of a few senior users, even though the catalog is online and no new cards have been filed for years.

To use a traditional library, the user almost invariably goes to the library. Some libraries provide services to deliver books or photocopies to offices of privileged users, but even these users must be near the library and known to the library staff. Users of a digital library have no incentive to visit any particular location. The librarians, webmasters, and other professionals who manage the collections have offices where they work, but there is no reason why they should ever see a user. The New York Public Library must be in New York, but the New York digital library could store its collections in Bermuda.

The dilemma is to know what will make a good library building in future years. The trials and tribulations of the new British Library building in London show the problems that happen without good planning. Library buildings typically last at least fifty years, but nobody can anticipate what an academic library will look like even a few years from now. Therefore the emphasis in new library buildings must be on flexibility. Since modern library buildings must anticipate communications needs that are only glimpsed at today, general purpose network wiring and generous electrical supplies must be led to all spaces. Yet the same structures must be suitable for traditional stacks.

Panel 5.7
The renovation of Harvard Law Library

Langdell Hall is the main library of Harvard Law School. As such, it is the working library of a large law school and one of the great collections of the history of law. During 1996/97, the library was fully renovated. The project illustrates the challenges of building for the long term during a period of rapid change.

At the time when the renovations were being planned, Harvard made two bold proposals. The first was that the library should provide no public computers. The second was that every working space should support laptop computers. It was assumed that people who visit the library will bring their own computers. In the new library, every place where users might work has a power outlet and a network connection. Users can use any of 540 Ethernet ports at tables, carrels, lounges, or study rooms throughout the building. In total, 1,158 network connections are installed within the building; almost all were activated immediately, the rest kept in reserve. In practice, the idea of having no public computers was relaxed slightly. There are about one hundred public computers, including one computer classroom and two small training labs.

Even for Harvard Law School, the cost of this project, $35 million, is a great deal of money. The school has effectively gambled on its assumptions that all users will have laptop computers and that the network installation has enough flexibility to adapt to changes with time. However, the biggest gamble was probably never stated explicitly. The school actually increased the amount of space devoted to library users. The assumption is that people will continue to come to the library to study. Law school faculty are known to prefer working in their offices rather than walk to the library. Many years ago, the librarian, Harry S. Martin III, stated, "Our aim is to keep the faculty out of the library." This witticism describes a commitment to provide faculty with service in their offices, including both online information and an excellent courier service. However, the attention given to reading spaces in Langdell implies a belief that, for many years, legal scholars and law school students will come to the library to do serious work.

Not all of the money went to technology. Langdell is an elegant old building, with historic books and valuable art. Much of the budget went into elevators, heating, air-conditioning and plumbing. The dignity of the library was not forgotten. Old tables were restored, chandeliers replaced an elderly and ineffective dropped ceiling, Latin inscriptions were re-located, and bas relief symbols of the law highlighted. This vision of a great law library combines the traditional view of a library with access to modern technology. Hopefully, Harvard has judged correctly and will be proud of its new law library for many years.

Buildings are one of many aspects of libraries that are long-term investments. Collections, catalogs, and indexes also represent sustained efforts over many years. Digital libraries are forcing rapid changes which provide both opportunities and challenges. This is not easy for individuals or for organizations, but the manner in which they react to digital libraries will determine which of them flourish in future years.

Chapter 6
Economic and legal issues

Introduction

Digital libraries poses challenges in the fields of economics, public policy, and the law. Publishing and libraries exist in a social and economic context where the operating rules and conventions have evolved over many years. As electronic publication and digital libraries expand, the business practices and legal framework are changing quickly.

Because digital libraries are based on technology, some people hope to solve all challenges with technical solutions. Other people believe that everything can be achieved by passing new laws. Both approaches are flawed. Technology can contribute to the solutions, but it can not resolve economic or social issues. Changes in laws may be helpful, but bad laws are worse than no laws. Laws are effective only when they codify a framework that people understand and are willing to accept. In the same manner, business models for electronic information will fail unless they appeal to the interests of all interested parties. The underlying challenge is to establish social customs for using information that are widely understood and generally followed. If they allow reasonable people to carry out their work, then reasonable people will observe them. The economic and legal frameworks will follow.

Economic forces

Libraries and publishing are big businesses. Huge industries create information or entertainment for financial gain. They include feature films, newspapers, commercial photographs, novels and text books, computer software, and musical recordings. Some estimates suggest that these industries comprise five percent of the economy of the United States. In 1997, Harvard University Libraries had a budget well over fifty million dollars and employed a thousand people. The Library of Congress employs about 4,500 people. Companies, such as Time Warner are major forces on the world's stock markets. In 1996, the Thomson Corporation paid more than three billion dollars for West Publishing, the legal publisher. The stock market valuation of Yahoo, the Internet search firm, is even higher though its profits are tiny.

The publishers' concerns about the business and legal framework for online information derive from the two usual economic forces of greed and fear. The greed comes from a belief that publishers can make huge sums of money from electronic information, if only they knew how. The fear is that the changing economic picture will destroy traditional sources of revenue, organizations will wither away, and people will lose their jobs. To computing professionals, this fear is completely reasonable. Most of the companies that succeeded in one generation of computing have failed in the next. Mainframe companies such as Univac, CDC, Burroughs, and Honeywell were supplanted by minicomputers. Minicomputer companies, such as Prime, Wang, and Data General did not survive the transition to personal computers. Early personal computer companies have died, as have most of the software pioneers. Even IBM has lost its dominance. Will the same pattern happen with electronic information? The publishers and information services who dominated traditional markets are not necessarily those who will lead in the new ones.

Whichever organizations thrive, large or small, commercial or not-for-profit, they have to cover their costs. Every stage in the creation, distribution, and use of digital libraries is expensive. Authors, photographers, composers, designers, and editors need incentives for their efforts in creating information. In many circumstances, but not all, the incentives are financial. Publishers, librarians, archivists, booksellers, subscription agents, and computing specialists - all these people require payment. As yet, there is no consensus on how best to pay for information on the Internet. Almost every conceivable method is being tried. For discussion of the economics of digital libraries, the various approaches can be divided into open access, in which the funds come from the creator or producer of the information, and models in which the user or the user's library pays for access to the collections.

Chapter 5 noted that the various people who are involved in digital libraries and electronic publishing have many motives. In particular, it noted that creators and publishers often have different financial objectives. When the creators are principally motivated by financial reward, their interests are quite well aligned with those of the publisher. The aim is to generate revenue. The only question is how much should go to the creator and how much to the publisher. If, however, the creator's objective is non-financial while the publisher is concentrating on covering costs or generating profits for shareholders, then they may have conflicting objectives.

Open access digital libraries collections

A remarkable aspect of the web is the huge amounts of excellent material that are openly available, with no requirement for payment by the user. First predictions were that open access would restrict the web to inferior quality material, but a remarkable amount of high-quality information is to be found on the networks, paid for and maintained by the producers. In retrospect this is not surprising. Many creators and suppliers of information are keen that their materials should be seen and are prepared to meet the costs from their own budgets. Creators who invest their own resources to make their materials openly available include researchers seeking for professional recognition, government agencies informing the public, all types of marketing, hobbyists, and other recreational groups.

Whenever creators are principally motivated by the wish to have their work widely used, they will prefer open access, but the money to maintain these collections must come from somewhere. Grants are one important source of funds. The Perseus project has relied heavily on grants from foundations. At the Library of Congress, a combination of grants and internal funds pay for American Memory and the National Digital Library Program. Grants are usually short-term, but they can be renewed. The Los Alamos E-Print Archives receive an annual grant from the National Science Foundation, and Netlib has received funding from DARPA since its inception. In essence, grant funding for these digital libraries has become institutionalized.

The web search firms, such as Yahoo, Infoseek, and Lycos, provide open access paid for by advertising. They have rediscovered the financial model that is used by broadcast television in the United States. The television networks pay to create television programs and broadcast them openly. The viewer, sitting at home in front of a television, does not pay directly. The revenues come from advertisers. A crucial point in the evolution of digital libraries occurred when these web search programs first became available. Some of the services attempted to charge a monthly fee, but the creator of Lycos was determined to offer open access to everybody. He set out to find alternative sources of revenue. Since Lycos was available with no charge, competing services could not charge access fees for comparable products. The web search services remain open to everybody. After a rocky few years, the companies are now profitable, using advertising and revenue from licensing to support open access to their services.

Research teams have a particular interest in having their work widely read. CNRI is a typical research organization that uses its own resources to maintain a high-quality web site. Most of the research is first reported on the web. In addition, the corporation has a grant to publish *D-Lib Magazine*. The Internet Draft series is also maintained at CNRI; it is paid for by a different method. The conference fees for meetings of the Internet Engineering Task Force pay the salaries of the people who manage the publications. All this information is open access.

Government departments are another important source of open access collections. They provide much information that is short-lived, such as the hurricane tracking service provided by the United States, but many of their collections are of long-term value. For example, the Trade Compliance Center of the U.S. Department of Commerce maintains a library collection of international treaties. The U.S. Department of State provides a grant to librarians at the University of Illinois at Chicago, who operates the web site "www.state.gov" for the department.

Private individuals maintain many open access library collections. Some excellent examples include collections devoted to sports and hobbies, fan clubs, and privately published poetry, with an increasingly large number of online novels. Payment by the producer is not new. In book publishing it is given the disparaging name of "vanity press", but on the Internet it is often the source of fine collections.

Payment for access to digital library collections

When the publisher of a digital library collection wishes to collect revenue from the user, access to the collections is almost always restricted. Users have access to the materials only after payment has been received. The techniques used to manage such access are a theme of the next chapter. This section looks at the business practices.

Book and journal publishing have traditionally relied on payment by the user. When a copy of a book is sold to a library or to an individual, the proceeds are divided amongst the bookseller, the publisher, the author, and the other contributors. Feature films follow the same model. The costs of creating and distributing the film are recovered from users through sales at cinemas and video rentals. Extending this model to online information, leads to fees based on usage. Most users of the legal information services, Lexis and Westlaw, pay a rate that is based on the number of hours that they use the services. Alternative methods of charging for online information, which are sometimes tried, set a fee that is based on peak usage, such as the number of computers that could connect to the information, or the maximum number of simultaneous users. With the Internet and web protocols, these charging methods that are based on computer or network usage are all rather artificial.

An alternative is to charge for the content transmitted to the user. Several publishers provide access to the text of an article if the user pays a fee, perhaps $10. This could be charged to a credit card, but credit card transactions are awkward for the user and expensive to process. Therefore, there is research interest in automatic payment systems for information delivered over the networks. The aim of these systems is to build an Internet billing service with secure, low-cost transactions. The hope is that this would allow small organizations to set up network service, without the complexity of developing private billing services. If such systems became established, they would support high volumes of very small transactions. Whereas physical items, such as books, come in fixed units, a market might be established for small units of electronic information.

At present, the concept of automatic payment systems is mainly conjecture. The dominant form of payment for digital library materials is by subscription, consisting of scheduled payments for access to a set of materials. Unlimited use is allowed so long as reasonable conditions are observed. The Wall Street Journal has developed a good business selling individual subscriptions to its online editions. Many large scientific publishers now offer electronic journal subscriptions to libraries; society publishers, such as the Association for Computing Machinery (ACM), sell subscriptions both to libraries and to individual members. Some of the digital libraries described in earlier chapters began with grants and have steadily moved towards self-sufficiency through subscriptions. JSTOR and the Inter-university Consortium for Political and Social Research (ICPSR) have followed this path.

Television again provides an interesting parallel. In the United States, two alternatives to advertising revenue have been tried by the television industry. The first, pay-by-view, requires viewers to make a separate payment for each program that they watch. This has developed a niche market, but not become widespread. The second, which is used by the cable companies, is to ask viewers to pay a monthly subscription for a package of programs. This second business model has been extremely successful.

It appears that the users of digital libraries, like television viewers, welcome regular, predictable charges. Payment by subscription has advantages for both the publisher and the user. Libraries and other subscribers know the costs in advance and are able to budget accurately. Publishers know the revenue to expect. For the publisher, subscriptions overcome one of the problems of use-based pricing, that popular items make great profits while specialist items with limited demand lose money.

Libraries are also attracted by the subscription form of payment because it encourages wide-spread use. Libraries wish to see their collections used as heavily as possible, with the minimum of obstacles. Digital libraries have removed many of the barriers inherent in a traditional library. It would be sad to introduce new barriers, through use-based pricing.

Economic considerations

An economic factor that differentiates electronic information from traditional forms of publishing is that the costs are essentially the same whether or not anybody uses the materials. Many of the tasks in creating a publication - including soliciting, reviewing, editing, designing and formatting material - are much the same whether a print product is being created or an electronic one. With digital information, however, once the first copy has been mounted online, the distribution costs are tiny. In economic terms, the cost is almost entirely fixed cost. The marginal cost is near zero. As a result, once sales of a digital product reach the level that covers the costs of creation, all additional sales are pure profit. Unless this level is reached, the product is condemned to make a loss.

With physical materials, the standard method of payment is to charge for each physical copy of the item, which may be a book, a music CD, a photograph, or similar artifact. The production of each of these copies costs money and the customer feels that in some way this is reflected in the price. With digital information, the customer receives no artifact, which is one reason why use-based pricing is unappealing. Subscriptions match a fixed cost for using the materials against the fixed cost of creating them.

A further way in which the economics of electronic publications differ from traditional publishing is that the pricing needs to recognize the costs of change. Electronic publishing and digital libraries will be cheaper than print in the long term, but today they are expensive. Organizations who wish to enter this field have a dilemma. They have to continue with their traditional operations, while investing in the new technology at a time when the new systems are expensive and in many cases the volume of use is still comparatively low.

Many electronic publications are versions of materials that is also published in print. When publishers put a price on the electronic publication they want to know if sales of electronic information will decrease sales of corresponding print products. If a dictionary is online, will sales in book stores decline (or go up)? If a journal is online, will individual subscriptions change? After a decade of experience with materials online, firm evidence of such substitution is beginning to emerge, but is still hard to find for any specific publication. At the macroeconomic level, the impact is clear. Electronic information is becoming a significant item in library acquisition budgets. Electronic and print products often compete directly for a single pool of money. If one publisher generates extra revenue from electronic information, it is probably at the expense of another's print products. This overall transfer of money from print to electronic products is one of the driving force that is pressing every publisher towards electronic publication.

Some dramatic examples come from secondary information services. During the past decade, products such as *Chemical Abstracts* and *Current Contents* have changed their content little, but whereas they were predominantly printed products, now the various digital versions predominate. The companies have cleverly offered users and their libraries many choices: CD-ROMs, magnetic tape distributions, online services, and the original printed volumes. Any publisher that stays out of the electronic market must anticipate steadily declining revenues as the electronic market diverts funds from the purchase of paper products.

Fortunately, the librarian and the publisher do not have to pay for one of the most expensive part of digital libraries. Electronic libraries are being built around general purpose networks of personal computers that are being installed by organizations everywhere. Long distance communications use the international Internet. These investments, which are the foundation that make digital libraries possible, are being made from other budgets.

Alternative sources of revenue

Use-based pricing and subscriptions are not the only ways to recover revenue from users. Cable television redistributes programs created by the network television companies, initially against strenuous opposition from the broadcasters. Political lobbying, notably by Ted Turner, led to the present situation whereby the cable companies can redistribute any program but must pay a percentage of their revenues as royalty. When a radio program broadcasts recorded music, the process by which revenue accrues to the composers, performers, recording studios, and other contributors is based on a complex system of sampling. The British football league gains revenue from gambling by exercising copyright on its fixture list.

There is nothing inevitable about these various approaches. They are pragmatic resolutions of the complex question of how the people who create and distribute various kinds of information can be compensated by the people who benefit from them.

A case study: scientific journals in electronic format

Scientific journals in electronic format provide an interesting case study, since they are one of the pioneering areas where electronic publications are recovering revenue from libraries and users. They highlight the tension between commercial publishers, whose objective lies in maximizing profits, and authors whose interests are in widespread distribution of their work. Academic publishers and university libraries are natural partners, but the movement to electronic publication has aggravated some long standing friction. As described in Panel 6.1, the libraries can make a strong case that the publishers charge too much for their journals, though, in many ways, the universities have brought the problem on themselves. Many researchers and librarians hope that digital libraries will lead to new ways of scientific publication, which will provide wider access to research at lower costs to libraries. Meanwhile, the commercial publishers are under pressure from their shareholders to make higher profits ever year.

Panel 6.1
The economics of scientific journals

The highly profitable business of publishing scientific journals has come under increasing scrutiny in recent years. In the United States, the federal government uses money received from taxes to fund research in universities, government laboratories, medical centers, and other research organizations. The conventional way to report the results of such research is for the researcher to write a paper and submit it to the publisher of a scientific journal.

The first stage in the publication process is that an editor, who may be a volunteer or work for the publisher, sends out the paper for review by other scientists. The reviewers are unpaid volunteers, who read the paper critically for quality, checking for mistakes, and recommending whether the paper is of high enough quality to publish. This process is known as peer review. The editor selects the papers to publish and gives the author an opportunity to make changes based on the reviewers' comments.

Before publication, most publishers place some requirements on the authors. Usually they demand that copyright in the paper is transferred from the author to the publisher. In addition, many publishers prohibit the author from releasing the results of the research publicly before the journal article is published. As a result, without making any payment, the publisher acquires a monopoly position in the work.

Although a few journals, typically those published by societies, have individual subscribers, the main market for published journals is academic libraries. They are required to pay whatever price the publisher chooses to place on the journal. Many of these prices are high; more than a thousand dollars per year is common. Moreover, the annual increase in subscriptions over the past decade has averaged ten to fifteen percent. Yet the libraries have felt compelled to subscribe to the journals because their faculty and students need access to the research that is reported in them.

The economic system surrounding scientific journals is strange. The taxpayer pays the researcher, most of the reviewers, and many of the costs of the libraries, but the copyright is given away to the publisher. However, the universities must take much of the blame for allowing this situation. Since their faculty carry out the research and their libraries buy the journals, simple policy changes could save millions of dollar, while still providing reasonable compensation for publishers. Recently universities have begun to work together to remedy the situation.

The underlying reason for the reluctance of universities to act comes from a peculiarity of the academic system. The function of scientific articles is not only to communicate research; they also enhance the professional standing of the authors. Academic reputations are made by publication of journal articles and academic monographs. Professional recognition, which is built on publication, translates into appointments, promotion, and research grants. The most prominent hurdle in an academic career is the award of tenure, which is based primarily on peer-reviewed publications. Some people also write text books, which are occasionally very lucrative, but all successful academics write research papers (in the sciences) or monographs (in the humanities). The standing joke is, "Our dean can't read, but he sure can count." The dean counts papers in peer-reviewed journals. Publishers are delighted to make profits by providing things for the dean to count. All faculty are expected to publish several papers every year, whether or not they have important research to report. Because prestige comes from writing many papers, they have an incentive to write several papers that report slightly different aspects of a single piece of research. Studies have shown that most scientific papers are never cited by any other researchers. Many papers are completely unnecessary.

There are conflicting signs whether this situation is changing. Digital libraries provide researchers with alternative ways to tell the world about their research. The process of peer review has considerable value in weeding out bad papers and identifying obvious errors, but it is time consuming; the traditional process of review, editing, printing, and distribution often takes more than a year. An article in a journal that is stored on the shelves of a library is available only to those people who have access to the library and are prepared to make the effort to retrieve the article; a research report on the Internet is available to everybody. In some disciplines, an unstable pattern is emerging of communicating research by rapid, open access publishing of online reports or pre-prints, with traditional journals being used as an archive and for career purposes.

Subscriptions to online journals

The publishers of research journals in science, technology, and medicine include commercial companies, such as Elsevier, John Wiley, and Springer-Verlag, and learned societies, such as the American Association for the Advancement of Science, the publisher of Science. These publishers have been energetic in moving into electronic publication and are some of the first organizations to face the economic challenges. Initially, their general approach has been to retain the standard journal format and to publish electronic versions in parallel to the print.

Since 1996, many scientific publishers have provided electronic versions of their printed journals over the Internet. The online versions are similar to but not always identical to the printed versions. They may leave out some material, such as letters to the editor, or add supplementary data that was too long to include in print. To get started, the publishers have chosen variants on a familiar economic model, selling annual subscriptions to libraries, or library consortia. Publishers that have followed this approach include Academic Press, the American Chemical Society, the Association for Computing Machinery, Elsevier, the American Physical Society, Johns Hopkins University Press, Springer-Verlag, John Wiley, and others. HighWire Press offers a similar service for smaller publishers who publish high-quality journals but do not wish to invest in expensive computer systems. A common model is for the publisher to provide open access to a searchable index and abstracts, but to require payment for access to the full articles. The payment can be by subscription or by a fee per article.

Subscriptions that provide access to an online collection require an agreement between the publisher and the subscriber. If the subscriber is a library, the agreement is usually written as a contract. Although, the details of these agreements vary, a number of topics occur in every agreement. Some of the issues are listed below.

- **Material covered.** When a library subscribes to a print journal for a year, it receives copies of the journal issues for that year, to be stored and used by the library for ever. When a library subscribes to an electronic journal for a year, it typically receives a license to access the publisher's online collection, containing the current year's journal and those earlier issues that have been converted to digital form. The library is relieved of the burden of storing back issues, but it loses access to everything if it does not renew a subscription the following year, or if the publisher goes out of business.

- **The user community.** If the subscription is to a library, the users of that library must be delineated. Some libraries have a well-defined community of users. With corporate libraries and residential universities with full-time students, it is reasonably easy to identify who is authorized to use the materials, but many universities and community colleges have large populations of students who take part-time courses, and staff who are affiliated through hospitals or other local organizations. Public libraries, by definition are open to the public. Fortunately, most libraries have traditionally had a procedure for issuing library cards. One simple approach is for the subscription agreement to covers everybody who is eligible for a library card or physically in the library buildings.

- **Price for different sized organizations.** One problem with subscription-based pricing is how to set the price. Should a university with 5,000 students pay the same subscription as a big state system with a population of 100,000, or a small college with 500? What should be the price for a small research laboratory within a large corporation? Should a liberal arts university pay the same for its occasional use of a medical journal as a medical school where it is a core publication? There is no simple answer to these questions.

- **Pricing relative to print subscriptions.** When material is available in both print and online versions, how should the prices compare? In the long term, electronic publications are cheaper to produce, because of the savings in printing, paper, and distribution. In the short term, electronic publications represent a considerable investment in new systems. Initially, a few publishers attempted to charge higher costs for the electronic versions. Now a consensus is emerging for slightly lower prices. Publishers are experimenting with a variety of pricing options that encourage libraries to subscribe to large groups of journals.

- **Use of the online journals.** One of the more contentious issues is what use can subscribers make of online journals. A basic subscription clearly should allow readers to view journal articles on a computer screen and it would be

> poor service for a publisher not to expect readers to print individual copies for their private use. Conversely, it would be unreasonable for a subscriber to an online journal to make copies and sell them on the open market. Reasonable agreement lies somewhere between these two extremes, but consensus has not yet been reached.

The model of institutional subscriptions has moved the delivery of scientific journals from print to the Internet without addressing any of the underlying tensions between publishers and their customers. It will be interesting to see how well it stands the test of time.

Scientific journals and their authors

The difference in goals between authors and publishers is manifest in the restrictions that publishers place on authors. Many publishers make demands on the authors that restrict the distribution of research. Publishers, quite naturally, will not publish papers that have appeared in other journals, but many refuse to publish research results that have been announced anywhere else, such as at a conference or on a web site. Others permit pre-publication, such as placing a version of the paper on a server, such as the Los Alamos archives, but require the open access version to be removed once the edited journal article is published. This sort of assertiveness antagonizes authors, while there is no evidence that it has any effect on revenues. The antagonism is increased by the publishers insisting that authors transfer copyright to them, leaving the authors few rights in the works that they created, and the taxpayer whose money fuels the process with nothing.

At the Association for Computing Machinery (ACM), we attempted to find a balance between the authors' interest in widespread distribution and the need for revenue. The resulting policy is clearly seen as interim, but hopefully the balance will be acceptable for the next few years. The current version of the policy retains the traditional copyright transfer from author to publisher and affirms that ACM can use material in any format or way that it wishes. At the same time, however, it allows authors great flexibility. In particular, the policy encourages authors to continue to mount their materials on private servers, both before and after publication. In fast moving fields, such as those covered by ACM journals, preprints have always been important and recently these preprints have matured into online collections of technical reports and preprints, freely available over the network. These are maintained privately or by research departments. Since they are important to ACM members, ACM does not want to remove them, yet the association does not want private online collections to destroy the journal market.

The legal framework

This is not a legal text book (though two lawyers have checked this section), but some legal issues are so important that this book would be incomplete without discussing them. Digital libraries reach across almost many areas of human activity. It is unsurprising that many aspects of the law are relevant to digital libraries, since the legal system provides a framework that permits the orderly development of online services. Relevant areas of law include contracts, copyright and other intellectual property, defamation, obscenity, communications law, privacy, tax, and international law.

The legal situation in the United States is complicated by the number of jurisdictions - the Constitution, international treaties, federal and state statutes, and the precedents set by courts at each level. Many laws, such as those controlling obscenity, are at the state level. People who mounted sexual material on a server in California, where it is legal, were prosecuted in Louisiana, where it was deemed to be obscene. Some legal areas, such as copyright, are covered by federal statutes, but important topics have never been interpreted by the Supreme Court. When the only legal rulings have been made by lower courts, other courts are not bound by precedent and may interpret the law differently.

The two communities that are building digital libraries on the Internet - the computer scientists and the information professionals - both have a tradition of responsible use of shared resources. The traditions are different and merging these traditions poses some problems, but a greater difficulty is the influx of people with no traditions to build on. Inevitably, a few of these people are malicious or deliberately exploit the networks for personal gain. Others are thoughtless and unaware of what constitutes reasonable behavior.

Until the late 1980s, the Internet was an academic and research network. There was a policy on who could use it and what was appropriate use, but more importantly the users policed themselves. Somebody who violated the norms was immediately barraged with complaints and other, less subtle forms of peer pressure. The social conventions were somewhat unconventional, but they worked as long as most people were members of a small community and learned the conventions from their colleagues. The conventions began to fail when the networks grew larger. At Carnegie Mellon University, we noticed a change when students who had learned their networked computing outside universities became undergraduates.

Many of the legal issues are general Internet questions and not specific to digital libraries. Currently, the Internet community is working on technical methods to control junk electronic mail, which is a burden on the network suppliers and an annoyance to users. Pornography and gambling are two areas which pit commercial interests against diverse social norms, and advocates of civil liberties against religious groups.

International questions

The Internet is worldwide. Any digital library is potentially accessible from anywhere in the world. Behavior that is considered perfectly normal in one country is often illegal in another. For example, the United States permits the possession of guns, but limits the use of encryption software. Most countries in Europe have the opposite rules.

Attitudes to free speech vary greatly around the world. Every country has laws that limit freedom of expression and access to information; they cover slander, libel, obscenity, privacy, hate, racism, or government secrets. The United States believes fervently that access to information is a fundamental democratic principle. It is enshrined in the First Amendment to the Constitution and the courts have consistently interpreted the concept of free speech broadly, but there are limits, even in the United States. Germany has strong laws on Nazism; Arabic countries are strict about blasphemy. Every jurisdiction expects to be able to control such activities, yet the Internet is hard to control. For years, there was a server in Finland that would act as a relay and post messages on the Internet anonymously. After great pressure from outside, the courts forced this server to disclose the names of people who posted particularly disreputable materials anonymously.

Internet commerce, including electronic information, is a broad area where the international nature of the Internet creates difficulty. In the United States, consumers and suppliers are already protected by laws that cover inter-state commerce, such as financial payments and sales by mail order across state boundaries. The situation is much more complex over a world-wide network, where the trade is in digital materials which can easily be duplicated or modified. On the Internet, the parties to a transaction do not even need to declare from what country they are operating.

Liability

The responsible for the content of library materials is a social and legal issue of particular importance to digital libraries. Society expects the creators of works to be responsible for their content, and that those who make decisions about content should behave responsibility. However, digital libraries will not thrive if legal liability for content is placed upon parties whose only function is to store and transmit information.

Because of the high esteem in which libraries are held, in most democratic countries they have a privileged legal position. It is almost impossible to hold a library liable for libelous statements or subversive opinions expressed in the books that it holds. In a similar way, telecommunications law protects common carriers, so that a telephone company does not have to monitor the conversations that take place on its lines. In fact, the telephone companies are prohibited from deliberately listening to such conversations.

Traditionally, the responsibility for the content has fallen on the creators and publishers, who are aware of the content, rather than on the libraries. This allows libraries to collect material from all cultures and all periods without having to scrutinize every item individually for possible invasions of privacy, libel, copyright violations, and so on. Most people would accept that this is good policy which should be extended to digital libraries. Organizations have a responsibility to account for the information resources they create and distribute, but it is unreasonable for libraries or Internet service companies to monitor everything that they transmit.

Liability of service providers is one of the central topics of the Digital Millennium Copyright Act of 1998, described in Panel 6.2. This act removes the liability for copyright violations from online service providers, including libraries and educational establishments. As usual with emerging law, this was not a simple process that does not end with legislation. Digital libraries naturally welcome the freedom of action, but it can be argued that it goes too far in the protection that it gives to service providers. It will interesting to see how the courts interpret the act.

Panel 6.2
The Digital Millennium Copyright Act

In 1998, the United States Congress passed an act that made significant changes in copyright law. Apart from a section dealing with the design of boat hulls, almost all the act is about digital works on networks. On the surface, the act appears a reasonable balance between commercial interests that wish to sell digital works, and the openness of information that is central to libraries and education. Some of the more important provision are listed below.

Online service providers, libraries, and educational establishments

Under the act, online service providers in the United States, including libraries and educational establishments, are largely protected from legal claims of copyright infringement that take place without their knowledge. To qualify for this exception, specific rules must be followed: the organization must provide users with information about copyright, have a policy for termination of repeat offenders, comply with requirements to take-down infringing materials, support industry-standard technical measures, and advise the Copyright Office of an agent to receive statutory notices under the act.

This section explicitly permits many of the activities that are fundamental to the operation of digital libraries. It allows services for users to store materials such as web sites, to follow hyperlinks, and to use search engines. It recognizes that service providers make copies of materials for technical reasons, notably system caching and to transmit or route material to other sites. These activities are permitted by the act, and service providers are not liable for violations by their users if they follow the rules and correct problems when notified.

For universities and other institutes of higher education, the act makes an important exception to the rule that organizations are responsible for the actions of their employees. It recognizes that administrators do not oversee the actions of faculty and graduate students, and are not necessarily liable for their acts.

Copyright protection systems and copyright management information

The act prohibits the circumvention of technical methods used by copyright owners to restrict access to works. It also prohibits the manufacture or distribution of methods to defeat such technology. However, the act recognizes several exceptions, all of which are complex and need careful interpretation: software developers can reverse engineer protection systems to permit interoperability, researchers can study encryption and system security, law enforcement agencies can be authorized to circumvent security technology, and libraries can examine materials to decide whether to acquire them. Finally, users are permitted to identify and disable techniques that collect private information about users and usage.

The act provides rules on tampering with copyright management information about a work, such as the title, author, performer, and the copyright owner. This information must not be intentional altered or removed.

Copyright

Some of the most heated legal discussions concern the interaction between economic issues and copyright law. Such arguments seem to emerge every time that a new technology is developed. In the early days of printing there was no copyright. Shakespeare's plays were freely pirated. In the nineteenth century, the United States had copyright protection for American authors but none for foreigners; the books of European authors were not protected and were shamelessly copied, despite the helpless pleas of authors, such as Dickens and Trollope.

In United States law, copyright applies to almost all literary works, including textual materials, photographs, computer programs, musical scores, videos and audio tapes. A major exception is materials created by government employees. Initially, the creator of a work or the employer of the creator owns the copyright. In general, this is considered to be intellectual property that can be bought and sold like any other property. Other countries have different approaches; in some countries, notably France, the creator has personal rights (known as "moral rights") which can not be transferred. Historically, copyright has had a finite life, expiring a certain number of years after the creator's death, but Congress has regularly extended that period, most recently when copyright on Mickey Mouse was about to expire - a sad example of the public good being secondary to the financial interests of a few corporations.

The owner of the copyright has an exclusive right to make copies, to prepare derivative works, and to distribute the copies by selling them or in other ways. This is important to authors, helping them to ensure that their works do not get corrupted, either accidentally or maliciously. It also allows publishers to develop products without fear that their market will be destroyed by copies from other sources.

Copyright law is not absolute, however. Although the rights holder has considerable control over how material may be used, the control has boundaries. Two important concepts in United States law are the first sale doctrine and fair use. First sale applies to a physical object, such as a book. The copyright owner can control the sale of a new book, and set the price, but once a customer buys a copy of the book, the customer has full ownership of that copy and can sell the copy or dispose of it in any way without needing permission.

Fair use is a legal right in the United States law that allows certain uses of copyright information without permission of the copyright owner. Under fair use, reviewers or scholars have the right to quote short passages, and photocopies can be made of an article of part of a book for private study. The boundaries of fair use are deliberately vague, but there are four basic factors that are considered:

- the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
- the nature of the copyrighted work;
- the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
- the effect of the use upon the potential market for or value of the copyrighted work.

Because these factors are imprecise, judges have discretion in how they are interpreted. In general, fair use allows reproduction of parts of work rather than the whole, single copies rather than many, and private use rather than commercial use.

The exact distinctions can be clarified only by legal precedence. Although there have been a few, well-publicized court cases, such cases are so expensive that, even for traditional print materials, many important issues have never been tested in court or only in the lower courts.

The first sale doctrine and the concept of fair use do not transfer easily to digital libraries. While the first sale doctrine can be applied to physical media that store electronic materials, such as CD-ROMs, there is no parallel for information that is delivered over networks. The guidelines for fair use are equally hard to translate from physical media to the online world.

This uncertainty was one of the reasons that led to a series of attempts to rewrite copyright law, both in the United States and internationally. While most people accepted that copyright law should provide a balance between the public's right of access to information and economic incentives to creators and publishers of information, there was no consensus what that balance should be, with extremists on both sides. This led to an unsavory situation of vested interests attempting to push through one-sided legislation, motivated by the usual forces of fear and greed. For several years, legislation was introduced in the United States Congress to change or clarify the copyright law relating to information on networks, usually to favor some group of interested parties. At the worst, some commercial organizations lobbied for draconian rights to control information and with criminal sanctions on every activity that is not explicitly authorized. In response, public interest groups argued that there is no evidence that fair use hurts any profits and that unnecessary restrictions on information flow are harmful.

Until 1998, the results were a stalemate, which was probably good. Existing legislation was adequate to permit the first phase of electronic publishing and digital libraries. The fundamental difficulty was to understand the underlying issues. Legal clarification was needed eventually, but it was better to observe the patterns that emerge rather than to rush into premature legislation. The 1998 legislation described in Panel 6.2, above, is probably good enough to allow digital libraries to thrive.

Panel 6.2
Events in the history of copyright

Outside the U.S. Copyright Office there is a sequence of display panels that summarize some major legal decisions about copyright law that have been decided by U.S. federal courts, including the Supreme Court. They illustrate how, over the years, legal precedents shape and clarify the law, and allow it to evolve into areas, such as photography, broadcasting, and computing that were not thought of when the Constitution was written and the laws enacted.

Even these major decisions can not be considered irrevocable. Many were never tested by the Supreme Court and could be reversed. Recently, a federal court made a ruling that explicitly disagreed with the King vs. Mr. Maestro, Inc. case listed below.

Wheaton vs. Peters, 1834. This landmark case, established the principle that copyright is not a kind of natural right but rather is the creation of the copyright statute and subject to the conditions it imposes.

Baker vs. Selden, 1880. This case established that copyright law protects what an author writes and the way ideas are expressed, but the law does not protect the ideas themselves.

Burrow-Giles Lithographic Co. vs. Sarony, 1884. This decision expanded the scope of copyright to cover media other than text, in this case a photograph of Oscar Wilde.

Bleistein vs. Donaldson Lithographic Co., 1903. This case concerned three circus posters. The court decided that they were copyrightable, whether or not they had artistic value or were aesthetically pleasing.

Fred Fisher, Inc. vs. Dillingham, 1924. This dispute concerned the similarity in two musical passages. The court ruled that unconscious copying could result in an infringement of copyright.

Nichols vs. Universal Pictures Corp., 1931. The court ruled that it was not an infringement of copyright for a film to copy abstract ideas of plot and characters from a successful Broadway play.

Sheldon vs. Metro-Goldwyn Pictures Corp., 1936. The court ruled that "no plagiarist can excuse the wrong by showing how much of his work he did not pirate."

G. Ricordi & Co. vs. Paramount Pictures, Inc., 1951. This was a case about how renewal rights and rights in derivative works should be interpreted, in this instance the novel Madame Butterfly by John Luther Long, Belasco's play based on the novel, and Puccini's opera based on the play. The court ruled that copyright protection in derivative works applies only to the new material added.

Warner Bros. Pictures, Inc. vs. Columbia Broadcasting System, Inc., 1955. This case decided that the character Sam Spade in the story The Maltese Falcon was a vehicle for the story, not a copyrightable element of the work.

Mazer vs. Stein, 1954. The court decided that copyright does not protect utilitarian or useful objects, in this case a sculptural lamp. It is possible to register the separable pictorial, graphic, or sculptural features of a utilitarian piece.

King vs. Mr. Maestro, Inc., 1963. This was a case about the speech "I have a dream" by Martin Luther King, Jr.. Although he had delivered the speech to a huge crowd with simultaneous broadcast by radio and television, the court decided that this public performance did not constitute publication and the speech could be registered for copyright as an unpublished work.

Letter Edged in Black Press, Inc., vs. Public Building Commission of Chicago, 1970. This case, about the public display of a Picasso sculpture, has been superseded by later legislation.

Williams Electronics, Inc. vs. Artic International, Inc., 1982. This case involved copying a video game. The court ruled that video game components were copyrightable and that computer read-only memory can be considered a copy.

Norris Industries, Inc. vs. International Telephone and Telegraph Corp., 1983. The court ruled that, even if the Copyright Office rejects a work because it is not copyrightable, the owner is still entitled to file suit and to ask for a court ruling.

Privacy

Libraries, at least in the United States, feel strongly that users have a right to privacy. Nobody should know that a user is consulting books on sensitive issues, such as unpleasant diseases. Libraries have gone to court, rather than divulge to the police whether a patron was reading books about communism. Many states have laws that prohibit libraries from gathering data that violates the privacy of their users. The Internet community has a similar tradition. Although corporations have the legal right to inspect the activities of their employees, most technical people expect their electronic mail and their computer files to be treated as private under most normal circumstances.

Problems arise because much of the technology of digital libraries is also used for electronic commerce. Advertisers and merchants strive to gather the maximum amount of information about their customers, often without the knowledge of the customer. They sell such information to each other. The web has the concept of "cookies", which are useful for such purposes as recording when a user has been authenticated. Unfortunately, the same technology can also be used as a tool for tracking users' behavior without their knowledge.

As discussed in Panel 6.4, digital libraries must gather data on usage. Good data is needed to tune computer systems, anticipate problems, and plan for growth. With care, usage statistics can be gathered without identifying any specific individuals, but not everybody takes care. When a computer system fails, system administrators have the ability to look at any file on a server computer or inspect every message passing over a network. Occasionally they stumble across highly personal information or criminal activities. What is the correct behavior in these circumstance? What should the law say?

Panel 6.4
Digital library statistics and privacy

Anybody would runs a service needs to have data about how it is used. To manage a computer system, such as a digital library, requires data about performance and reliability, about utilization of capacity, about trends and peaks. Libraries and publishers need statistical data about how the collections are used and which services are popular. Designers of user interfaces depend upon knowledge of how people interact with the library. Security requires careful tracking of use patterns and analysis of anomalies.

This information is not easy to gather. Web sites gather data about how often each file is accessed. A common, though doubtful practice, is for sites to boast about how many "hits" they have. Since every graphic image is usually a separate file, a single user who reads one page may generate many hits. This statistic is important for configuring the computer system, but for nothing else. The manager of an online journal would like to know how many people read the journal, with data about the use of individual articles. The number of times that the contents page is accessed is probably useful information. So is the frequency with which the first page of each article is accessed.

These are indirect measures that can not distinguish between one person who reads a story several times and several different people who each read it once. They measure how often the computer system is accessed, but they have no way of estimating how many readers are accessing copies of the materials through caches or mirror sites. If the users are identified it becomes much easier to gather track usage and gather precise statistics, but then privacy becomes a problem.

In our work at Carnegie Mellon we handled this problem as follows. Every time that a user accessed an item a temporary record was created that included a hashed version of the user's ID. The temporary records were kept in a special part of the file system, which was never copied onto back-up tapes. Once a week, the records were analyzed, creating a report which could not be used to identify individual users. It contained information such as the number of different people who had searched each collection and the number of total searches. The temporary records with all trace of the users' identity were then discarded.

Software patents

Although the legal system has its problems, overall it has dealt well with the rapid growth of computing and the Internet. The worst exception is software patents. Few areas of the law are so far removed from the reality they are applied to. In too many cases the Patent Office approves patents that the entire computer industry knows to be foolish. Until recently, the Patent Office did not even employ trained computer scientists to evaluate patent applications. The examiners still award patents that are overly broad, that cover concepts that have been widely known for years, or that are simple applications of standard practice.

Part of the problem is that patent law is based on a concept of invention that does not fit computer science: Archimedes leaps from his bath crying, "Eureka." New ideas in software are created incrementally. The computer science community is quite homogeneous. People are trained at the same universities, use the same computers, and software. There is an open exchange of ideas through many channels. As a result, parallel groups work on the same problems and adapt the same standard techniques in the same incremental ways.

In one of our digital library projects, we did the initial design work as a small team working by itself. A few months later, we met two other groups who had worked on some of the same issues. The three groups had independently found solutions which were remarkably similar. One of the three groups kept their work private and filed a patent application. The other two followed the usual academic tradition of publishing their ideas to the world. Some of these concepts are now widely deployed in digital libraries. In this instance the patent application was turned down, but had it been approved, one group - the smallest contributor in our opinion - would have been in a position to dictate the development of this particular area.

The success of the Internet and the rapid expansion of digital libraries have been fueled by the open exchange of ideas. Patent law, with its emphasis on secrecy, litigation, and confrontation, can only harm such processes.

Footnote

This chapter, more than any other in the book, is a quick review of an extremely complex area. No attempt has been made to describe all the issues. The discussion reflects the author's viewpoint, which will probably need revision in time. Hopefully, however, the basic ideas will stand. Digital libraries are being developed in a world in which issues of users, content, and technology are interwoven with the economic, social, and legal context. These topics have to be studied together and can not be understood in isolation.

Chapter 7
Access management and security

Why control access?

This chapter looks at two related topics: methods for controlling who has access to materials in digital libraries, and techniques of security in networked computing. This book uses the term **access management** to describe the control of access to digital libraries, but other words are also used. Some people refer to "terms and conditions." In publishing, where the emphasis is usually on generating revenue, the strange expression "rights management" is common. Each phrase has a different emphasis, but they are essentially synonymous.

An obvious reason for controlling access is economic. When publishers expect revenue from their products, they permit access only to users who have paid. It might be thought that access management would be unnecessary except when revenue is involved, but that is not the case; there are other reasons to control access to materials in a digital library. Materials donated to a library may have conditions attached, perhaps tied to external events such as the lifetime of certain individuals. Organizations may have information in their private collections that they wish to keep confidential, such as commercial secrets, police records, and classified government information. The boundaries of art, obscenity, and the invasion of privacy are never easy to draw. Even when access to the collections is provided openly, controls are needed over the processes of adding, changing, and deleting material, both content and metadata. A well-managed digital library will keep a record of all changes, so that the collections can be restored if mistakes are made or computer files are corrupted.

Uncertainty is a fact of life in access management. People from a computing background sometimes assume that every object can be labeled with metadata that lists all the rights, permissions, and other factors relevant to access management. People who come from libraries, and especially those who manage historic collections or archives, know that assembling such information is always time consuming and frequently impossible. Projects, such as the American Memory project at the Library of Congress, convert millions of items from historic collections. For these older materials, a natural assumption is that copyright has expired and there need be no access restrictions, but this is far from true. For published materials, the expiration of copyright is linked to the death of the creator, a date which is often hard to determine, and libraries frequently do not know whether items have been published.

As explained in Chapter 6, many of the laws that govern digital libraries, such as copyright, have fuzzy boundaries. Access management policies that are based on these laws are subject to this fuzziness. As the boundaries become clarified through new laws, treaties, or legal precedents, policies have to be modified accordingly.

Elements of access management

Figure 7.1 shows a framework that is useful for thinking about access management. At the left of this figure, information managers create policies for access. **Policies** relate **users** (at the top) to **digital material** (at the bottom). **Authorization**, at the center of the figure, specifies the **access**, at the right. Each of these sections requires elaboration. Policies that the information managers establish must take into account relevant laws, and agreements made with others, such as licenses from copyright holders. Users need to be authenticated and their role in accessing materials established. Digital material in the collections must be identified and its authenticity established. Access is expressed in terms of permitted operations.

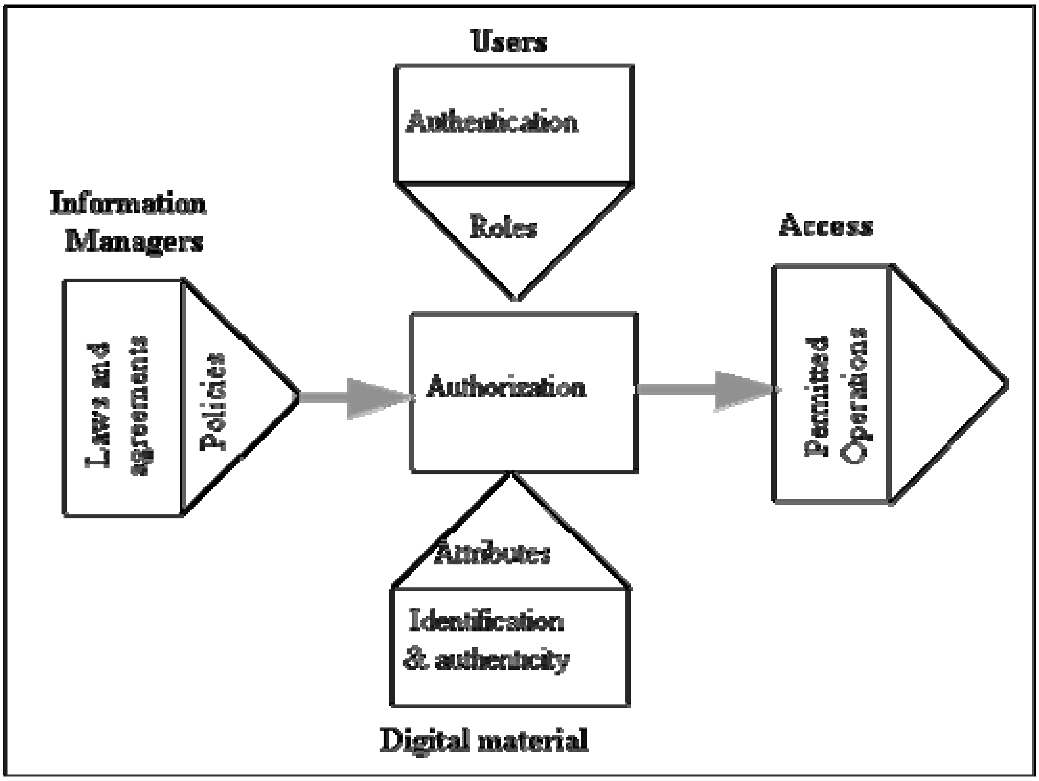


Figure 7.1. A framework for access management

When users request access to the collections, each request passes through an access management process. The users are authenticated; authorization procedures grant or refuse them permission to carry out specified operations.

The responsibility for access lies with whoever manages the digital material. The manager may be a library, a publisher, a webmaster, or the creator of the information. Parts of the responsibility may be delegated. If a library controls the materials and makes them available to users, the library sets the policies and implements them, usually guided by external restraints, such as legal restrictions, licenses from publishers, or agreements with donors. If a publisher mounts materials and licenses access, then the publisher is the manager, but may delegate key activities, such as authorization of users, to others.

Users

Authentication

When a user accesses a computer system, a two-step process of identification usually takes place. The first is **authentication** which establishes the identify of the individual user. The second is to determine what a user is **authorized** to do. A wide variety of techniques are used to authenticate users; some are simple but easy to circumvent, while others are more secure but complex. The techniques divide into four main categories:

- **What does the user know?** A standard method of authentication is to provide each user with a login name and a password. This is widely used but has weaknesses. Passwords are easily stolen. Most people like to select their own password and often select words that are easy to remember and hence easy to guess, such as personal names, or everyday words.

- **What does the user possess?** Examples of physical devices that are used for authentication include the magnetically encoded cards used by bank teller machines, and digital smart-cards that execute an authentication program. Smart-cards are one of the best systems of authentication; they are highly secure and quite convenient to use.

- **What does the user have access to?** A common form of authentication is the network address of a computer. Anybody who has access to a computer with an approved IP address is authenticated. Data on many personal computers is unprotected except by physical access; anybody who has access to the computer can read the data.

- **What are the physical characteristics of the user?** Authentication by physical attributes such as voice recognition is used in a few esoteric applications, but has had little impact in digital libraries.

Roles

Policies for access management rarely specify users by name. They are usually tied to categories of users or the role of a user. An individual user can have many roles. At different times, the same person may use a digital library for teaching, private reading, or to carry out a part-time business. The digital library may have different policies for the same individual in these various roles. Typical roles that may be important include:

- **Membership of a group**. The user is a member of the Institute of Physics. The user is a student at the U.S. Naval Academy.

- **Location.** The user is using a computer in the Carnegie Library of Pittsburgh. The user is in the USA.

- **Subscription.** The user has a current subscription to Journal of the Association for Computing Machinery. The user belongs to a university that has a site license to all JSTOR collections.

- **Robotic use.** The user is an automatic indexing program, such as a Web crawler.

- **Payment.** The user has a credit account with Lexis. The user has paid $10 to access this material.

Most users of digital libraries are people using personal computers, but the user can be a computer with no person associated, such as a program that is indexing web pages or a mirroring program that replicates an entire collection. Some sites explicitly ban access by automatic programs.

Digital material

Identification and authenticity

For access management, digital materials must be clearly identified. Identification associates some name or identifier with each item of material. This is a major topic in both digital libraries and electronic publishing. It is one of the themes of Chapter 12.

Authentication of digital materials assures both users and managers of collections that materials are unaltered. In some contexts this is vital. In one project, we worked with a U.S. government agency to assemble a collection of documents relevant to foreign affairs, such as trade agreements and treaties. With such documents the exact wording is essential; if a document claims to be the text of the North America Free Trade Agreement, the reader must be confident that the text is accurate. A text with wrong wording, whether created maliciously or by error, could cause international problems.

In most digital libraries, the accuracy of the materials is not verified explicitly. Where the level of trust is high and the cost of mistakes are low, no formal authentication of documents is needed. Deliberate alterations are rare and mistakes are usually obvious. In some fields, however, such as medical records, errors are serious. Digital libraries in these areas should seriously consider using formal methods of authenticating materials.

To ensure the accuracy of an object, a **digital signature** can be associated with it, using techniques described at the end of this chapter. A digital signature ensures that a file or other set of bits has not changed since the signature was calculated. Panel 7.1 describes the use of digital signatures in the U.S. Copyright Office.

Panel 7.1
Electronic registration and deposit for copyright

The U.S. Copyright Office is a separate department of the Library of Congress. Under United States law, all works published in the U.S. are subject to mandatory deposit and the library is entitled to receive two copies for its collections. While deposit of published works is mandatory, the copyright law since 1978 has not required registration of copyrighted works, though registration is encouraged by conferring significant benefits.

The method of copyright registration is straightforward. The owner of the copyright sends two copies of the work to the Copyright Office, with an application form and a fee. The copyright claim and the work are examined and, after approval, a registration certificate is produced. Published works are transferred to the Library of Congress, which decides whether to retain the works for its collections or use them in its exchange program.

In 1993, the Copyright Office and CNRI began work on a system, known as CORDS

(Copyright Office Electronic Registration, Recordation and Deposit System) to register and deposit electronic works. The system mirrors the traditional procedures. The submission consists of a web form and a digital copy of the work, delivered securely over the Internet. The fee is processed separately.

Digital signatures are used to identify claims that are submitted for copyright registration in CORDS. The submitter signs the claim with the work attached, using a private key. The submission includes the claim, the work, the digital signature, the public key, and associated certificates. The digital signature verifies to the Copyright Office that the submission was received correctly and confirms the identity of the submitter. If, at any future date, there is a copyright dispute over the work, the digital signature can be used to authenticate the claim and the registered work.

A group of methods that are related to authentication of materials are described by the term watermarking. They are defensive techniques used by publishers to deter and track unauthorized copying. The basic idea is to embed a code into the material in a subtle manner that is not obtrusive to the user, but can be retrieved to establish ownership. A simple example is for a broadcaster to add a corporate logo to a television picture, to identify the source of the picture if is copied. Digital watermarks can be completely imperceptible to a users, yet almost impossible to remove without trace.

Attributes of digital material

Access management policies frequently treat different material in varying ways, depending upon properties or attributes of the material. These attributes can be encoded as administrative metadata and stored with the object, or they can be derived from some other source. Some attributes can also be computed. Thus the size of an object can be measured when required. Here are some typical examples:

- **Division into sub-collections.** Collections often divide material into items for public access and items with restricted access. Publishers may separate the full text of articles from indexes, abstracts and promotional materials. Web sites have public areas, and private areas for use within an organization.

- **Licensing and other external commitments.** A digital library may have material that is licensed from a publisher or acquired subject to terms and conditions that govern access, such as materials that the Library of Congress receives through copyright deposit.

- **Physical, temporal, and similar properties.** Digital libraries may have policies that depend upon the time since the date of publication or physical properties, such as the size of the material. Some newspapers provide open access to selected articles when they are published while requiring licenses for the same articles later.

- **Media types.** A digital library may have access policies that depend upon format or media type, for example treating digitized sound differently from textual material, or computer programs from images.

Attributes need to be assigned at varying granularity. If all the materials in a collection have the same property, then it is convenient to assign the attribute to the collection as a whole. At the other extreme, there are times when parts of objects may have specific properties. The rights associated with images are often different from those associated with the text in which they are embedded and will have to be distinguished. A donor may donate a collection of letters to a library for public access, but request that access to certain private correspondence be restricted. Digital libraries need to offer flexibility, so that attributes can be associated with entire collections, sub-collections, individual library objects, or elements of individual objects.

Operations

Access management policies often specify or restrict the operations and the various actions that a user is authorized to carry out on library materials. Some of the common categories of operation include:

- **Computing actions.** Some operations are defined in computing terms, such as to write data to a computer, execute a program, transmit data across a network, display on a computer screen, print, or copy from one computer to another.

- **Extent of use.** A user may be authorized to extract individual items from a database, but not copy the entire database.

These operations can be controlled by technical means, but many policies that an information manager might state are essentially impossible to enforce technically. They include:

- **Business or purpose.** Authorization of a user might refer to the reason for carrying out an operation. Examples include commercial, educational, or government use.

- **Intellectual operations.** Operations may specify the intellectual use to be made of an item. The most important is the rules that govern the creation of a new work that is derived from the content of another. The criteria may need to consider both the intent and the extent of use.

Subsequent use

Systems for access management have to consider both direct operations and subsequent use of material. Direct operations are actions initiated by a repository, or another computer system that acts as a agent for the managers of the collection. Subsequent use covers all those operations that can occur once material leaves the control of the digital library. It includes all the various ways that a copy can be made, from replicating computer files to photocopying paper documents. Intellectually, it can include everything from extracting short sections, the creation of derivative works, to outright plagiarism.

When an item, or part of a item, has been transmitted to a personal computer it is technically difficult to prevent a user from copying what is received, storing it, and distributing it to others. This is comparable to photocopying a document. If the information is for sale, the potential for such subsequent use to reduce revenue is clear. Publishers naturally have concerns about readers distributing unauthorized copies of materials. At an extreme, if a publisher sells one copy of an item that is subsequently widely distributed over the Internet, the publisher might end up selling only that one copy. As a partial response to this fear, digital libraries are often designed to allow readers access to individual records, but do not provide any way to copy complete collections. While this does not prevent a small loss of revenue, it is a barrier against anybody undermining the economic interests of the publisher by wholesale copying.

Policies

The final branch of Figure 7.1 is the unifying concept of a policy, on the left-hand side of the figure. An informal definition of a policy is that it is a rule, made by information managers that states who is authorized to do what to which material. Typical policies in digital libraries are:

- A publication might have the policy of open access. Anybody may read the material, but only the editorial staff may change it.
- A publisher with journals online may have the policy that only subscribers have access to all materials. Other people can read the contents pages and abstracts, but have access to the full content only if they pay a fee per use.
- A government organization might classify materials, e.g., "top secret", and have strict policies about who has access to the materials, under what conditions, and what they can do with them.

Policies are rarely as simple as in these examples. For example, while *D-Lib Magazine* has a policy of open access, the authors of the individual articles own the copyright. The access policy is that everybody is encouraged to read the articles and print copies for private use, but some subsequent use, such as creating a derivative work, or selling copies for profit requires permission from the copyright owner. Simple policies can sometimes be represented as a table in which each row relates a user role, attributes of digital material, and certain operations.

Because access management policies can be complex, a formal method is needed to express them, which can be used for exchange of information among computer system. Perhaps the most comprehensive work in this area has been carried out by Mark Stefik of Xerox. The Digital Property Rights Language, which he developed, is a language for expressing the rights, conditions, and fees for using digital works. The purpose of the language is to specify attributes of material and policies for access, including subsequent use. The manager of a collection can specify terms and conditions for copying, transferring, rendering, printing, and similar operations. The language allows fees to be specified for any operation, and it envisages links to electronic payment mechanisms. The notation used by the language is based on Lisp, a language used for natural language processing. Some people have suggested that, for digital libraries, a more convenient notation would use XML, which would be a straightforward transformation. The real test of this language is how effective it proves to be when used in large-scale applications.

Enforcing access management policies

Access management is not simply a question of developing appropriate policies. Information managers want the policies to be followed, which requires some form of enforcement.

Some policies can be enforced technically. Others are more difficult. There are straightforward technical methods to enforce a policy of who is permitted to change material in a collection, or search a repository. There are no technical means to enforce a policy against plagiarism, or invasion of privacy, or to guarantee that all use is educational. Each of these is a reasonable policy that is extremely difficult to enforce by technical means. Managing such policies is fundamentally social.

There are trade-offs between strictness of enforcement and convenience to users. Technical method of enforcing policies can be annoying. Few people object to typing in a password when they begin a session, but nobody wants to be asked repeatedly for passwords, or other identification. Information managers will sometimes decide to be relaxed about enforcing policies in the interests of satisfying users. Satisfied customers will help grow the size of the market, even if some revenue is lost from unauthorized users. The publishers who are least aggressive about enforcement keep their customers happy and often generate most total revenue. As discussed in Panel 7.2, this is the strategy now used for most personal computer software. Data from publishers such as HighWire Press is beginning to suggest the same result with electronic journal publishing.

If technical methods are relaxed, social and legal pressures can be effective. The social objective is to educate users about the policies that apply to the collections, and coax or persuade people to follow them. This requires policies that are simple to understand and easy for users to follow. Users must be informed of the policies and educated as to what constitutes reasonable behavior. One useful tool is to display an **access statement** when the material is accessed; this is text that states some policy. An example is, "For copyright reasons, this material should not be used for commercial purposes." Other non-technical methods of enforcement are more assertive. If members of an organization repeatedly violate a licensing agreement or abuse policies that they should respect, a publisher can revoke a license. In extreme cases, a single, well-publicized legal action will persuade many others to behave responsibly.

Panel 7.2
Access management policies for computer software

Early experience with software for personal computers provides an example of what happens when attempts to enforce policies are unpleasant for the users.

Software is usually licensed for a single computer. The license fee covers use on one computer only, but software is easy to copy. Manufacturers lose revenue from unlicensed copying, particularly if there develops widespread distribution of unlicensed copies.

In the early days of personal computers, software manufacturers attempted to control unlicensed copying by technical means. One approach was to supply their products on disks that could not be copied easily. This was called copy-protection. Every time that a program was launched, the user had to insert the original disk. This had a powerful effect on the market, but not what the manufacturers had hoped. The problem was that it became awkward for legitimate customers to use the software that they had purchased. Hard disk installations were awkward and back-up difficult. Users objected to the inconvenience. Those software suppliers who were most assertive about protection lost sales to competitors who supplied software without copy-protection.

Microsoft was one of the companies that realized that technical enforcement is not the only option. The company has become extremely rich by selling products that are not technically protected against copying. Instead, Microsoft has worked hard to stimulate adherence to its policies by non-technical methods. Marketing incentives, such as customer support and low-cost upgrades, encourage customers to pay for licenses. Social pressures are used to educate people and legal methods are used to frighten the worst offenders.

Unlicensed copying still costs the software manufacturers money, but, by concentrating on satisfying their responsible customers, the companies are able to thrive.

Access management at a repository

Most digital libraries implement policies at the repository or collection level. Although there are variations in the details, the methods all follow the outline in Figure 7.1. Digital libraries are distributed computer systems, in which information is passed from one computer to another. If access management is only at the repository, access is effectively controlled locally, but once material leaves the repository problems multiply.

The issue of subsequent use has already been introduced; once the user's computer receives information it is hard for the original manager of the digital library to retain effective control, without obstructing the legitimate user. With networks, there is a further problem. Numerous copies of the material are made in networked computers, including caches, mirrors, and other servers, beyond the control of the local repository.

To date, most digital libraries have been satisfied to provide access management at the repository, while relying on social and legal pressure to control subsequent use. Usually this is adequate, but some publishers are concerned that the lack of control could damage their revenues. Therefore, there is interest in technical methods that control copying and subsequent, even after the material has left the repository. The methods fall into two categories: trusted systems and secure containers.

Trusted systems

A repository is an example of a trusted system. The managers of a digital library have confidence that the hardware, software, and administrative procedures provide an adequate level of security to store and provide access to valuable information. There may be other systems, linked to the repository, that are equally trusted. Within such a network of trusted systems, digital libraries can use methods of enforcement that are simple extensions of those used for single repositories. Attributes and policies can be passed among systems, with confidence that they will be processed effectively.

Implementing networks of trusted systems is not easy. The individual systems components must support a high level of security and so must the processes by which information is passed among the various computers. For these reasons, trusted systems are typically used in restricted situations only or on special purpose computers. If all the computers are operated by the same team or by teams working under strict rules, many of the administrative problems diminish. An example of a large, trusted system is the network of computers that support automatic teller machines in banks.

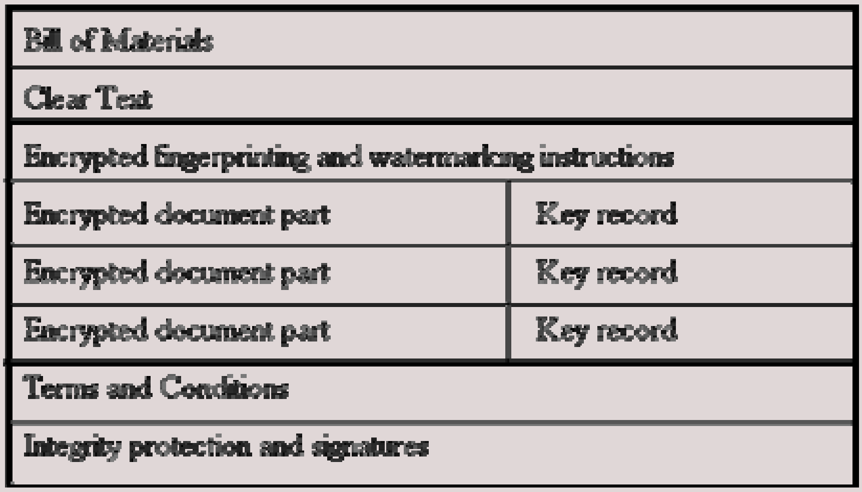
No assumptions can be made about users' personal computers and how they are managed. In fact, it is reasonable not to trust them. For this reason, early applications of trusted systems in digital libraries are likely to be restricted to special purpose hardware, such as smart cards or secure printers, or dedicated servers running rightly controlled software.

Secure containers

Since networks are not secure and trusted system difficult to implement, several groups are developing secure containers for transmitting information across the Internet. Digital material is delivered to the user in a package that contains data and metadata about access policies. Some or all of the information in the package is encrypted. To access the information requires a digital key, which might be received from an electronic payment system or other method of authentication. An advantage of this approach is that it provides some control over subsequent use. The package can be copied and distributed to third parties, but the contents can not be accessed without the key. Panel 7.3 describes one such system, IBM's Cryptolopes.

Panel 7.3. Cryptolopes

IBM's Cryptolope system is an example of how secure containers can be used. Cryptolopes are designed to let Internet users buy and sell content securely over the Internet. The figure below gives an idea of the structure of information in a Cryptolope.



Bill of Materials	
Clear Text	
Encrypted fingerprinting and watermarking instructions	
Encrypted document part	Key record
Encrypted document part	Key record
Encrypted document part	Key record
Terms and Conditions	
Integrity protection and signatures	

Figure 7.2. The structure of a Cryptolope

Information is transmitted in a secure cryptographic envelope, called a Cryptolope container. Information suppliers seal their information in the Cryptolope container. It can be opened by recipients only after they have satisfied any access management requirements, such as paying for use of the information. The content is never separated from the access management and payment information in the envelope. Thus, the envelope can later be passed on to others, who also must pay for usage if they want to open it; each user must obtain the code to open the envelope.

In addition to the encrypted content, Cryptolope containers can include subfiles in clear text to provide users with a description of the product. The abstract might include the source, summary, author, last update, size, and price, and terms of sale. Once the user has decided to open the contents of a Cryptolope container, a digital key is issued unlocking the material contained within. To view a free item, the user clicks on the abstract and the information appears on the desktop. To view priced content, the user agrees to the terms of the Cryptolope container as stated in the abstract.

The content in a Cryptolope container can be dynamic. The system has the potential to wrap JavaScripts, Java programs, and other live content into secure containers. In the interest of standardization, IBM has licensed Xerox's Digital Property Rights Language for specifying the rules governing the use and pricing of content.

Secure containers face a barrier to acceptance. They are of no value to a user unless the user can acquire the necessary cryptographic keys to unlock them and make use of the content. This requires widespread deployment of security service and methods of electronic payment. Until recently, the spread of such services has been rather slow, so that publishers have had little market for information delivered via secure containers.

Security of digital libraries

The remainder of this chapter looks at some of the basic methods of security that are used in networked computer systems. These are general purpose methods with applications far beyond digital libraries, but digital libraries bring special problems because of the highly decentralized networks of suppliers and users of information.

Security begins with the system administrators, the people who install and manage the computers and the networks that connect them. Their honesty must be above suspicion, since they have privileges that provide access to the internals of the system. Good systems administrators will organize networks and file systems so that user have access to appropriate information. They will manage passwords, install firewalls to isolate sections of the networks, and run diagnostic programs to search for problems. They will back-up information, so that the system can be rebuilt after a major incident whether it is an equipment breakdown, a fire, or a security violation.

The Internet is basically not secure. People can tap into it and observe the packets of information traveling over the network. This is often done for legitimate purposes, such as trouble-shooting, but it can also be done for less honest reasons. The general security problem can be described as how to build secure applications across this insecure network.

Since the Internet is not secure, security in digital libraries begins with the individual computers that constitute the library and the data on them, paying special attention to the interfaces between computers and local networks. For many personal computers, the only method of security is physical restrictions on who uses the computer. Other computers have some form of software protection, usually a simple login name and password. When computers are shared by many users, controls are needed to determine who may read or write to each file.

The next step of protection is to control the interface between local networks and the broader Internet, and to provide some barrier to intruders from outside. The most complete barrier is isolation, having no external network connections. A more useful approach is to connect the internal network to the Internet through a special purpose computer called a **firewall**. The purpose of a firewall is to screen every packet that attempts to pass through and to refuse those that might cause problems. Firewalls can refuse attempts from outside to connect to computers within the organization, or reject packets that are not formatted according to a list of approved protocols. Well-managed firewalls can be quite effective in blocking intruders.

Managers of digital libraries need to have a balanced attitude to security. Absolute security is impossible, but moderate security can be built into networked computer systems, without excessive cost, though it requires thought and attention. Universities have been at the heart of networked computing for many years. Despite their polyglot communities of users, they have succeeded in establishing adequate security for campus networks with thousands of computers. Incidents of abusive, anti-social, or malicious behavior occur on every campus, yet major problems are rare.

With careful administration, computers connected to a network can be made reasonably secure, but that security is not perfect. There are many ways that an ill-natured person can attempt to violate security. In universities, most problems come from insiders: disgruntled employees or students who steal a user's login name and password. More sophisticated methods of intrusion take advantage of the complexity of computer software. Every operating system has built-in security, but design errors or programming bugs may have created gaps. Some of the most useful programs for digital libraries, such as web servers and electronic mail, are some of the most difficult to secure. For these reasons, everybody who builds a digital library must recognize that security can never be guaranteed. With diligence, troubles can be kept rare, but there is always a chance of a flaw.

Encryption

Encryption is the name given to a group of techniques that are used to store and transmit private information, encoding it in a way that the information appears completely random until the procedure is reversed. Even if the encrypted information is read by somebody who is unauthorized, no damage is done. In digital libraries, encryption is used to transmit confidential information over the Internet, and some information is so confidential that it is encrypted wherever it is stored. Passwords are an obvious example of information that should always be encrypted, whether stored on computers or transmitted over networks. In many digital libraries, passwords are the only information that needs to be encrypted.

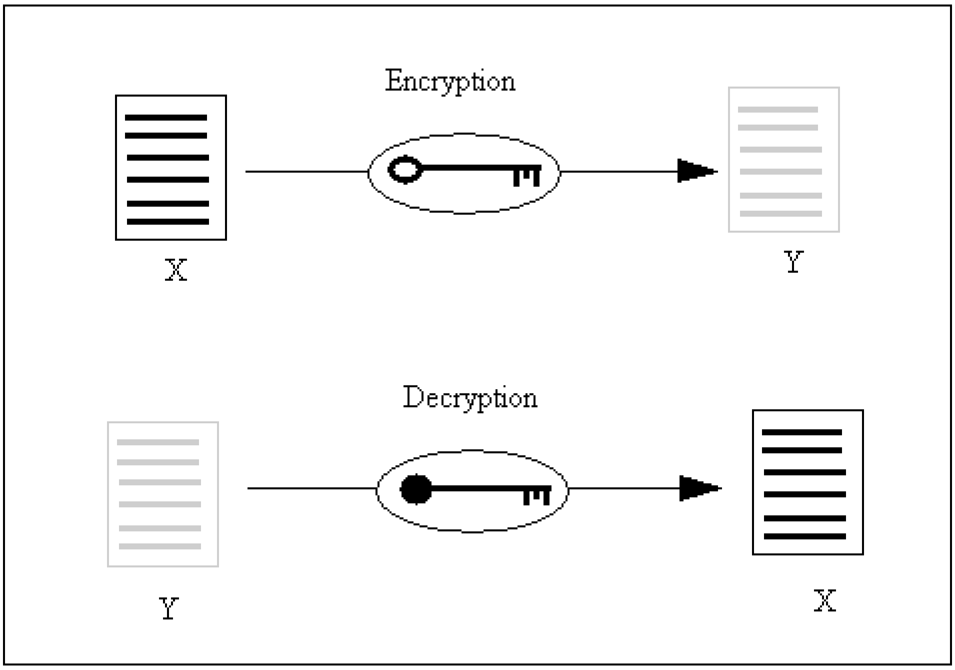


Figure 7.3. Encryption and decryption

The basic concept of encryption is shown in Figure 7.3. The data that is to be kept secret, X, is input to an encryption process which performs a mathematical transformation and creates an encrypted set of data, Y. The encrypted set of data will have the same number of bits as the original data. It appears to be a random collection of bits, but the process can be reversed, using a reverse process which regenerates the original data, X. These two processes, encryption and decryption, can be implemented as computer programs, in software or using special purpose hardware.

The commonly used methods of encryption are controlled by a pair of numbers, known as **keys**. One key is used for encryption, the other for decryption. The methods of encryption vary in the choice of processes and in the way the keys are selected. The mathematical form of the processes are not secret. The security lies in the keys. A key is a string of bits, typically from 40 to 120 bits or more. Long keys are intrinsically much more secure than short keys, since any attempt to violate security by guessing keys is twice as difficult for every bit added to the key length.

Historically, the use of encryption has been restricted by computer power. The methods all require considerable computation to scramble and unscramble data. Early implementations of DES, the method described in Panel 7.4, required special hardware to be added to every computer. With today's fast computers, this is much less of a problem, but the time to encrypt and decrypt large amounts of data is still noticeable. The methods are excellent for encrypting short message, such as passwords, or occasional highly confidential messages, but the methods are less suitable for large amounts of data where response times are important.

Private key encryption

Private key encryption is a family of methods in which the key used to encrypt the data and the key used to decrypt the data are the same, and must be kept secret. Private key encryption is also known as single key or secret key encryption. Panel 7.4 describes DES, one of the most commonly used methods.

Panel 7.4
The Data Encryption Standard (DES)

The Data Encryption Standard (DES) is a method of private key encryption originally developed by IBM. It has been a U.S. standard since 1977. The calculations used by DES are fairly slow when implemented in software, but the method is fast enough for many applications. A modern personal computer can encrypt about one million bytes per second.

DES uses keys that are 56 bits long. It divides a set of data into 64 bit blocks and encrypts each of them separately. From the 56 bit key, 16 smaller keys are generated. The heart of the DES algorithm is 16 successive transformations of the 64 bit block, using these smaller keys in succession. Decryption uses the same 16 smaller keys to carry out the reverse transactions in the opposite order. This sounds like a simple algorithm, but there are subtleties. Most importantly, the bit patterns generated by encryption appear to be totally random, with no clues about the data or the key.

Fanatics argue that DES with its 56 bit keys can be broken simply by trying every conceivable key, but this is a huge task. For digital library applications it is perfectly adequate.

Private key encryption is only as secure as the procedures that are used to keep the key secret. If one computer wants to send encrypted data to a remote computer it must find a completely secure way to get the key to the remote computer. Thus private key encryption is most widely used in applications where trusted services are exchanging information.

Dual key encryption

When using private key encryption over a network, the sending computer and the destination must both know the key. This poses the problem of how to get started if one computer can not pass a key secretly to another. **Dual key encryption** permits all information to be transmitted over a network, including the public keys, which can be transmitted completely openly. For this reason, it has the alternate name of **public key encryption**. Even if every message is intercepted, the encrypted information is still kept secret.

The RSA method is the best known method of dual key encryption. It requires a pair of keys. The first key is made public; the second is kept secret. If an individual, A, wishes to send encrypted data to a second individual, B, then the data is encrypted using the public key of B. When B receives the data it can be decrypted, using the private key, which only B knows.

This dual key system of encryption has many advantages and one major problem. The problem is to make sure that a key is genuinely the public key of a specific individual. The normal approach is to have all keys generated and authenticated by a trusted authority, called a certification authority. The certification authority generates certificates, which are signed messages specifying an individual and a public key. This works well, so long as security at the certificate authority is never violated.

Digital signatures

Digital signatures are used to check that a computer file has not been altered. Digital signatures are based on the concept of a hash function. A **hash** is a mathematical function that can be applied to the bytes of a computer file to generate a fixed-length number. One commonly used hash function is called MD5. The MD5 function can be applied to any length computer file. It carries out a special transformation on the bits of the file and ends up with an apparently random 128 bits.

If two files differ by as little as one bit, their MD5 hashes will be completely different. Conversely, if two files have the same hash, there is an infinitesimal probability that they are not identical. Thus a simple test for whether a file has been altered is to calculate the MD5 hash when the file is created; at a later time, to check that no changes have taken place, recalculate the hash and compare it with the original. If the two are the same then the files are almost certainly the same.

The MD5 function has many strengths, including being fast to compute on large files, but, as with any security device, there is always a possibility that some bright person may discover how to reverse engineer the hash function, and find a way to create a file that has a specific hash value. At the time that this book was being written there were hints that MD5 may be vulnerable in this way. If so, other hash functions are available.

A hash value gives no information about who calculated it. A **digital signature** goes one step further towards guaranteeing the authenticity of a library object. When the hash value is calculated it is encrypted using the private key of the owner of the material. This together with the public key and the certificate authority creates a digital signature. Before checking the hash value the digital signature is decrypted using the public key. If the hash results match, then the material is unaltered and it is known that the digital signature was generated using the corresponding private key.

Digital signatures have a problem. While users of a digital library want to be confident that material is unaltered, they are not concerned with bits; their interest lies in the content. For example, the Copyright Office pays great attention to the intellectual content, such as the words in text, but does not care that a computer system may have attached some control information to a file, or that the font used for the text has been changed, yet the test of a digital signature fails completely when one bit is changed. As yet, nobody has suggested an effective way to ensure authenticity of content, rather than bits.

Deployment of public key encryption

Since the basic mathematics of public key encryption are now almost twenty years old, it might be expected that products based on the methods would have been widely deployed for many years. Sadly this is not the case.

One reason for delay is that there are significant technical issues. Many of them concern the management of the keys, how they are generated, how private keys are stored, and what precautions can be taken if the agency that is creating the keys has a security break-in. However, the main problems are policy problems.

Patents are part of the difficulty. Chapter 6 discussed the problems that surround software patents. Public key encryption is one of the few areas where most computer scientists would agree that there were real inventions. These method are not obvious and their inventors deserve the rewards that go with invention. Unfortunately, the patent holders and their agents have followed narrow licensing policies, which have restricted the creative research that typically builds on a break-through invention.

A more serious problem has been interference from U.S. government departments. Agencies such as the CIA claim that encryption technology is a vital military secret and that exporting it would jeopardize the security of the United States. Police forces claim that public safety depends upon their ability to intercept and read any messages on the networks, when authorized by an appropriate warrant. The export argument is hard to defend when the methods are widely published overseas, and reputable companies in Europe and Japan are building products that incorporate them. The public safety augment is more complicated, but it is undercut by the simple fact that the American public does not trust the agencies, neither their technical competence nor their administrative procedures. People want the ability to transmit confidential information without being monitored by the police.

The result of these policy problems has been to delay the deployment of the tools that are needed to build secure applications over the Internet. Progress is being made and, in a few years time we may be able to report success. At present it is a sorry story.

It is appropriate that this chapter ends with a topic in which the technical solution is held up by policy difficulties. This echoes a theme that recurs throughout digital libraries and is especially important in access management. People, technology, and administrative procedures are intimately linked. Successful digital libraries combine aspects of all three and do not rely solely on technology to solve human problems.

Chapter 8
User interfaces and usability

The person who uses a library is conventionally called a "patron" or "reader." In computing, the usual term is "user" or "end user". Whatever word is chosen, digital libraries are of little value unless they are easy to use effectively.

The relationship between people and computers is the subject of intensive research, drawing on field as diverse as cognitive science, graphic design, rhetoric, and mathematical modeling of computer systems. Some of the research aims to develop a theoretical understanding of how people interact with computers, so that models of how people process information can be used to design appropriate computer systems. Other research helps the user comprehend the principles behind a computer system, to stimulate productive use of the services and information that it provides. This chapter, however, is less ambitious. It concentrates on the methods that are widely used today, with some illustrations of experimental systems that show promise.

Change is one of the themes of digital libraries and change is one of the problems in designing user interfaces. Traditional libraries are not easy to use effectively, but they change slowly; users develop expertise over many years. Digital libraries evolve so quickly that every month brings new services, new collections, new interfaces, and new headaches. Users do not enjoy having to relearn basic skills, but change appears to be a necessary evil.

Partly because of this rapid change, users have an enormous variety in their levels of expertise. Much of the development of digital libraries has come out of universities, where there are many experts. Colleagues and librarians are at hand to help users, while system administrators configure computers, install software, and track changes in the market place. As the Internet has spread more widely, digital libraries are being used by people who do not have access to such expertise and do not want to spend their own time learning techniques that may be of transitory value. This creates a tension. Advanced features of a library are valuable for the specialist. They enable the skilled user to work faster and be more effective, but digital libraries must also be usable by people with minimal training.

Aspects of usability and user interface design

In discussing the **usability** of a computer system, it is easy to focus on the design of the interface between the user and the computer, but usability is a property of the total system. All the components must work together smoothly to create an effective and convenient digital library, for both the patrons, and for the librarians and systems administrators.

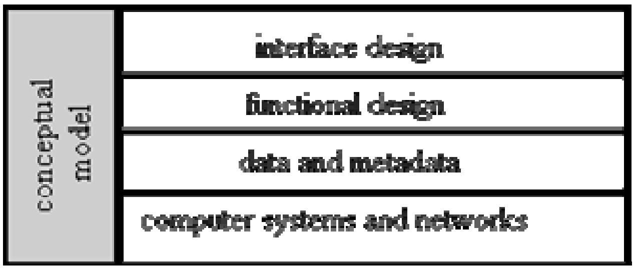


Figure 8.1. Aspects of usability

Figure 8.1 shows a way to think about usability and the design of user interfaces. In any computer system, the user interface is built on a **conceptual model** that describes the manner in which the system is used. Here are some typical conceptual models that are used to design digital libraries. In practice, most digital libraries combine concepts from several such models.

- The classical library model is to search a catalog or index, select objects from the results, and retrieve them from a repository.
- The basic web model is for the user to follow hyperlinks between files.
- The Z39.50 protocol supports a conceptual model of searching a collection, and storing sets of results for subsequent manipulation and retrieval.

The right-hand side of Figure 8.1 shows the layers that are needed to implement any conceptual model. At the top is the **interface design**, the appearance on the screen and the actual manipulation by the user. This covers such considerations as fonts, colors, logos, key board controls, menus, and buttons. The next level down, the **functional design**, represents the functions that are offered to the user. Typical functions include selection of parts of a digital object, searching a list or sorting the results, help information, and the manipulation of objects that have been rendered on a screen. These functions are made possible by the two bottom sections of the figure: the **data and metadata** that are provided by the digital library, and the underlying **computer systems and networks**. Panel 8.1 illustrates these five aspects by showing how they apply to an important application.

Panel 8.1
Aspects of a user interface: page turning

Page turning is an example that illustrates the distinction between the various aspects of user interface design shown in Figure 8.1. Conversion projects, such as JSTOR or American Memory, have collections of digital objects, each of which is a sets of page images scanned from a book or other printed materials.

Conceptual model

The conceptual model is that the user interacts with an object in much the same manner as with a book. Often the pages will be read in sequence, but the reader may also go back a page or jump to a different page. Some pages may be identified as special, such as the table of contents or an index. Since many personal computers have screens that are smaller than printed pages and have lower resolution, the conceptual model includes zooming and panning across a single page.

Interface design

The interface design defines the actual appearance on the screen, such as the choice of frames, icons, colors, and visual clues to help the user. It also includes decisions about how the individual functions are offered to the user. The interface design determines the appearance of icons, the wording on the buttons, and their position on the screen. The design also specifies whether panning and zooming are continuous or in discrete steps.

When we built a page turner at Carnegie Mellon University, the interface design maximized the area of the screen that was available for displaying page images. Most manipulations were controlled from the keyboard; the arrow keys were used for panning around an individual page, with the tab key used for going to the next page. An alternative design would have these functions controlled by buttons on the screen, but with less space on the screen for the page images.

Functional design

To support this conceptual model, the design provides functions that are equivalent to turning the pages of a book. These functions include: go to the first, next, previous, or last page. There will be functions that relate to the content of a specific page, such as: go to the page that has a specific page number printed on it, or go to a page that is designated as the contents page. To support panning and zooming within a page, other functions of the user interface move the area displayed up or and down one screen, and zoom in or out.

Data and metadata

The functions offered to the user depend upon the digital objects in the collections and especially on the structural metadata. The page images will typically be stored as compressed files which can be retrieved in any sequence. To turn pages in sequence, structural metadata must identify the first page image and list the sequence of the other images. To go to the page with a specific page number printed on it requires structural metadata that relates the page numbers to the sequence of page images, since it is rare that the first page of a set is the page with number one. Zooming and panning need metadata that states the dimensions of each page.

Computer systems and networks

The user interface is only as good as the performance of the underlying system. The time to transmit a page image across a network can cause annoying delays to the user. One possible implementation is to anticipate demands by sending pages from the repository to the user's computer before the user requests them, so that at least the next pages in sequence is ready in memory. This is known as "pre-fetching".

For the Carnegie Mellon page turner, priority was given to quick response, about one second to read and display a page image, even when transmitted over the busy campus network. This led to a pipelined implementation in which the first part of a page was being rendered on the user's computer even before the final section was read into the repository computer from disk storage.

The desk-top metaphor

Almost all personal computers today have a user interface of the style made popular on Apple's Macintosh computers, and derived from earlier research at Xerox's Palo Alto Research Center. It uses a metaphor of files and folders on a desktop. Its characteristics include overlapping windows on the screen, menus, and a pointing device such as a mouse. Despite numerous attempts at improvements, this style of interface dominates the modern computer market. A user of an early Macintosh computer who was transported through fifteen years and presented with a computer running Microsoft's latest system would find a familiar user interface. Some new conventions have been introduced, the computer hardware is greatly improved, but the basic metaphor is the same.

In the terminology of Figure 8.1, the conceptual model is the desktop metaphor; files are thought of as documents that can be moved to the desktop, placed in folders, or stored on disks. Every windows-based user interface has this same conceptual model, but the interface designs vary. For example, Apple uses a mouse with one button, Microsoft uses two buttons, and Unix systems usually have three. The functions to support this model include open and closing files and folders, selecting them, moving them from one place to another, and so on. These functions differ little between manufacturers, but the systems differ in the metadata that they use to support the functions. The desktop metaphor requires that applications are associated with data files. Microsoft and Unix systems use file naming conventions; thus files with names that end ".pdf" are to be used with a PDF viewer. Apple stores such metadata in a separate data structure which is hidden from users. Finally, differences in the underlying computer systems permit some, but not all, user interfaces to carry out several tasks simultaneously.

Browsers

The introduction of browsers, notably Mosaic in 1993, provided a stimulus to the quality of user interfaces for networked applications. Although browsers were designed for the web, they are so flexible that they are used as the interface to almost ever type of application on the Internet, including digital libraries. Before the emergence of general purpose browsers, developers had to provide a separate user interface for every type of computer and each different computing environment. These interfaces had to be modified whenever the operating systems changed, a monumental task that few attempted and essentially nobody did well. By relying on web browsers for the actual interaction with the user, the designer of a digital library can now focus on how to organize the flow of information to the user, leaving complexities of hardware and operating systems to the browser.

Basic functions of web browsers

The basic function of a browser is to retrieve a remote file from a web server and render it on the user's computer.

- To locate the file on the web server, the browser needs a URL. The URL can come from various sources. It can be typed in by the user, be a link within an HTML page, or it can be stored as a bookmark.

- From the URL, the browser extracts the protocol. If it is HTTP, the browser then extracts from the URL the domain name of the computer on which the file is stored. The browser sends a single HTTP message, waits for the response, and closes the connection.

- If all goes well, the response consists of a file and a MIME type. To render the file on the user's computer, the browser examines the MIME type and invokes the appropriate routines. These routines may be built into the browser or may be an external program invoked by the browser.

Every web browser offers support for the HTTP protocol and routines to render pages in the HTML format. Since both HTTP and HTML are simple, a web browser need not be a complex program.

Extending web browsers beyond the web

Browsers were developed for the web; every browser supports the core functions of the web, including the basic protocols, and standard formats for text and images. However, browsers can be extended to provide other services while retaining the browser interface. This extensibility is a large part of the success of browsers, the web, and indeed of the whole Internet. Mosaic had three types of extension which have been followed by all subsequent browsers:

- **Data types.** With each data type, browsers associate routines to render files of that type. A few types have been built into all browsers, including plain text, HTML pages, and images in GIF format, but users can add additional types through mechanisms such as helper applications and plug-ins. A **helper application** is a separate program that is invoked by selected data types. The source file is passed to the helper as data. For example, browsers do not have built-in support for files in the PostScript format, but many users have a PostScript viewer on their computer which is used as a helper application. When a browser receives a file of type PostScript, it starts this viewing program and passes it the file to be displayed. A **plug-in** is similar to a helper application, except that it is not a separate program. It is used to render source files of non-standard formats, within an HTML file, as part of a single display.

- **Protocols.** HTTP is the central protocol of the web, but browsers also support other protocols. Some, such as Gopher and WAIS, were important historically because they allowed browsers to access older information services. Others, such as Netnews, electronic mail and FTP, remain important. A weakness of most browsers is that the list of protocols supported is fixed and does not allow for expansion. Thus there is no natural way to add Z39.50 or other protocols to browsers.

- **Executing programs.** An HTTP message sent from a browser can do more than retrieve a static file of information from a server. It can run a program on a server and return the results to the browser. The earliest method to achieve this was the common gateway interface (CGI), which provides a simple way for a browser to execute a program on a remote computer. The CGI programs are often called **CGI scripts**. CGI is the mechanism that most web search programs use to send queries from a browser to the search system. Publishers store their collections in databases and use CGI scripts to provide user access. Consider the following URL:

  http://www.dlib.org/cgi-bin/seek?author='Arms'

  An informal interpretation of this URL is, "On the computer with domain name "www.dlib.org", execute the program in the file "cgi-bin/seek", pass it the parameter string "author='Arms'", and return the output." The program might search a database for records having the word "Arms" in the author field.

  The earliest uses of CGI were to connect browsers to older databases and other information. By a strange twist, now that the web has become a mature system, the roles have been reversed. People who develop advanced digital libraries often use CGI as a method to link the old system (the web) to their newer systems.

Mobile code

Browsers rapidly became big business, with new features added continually. Some of these features are clearly good for the user, while others are marketing features. The improvements include performance enhancements, elaboration to HTML, built-in support for other formats (an early addition was JPEG images), and better ways to add new formats. Two changes are more than incremental improvements. The first is the steady addition of security features. The other is mobile code which permits servers to send computer programs to the client, to be executed by the browser on the user's computer.

Mobile code gives the designer of a web site the ability to create web pages that incorporate computer programs. Panel 8.2 describes one approach, using small programs, called applets, written in the Java programming language. An applet is a small program that can be copied from a web site to a client program and executed on the client. Because Java is a full programming language, it can be used for complex operations. An example might be an authentication form sent from the web site to the browser; the user can type in an ID and password, which a Java applet encrypts and sends securely back to the server.

Panel 8.2
Java

Java is a general purpose programming language that was explicitly designed for creating distributed systems, especially user interfaces, in a networked environment.

If user interface software is required to run on several types of computer, the conventional approach has been to write a different version for each type of computer. Thus browsers, such as Netscape Navigator, and electronic mail programs, such as Eudora, have versions for different computers. Even if the versions are written in a standard language, such as C, the differences between the Microsoft Windows, or Unix, or Macintosh force the creator of the program to write several versions and modify them continually as the operating systems change. A user who wants to run a new user interface must first find a version of the user interface for the specific type of computer. This must be loaded onto the user's computer and installed. At best, this is an awkward and time-consuming task. At worst, the new program will disrupt the operation of some existing program or will introduce viruses onto the user's personal computer.

Computer programs are written in a high-level language, known as the source code, that is easily understood by the programmer. The usual process is then to compile the program into the machine language of the specific computer. A Java compiler is different. Rather than create machine code for a specific computer system, it transforms the Java source into an intermediate code, known as Java bytecode, that is targeted for a software environment called a Java Virtual Machine. To run the bytecode on a specific computer a second step takes place, to interpret each statement in the bytecode to machine code instructions for that computer as it is executed. Modern browsers support the Java Virtual Machine and incorporate Java interpreters.

A Java **applet** is a short computer program. It is compiled into a file of Java bytecode and can be delivered across the network to a browser, usually by executing an HTTP command. The browser recognizes the file as an applet and invokes the Java interpreter to execute it.

Since Java is a fully featured programming language, almost any computing procedure can be incorporated within a web application. The Java system also provides programmers with a set of tools that can be incorporated into programs. They include: basic programming constructs, such as strings, numbers, input and output, and data structures; the conventions used to build applets; networking services, such as URLs, TCP sockets, and IP addresses; help for writing programs that can be tailored for scripts and languages other than English; security, including electronic signatures, public/private key management, access control, and certificates; software components, known as JavaBeans, which can plug into other software architectures; and connections to databases. These tools help programmers. In addition, since the basic functions are a permanent part of the web browser, they do not have to be delivered across the Internet with every applet and can be written in the machine code of the individual computer, thus executing faster than the interpreted bytecode.

When a new computing idea first reaches the market, separating the marketing hype from the substance is often hard. Rarely, however, has any new product been surrounded by as much huff and puff as the Java programming language. Java has much to offer to the digital library community, but it is not perfect. Some of the defects are conventional technical issues. It is a large and complex programming language, which is difficult to learn. Interpreted languages always execute slowly and Java is no exception. Design decisions that prevent Java applets from bringing viruses across the network and infecting the user's computer also constrain legitimate programs. However, Java's biggest problems is non-technical. Java was developed by Sun Microsystems which set out to develop a standard language that would be the same in all versions. Unfortunately, other companies, notably Microsoft, have created incompatible variants.

Java is not the only way to provide mobile code. An alternative is for an HTML page to include a script of instructions, usually written in the language known as JavaScript. JavaScript is simpler to write than Java and executes quickly. A typical use of JavaScript is to check data that a user provides as input, when it is typed, without the delays of transmitting everything back to the server for validation. Do not be confused by the names Java and JavaScript. The two are completely different languages. The similarity of names is purely a marketing device. Java has received most of the publicity, but both have advantages and both are widely used.

Recent advances in the design of user interfaces

The design of user interfaces for digital libraries is part art and part science. Figure 8.1 provides a systematic framework for developing a design, but ultimately the success of an interface depends upon the designers' instincts and their understanding of users. Each part of the figure is the subject of research and new concepts are steadily being introduced to the repertoire. This section looks at some of the new ideas and topics of research. In addition, Chapter 12 describes recent research into structural metadata. This is a topic of great importance to user interfaces, since the manner in which digital objects are modeled and the structural metadata associated with them provide the raw material on which user interfaces act.

Conceptual models

Several research groups have been looking for conceptual models that help users navigate through the vast collections now available online. There are few landmarks on the Internet, few maps and signposts. Using hyperlinks, which are the heart of the web, the user is led to unexpected places and can easily get lost. Users of digital libraries often work by themselves, with little formal training, and nobody to turn to for help. This argues for interfaces to be based on conceptual models that guide users along well-established paths, although, with a little ingenuity, some people are remarkably adept at finding information on the Internet. Observations suggest that experienced people meet a high percentage of their library needs with networked information, but less experienced users often get lost and have difficult evaluating the information that they find. Panel 8.3 describes two research projects that have explored novel conceptual models for digital libraries.

Panel 8.3
New conceptual models: DLITE and Pad++

The standard user interface on personal computers is derived from an abstraction of a desktop. Several interesting experiments in digital libraries have searched for metaphors other than the desktop. Since these were research projects, the interfaces will probably never be used in a production system, but they are important for developing concepts that can be used in other interfaces and illustrating a design process that is based on systematic analysis of user needs and expectations.

DLITE

DLITE is an experimental user interface developed by Steve Cousins, who was then at Stanford University, as part of the Digital Libraries Initiative. DLITE was created as a user interface of the Stanford InfoBus. It uses concepts from object oriented programming, with each component being implemented as a CORBA object. The InfoBus and CORBA are described in Chapter 13.

Conceptual model

The conceptual model is based on an analysis of the tasks that a user of digital libraries carries out. The following key requirements were identified:

- Digital libraries consist of heterogeneous collections that must be accessible from anywhere on the Internet.
- Results created by one service may become the input to another.
- The interface must be extensible, so that new resources can be integrated easily with existing ones
- Resources may be retained over long periods of time.
- Users must be able to collaborate with each other.

The model describes digital libraries in terms of components; the four major types are documents, queries, collections, and services. These components are represented by icons that can be manipulated directly by the user in viewing windows on the screen. For example, dragging a query onto a search service, causes the search to be carried out, thus creating a collection of results. DLITE allows end users to create task-specific interfaces by assembling interface elements on the screen.

Functional design

The functional design of DLITE was motivated by two considerations: the ability to add new services with minimal effort, and rapid response for the user. For these reasons, the functional design is divided into two sections, known as the user interface clients and the user interface server. The clients carries out the manipulations of the four types of components. Several can run at the same time. The server provides the interface to the external services and can operate even when the clients are shut down.

Extensibility is provided by the server. When a new service is added, a new interface will need to be programmed in the server, but no modification is needed in existing clients. Support for a variety of computers and operating systems is provided by having separate client programs for each.

Pad++

Pad++ is a user interface concept that was conceived by Ken Perlin at New York University and has been developed by researchers from several universities and research centers. Its fundamental concept is that a large collection of information can be viewed at many different scales. The system takes the metaphor of the desktop far beyond the confines of a computer display, as though a collection of documents were spread out on an enormous wall.

User interactions are based on the familiar ideas of pan and zoom. A user can zoom out and see the whole collection but with little detail, zoom in part way to see sections of the collection, or zoom in to see every detail. This spatial approach makes extensive use of research in human perception. Since people have good spatial memory, the system emphasizes shape and position as clues to help people explore information and to recall later what they found.

When Pad++ zooms out, details do not grow ever smaller and smaller. This would create a fussy image for the user and the program would be forced to render huge numbers of infinitesimal details. Features have thresholds. Below a certain scale, they are merged into other features or are not displayed at all. Using a process known as "semantic zooming", objects change their appearance when they change size, so as to be most meaningful. This approach is familiar from map making. A map of a large area does not show individual buildings, but has a notation to represent urban areas.

Pad++ is not intended as a universal interface for all applications. For some applications, such as exploring large hierarchical collections in digital libraries, Pad++ may be the complete user interface. At other times it may be a visualization component, alongside other user interface concepts, within a conventional windowing system.

Computer systems

Pad++ provides an interesting example of user interface research into the relationship between system performance and usability. Panning and zooming are computationally complex operations; slow or erratic performance could easily infuriate the user. Several versions of the Pad++ concept have been developed, both freestanding and as part of web browsers. Each contains internal performance monitoring. The rendering operations are timed so that the frame refresh rate remains constant during pans and zooms. When the interface starts to be slow, medium-sized features are rendered approximately; the details are added when the system is idle.

Interface design

Interface design is partly an art, but a number of general principles have emerged from recent research. Consistency is important to users, in appearance, controls, and function. Users need feedback; they need to understand what the computer system is doing and why they see certain results. They should be able to interrupt or reverse actions. Error handling should be simple and easy to comprehend. Skilled users should be offered shortcuts, while beginners have simple, well-defined options. Above all the user should feel in control.

Control creates a continuing tension between the designers of digital libraries and the users, particularly control over graphic design and the appearance of materials. Many designers want the user to see materials exactly as they were designed. They want to control graphical quality, typography, window size, location of information within a window, and everything that is important in good design. Unfortunately for the designer, browsers are generic tools. The designer does not know which browser the user has, what type of computer, how fast the network, or whether the display is large or small. Users may wish to reconfigure their computer. They may prefer a large font, or a small window, they may turn off the display of images to reduce network delays. Therefore, good designs must be effective in a range of computing environments. The best designers have an knack of building interfaces that are convenient to use and attractive on a variety of computers, but some designers find difficulty in making the transition from traditional media, where they control everything, to digital libraries and the web. A common mistake is over-elaboration, so that an interface is almost unusable without a fast network and a high-performance computer.

One unnecessary problem in designing interfaces is that Netscape and Microsoft, the two leading vendors of browsers, deliver products that have deliberately chosen to be different from each other. A user interface that works beautifully on one may be a disaster on the other. The manufacturers' competitive instincts have been given priority over the convenience of the user. If a designer wishes to use some specialized features, the user must be warned that the application can not be used on all browsers.

Functional design

Digital libraries are distributed systems, with many computers working together as a team. Research into functional design provides designers with choices about what functions belong on which of the various computers and the relationships between them. Between a repository, where collections are stored, and the end user, who is typically using a web browser, lies an assortment of computer programs which are sometimes called **middleware**. The middleware programs act as intermediaries between the user and the repository. They interpret instructions from users and deliver them to repositories. They receive information from repositories, organize it, and deliver it to the user's computer. Often the middleware provides supporting services such as authentication or error checking, but its most important task is to match the services that a user requires with those provided by a repository.

Figures 8.2 and 8.3 show two common configurations. In both situations a user at a personal computer is using a standard web browser as the front end interface to digital library services, illustrated by a repository and a search system.

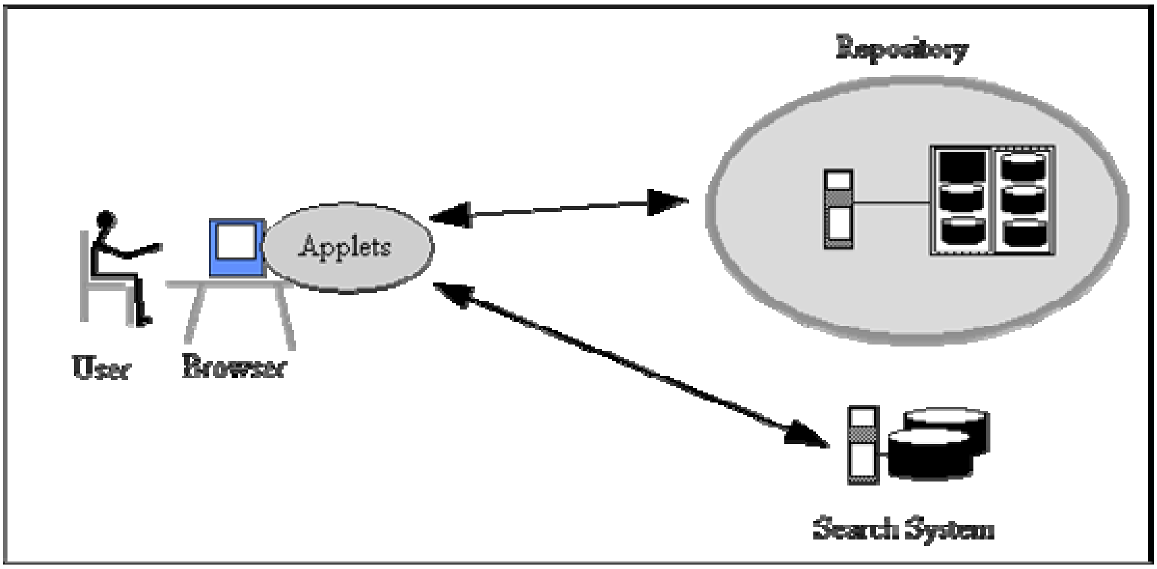


Figure 8.2. User interface with interface scripts

Figure 8.2 shows middleware implemented as independent computer programs, usually CGI scripts, that run on server computers somewhere on the network. The browser sends messages to the interface in a standard protocol, probably HTTP. The scripts can run on the remote service or they can communicate with the service using any convenient protocol. In the figure, the link between the interface scripts and the search system might use the Z39.50 protocol. The technical variations are important for the flexibility and performance of the system, but functionally they fill the same need. This configuration has the disadvantage that every action taken by the user must be transmitted across the network to be processed and the user then waits for a response.

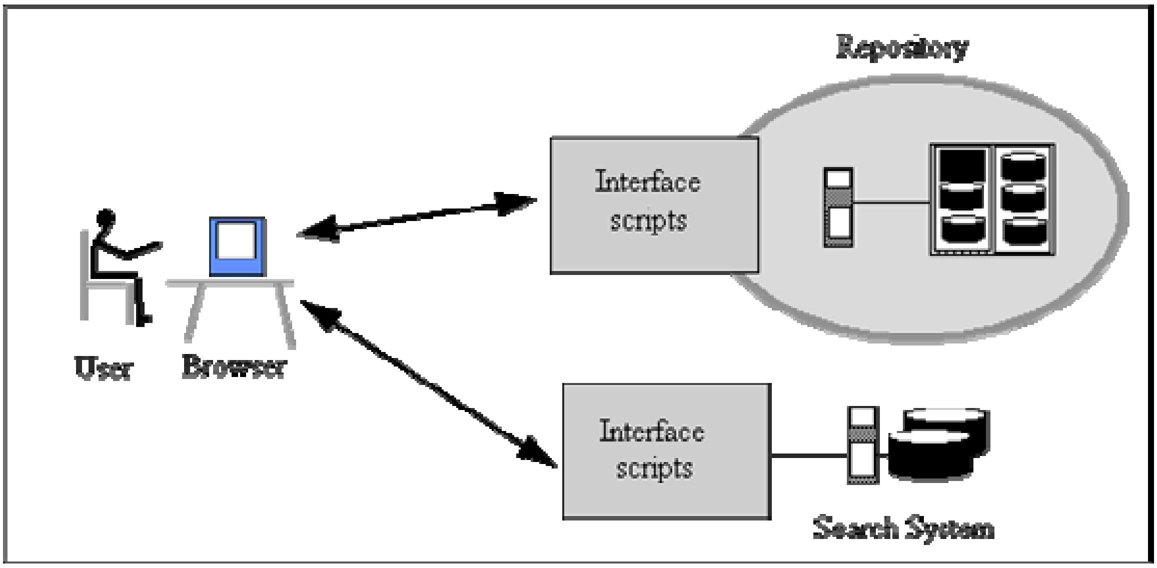


Figure 8.3. User interface with mobile code (applets)

Figure 8.3 shows a more modern configuration, where mobile code is executed on the user's computer by the browser. The code might be Java applets. Until used, the code is stored on the remote server; it is loaded into the user's computer automatically, when required. This configuration has many advantages. Since the user interface code runs on the user's computer it can be very responsive. The protocol between the user interface and the services can be any standard protocol or can be tailored specially for the application.

A **presentation profile** is an interesting concept which has recently emerged. Managers of a digital library associate guidelines with stored information. The guidelines suggest how the objects might be presented to the user. For example, the profile might recommend two ways to render an object, offering a choice of a small file size or the full detail. The user interface is encouraged to follow the profile in rendering the data, but has the option of following different approaches. An potential use of presentation profiles is to permit specialized interfaces to support people who have physical difficulties in using the standard presentation.

Computer systems and networks

The performance of computer systems and networks has considerable impact on the usability. The creator of a digital library can make few assumptions about the equipment that a user possesses, beyond the basic knowledge that every user has a personal computer attached to the Internet. However, this simple statement covers a multitude of situations. Some personal computers are more powerful than others; the quality of displays vary greatly. Some people have their own private computer, which they can configure as they wish; others may share a computer, or use more than one computer. Any digital library must assume that users will have a variety of computers, with various operating systems and environments. The environments include the various versions of Microsoft's Windows, the Macintosh, and a plethora of types of Unix. Differences between operating systems can be minimized by using a web browser for the user interface, but performance differences are not so simple.

The usability of a computer system depends upon the speed with which it responds to instructions. The designer of a digital library has little or no knowledge of the quality of network connections between the user and the library. Connections vary from spectacularly fast to frustratingly slow and even the best connections experience occasional long pauses. A user interface that is a delight to use over a 10 million bits/second local network may be unusable over an erratic, dial-up link that rarely reaches its advertised speed of 14 thousand bits/sec. Thus digital libraries have to balance effective use of advanced services, requiring fast equipment and up-to-date software, with decent service to the users who are less well provided.

Improving the responsiveness of browsers

Web browsers incorporate several tricks to improve the responsiveness seen by the user. One is internal caching. Information that has been used once is likely to be used again, for example when the user clicks the "back" button on a web browser; some graphics may be repeatedly constantly, such as logos and buttons. Web browsers retain recently used files by storing them temporarily on the disk of the user's personal computer. These files may be HTML pages, images, or files of mobile code, perhaps in JavaScript. When the user requests a file, the browser first checks to see if it can read the file from the local cache, rather than reach out across the Internet to retrieve it from a distant server.

Another family of methods that improve the responsiveness of browsers is to carry out many operations in parallel. Browsers display the first part of a long file before the whole file has arrived; images are displayed in outline with the details filled in later as more data arrives. Several separate streams of data can be requested in parallel. In aggregate, these techniques do much to mitigate the slow and erratic performance that often plagues the Internet.

Mirroring and caching

A key technique to enhance the performance of systems on the Internet is to replicate data. If a digital library collection is to be used by people around the world, duplicate copies of the collection are made at several sites, perhaps two in Europe, two in the United States, one in Australia, and one in Japan. This is called **mirroring**. Each user selects the mirror site that provides the best performance locally. Mirroring also provides back-up. If one site breaks down, or a section of the network is giving trouble, a user can turn to another site.

While mirroring is a technique to make copies of entire collections, **caching** is used to replicate specific information. A cache is any store that retains recently used information, to avoid delays the next time that it is used. Caches are found within the computer hardware to help computer processors run quickly; they are found in the controllers that read data from computer disks; they are used to improve the speed and reliability of the domain name system which converts domain names to Internet addresses. Digital libraries have caches in many places. Organizations may run local caches of documents that have been read recently; a digital library that store large collections on slow but cheap mass storage devices will have a cache of information stored on faster storage devices.

All methods of replicating data suffer from the danger that differences between the versions may occur. When information has been changed, some users may be receiving material that is out of date. Elaborate procedures are needed to discard replicated information after a stated time, or to check the source to see if changes have taken place. Caches are also vulnerable to security break-ins. A computer system is only as secure as its weakest link. A cache can easily be the weakest link.

Reliability and user interfaces

Few computer systems are completely reliable and digital libraries depend upon many subsystems scattered across the Internet. When the number of independent components is considered, it is remarkable that anything ever works. Well-designed computer systems provide the user with feedback about the progress in carrying out tasks. A simple form of feedback is to have an animation that keeps moving while the user interface is waiting for a response. This at least tells the user that something is happening. More advanced feedback is provided by an indication of the fraction of the task that has been completed with an estimate of the time to completion. In all cases, the user needs a control that allows a time consuming operation to be canceled.

The term "graceful degradation" describes methods that identify when a task is taking a long time and attempt to provide partial satisfaction to the user. A simple technique is to allow users of web browsers to turn off the images and see just the text in web pages. Several methods of delivering images allow a crude image to be displayed as soon as part of the data has been delivered, with full resolution added later.

User interfaces for multimedia information

The types of material in digital libraries are steadily becoming more varied. User interface designs for collections of textual material may be inadequate when faced with collections of music, of maps, of computer software, of images, of statistical data, and even of video games. Each is different and requires different treatment, but some general principles apply.

- **Summarization.** When a user has many items to choose from, it is convenient to represent each by some small summary information that conveys the essence of the item. With a book, the summary is usually the title and author; for a picture, a thumbnail-sized image is usually sufficient. More esoteric information can be difficult to summarize. How does a digital library summarize a computer program, or a piece of music?

- **Sorting and categorization.** Libraries organize information by classification or by ordering, but many types of information lack easily understood ways of organization. How does the library sort musical tunes? Fields guides to birds are easy to use because they divide birds into categories, such as ducks, hawks, and finches. Guides to wild flowers are more awkward, because flowers lack the equivalent divisions.

- **Presentations.** Many types of information require specialized equipment for the best presentation to the user. High-quality digitized video requires specialized displays, powerful computers, high-speed network connections, and appropriate software. A digital library should make full use of these capabilities, if they exist, but also needs to have an alternative way to present information to users with lesser equipment. This inevitably requires the digital library to offer a selection of presentations.

To design digital libraries for different types of information, it is abundantly clear that there is no single solution. User interfaces have to be tailored for the different classes of material and probably also for different categories of user. Panel 8.4 describes one good example of a user interface, the Informedia digital library of digitized segments of video.

Panel 8.4
Informedia

Informedia is a research program at Carnegie Mellon University into digital libraries of video. The leader is Howard Wactlar. The objects in the library are broadcast news and documentary programs, automatically broken into short segments of video, such as the individual items in a news broadcast. The emphasis is on automatic methods for extracting information from the video, in order to populate the library with minimal human intervention. As such it is definitely a research project, but it is a research project with a practical focus. It has a large collection of more than one thousand hours of digitized video, obtained from sources such as Cable Network News, the British Open University, and WQED television.

Chapter 12 describes some of the techniques that have been developed to index and search video segments automatically. Informedia has also proved to be a fertile source of user interface concepts.

Video skimming

Research projects often achieve results that go beyond what was expected at the start. Informedia's video skimming is an example. One of the challenges with video or audio material is to provide users with an quick overview of an item. The reader of a book can look at the contents page or flick through the pages to see the chapter headings, but there is no way to flick through a video.

Video skimming uses automatic methods to extract important words and images from the video. In combination, the selected words and images provide a video abstract that conveys the essence of the full video segment.

The user interface

The Informedia user interface uses the conceptual model of searching an index of the collection to provide a set of hits, followed by browsing through the items found. Queries can be entered by typing, or by speech which is converted to words by a speech recognition program.

After searching, the user interface presents ranked results. Each video clip is represented by an image; when the mouse is moved over the image, a text summary is provided. The image is selected automatically as representative of the video segment as it relates to the current query. The summary is created through natural language processing, forming a written representation of the video. The user can then click on an image to view the segment. The interface design is a video viewer with controls, such as "play" or "stop", that are similar to the control on a normal video player. Most users are familiar with these controls.

The Informedia interface is sufficiently intuitive that it can be used by people who have not been trained on the system. The interface maximizes feedback to the user at each step. Much of the testing has been with high school students in Pittsburgh.

User interfaces and the effectiveness of digital libraries

During the past few years, the quality of user interface design has improved dramatically. It is now assumed that new users can begin productive work without any training. Most importantly, there are now numerous examples of fine interfaces on the Internet that others can use as models and inspiration. Standards of graphical design gets better every year. For *D-Lib Magazine*, we redesigned our materials three times in as many years. Each was considered elegant in its day, but needed a face lift a year later.

Good support for users is more than a cosmetic flourish. Elegant design, appropriate functionality, and responsive systems make a measurable difference to the effectiveness of digital libraries. When a system is hard to use, the users may fail to find important results, may mis-interpret what they do find, or may give up in disgust believing that the system is unable to help them. A digital library is only as good as the interface it provides to its users.

Chapter 9
Text

The importance of text

Textual materials have a special place in all libraries, including digital libraries. While sometimes a picture may indeed be worth a thousand words, more often the best way to convey complex ideas is through words. The richness of concepts, the detail, and the precision of ideas that can be expressed in text are remarkable. In digital libraries, textual documents come from many sources. They can be created for online use. They can be converted from print or other media. They can be the digitized sound track from films or television programs. In addition, textual records have a special function as metadata to describe other material, in catalogs, abstracts, indexes, and other finding aids. This chapter looks at how textual documents are represented for storage in computers, and how they are rendered for printing and display to users. Metadata records and methods for searching textual documents are covered in later chapters.

Mark-up, page description, and style sheets

Methods for storing textual materials must represent two different aspects of a document: its structure and its appearance. The **structure** describes the division of a text into elements such as characters, words, paragraphs and headings. It identifies parts of the documents that are emphasized, material placed in tables or footnotes, and everything that relates one part to another. The structure of text stored in computers is often represented by a mark-up specification. In recent years, SGML (Standard Generalized Markup Language) has become widely accepted as a generalized system for structural mark-up.

The **appearance** is how the document looks when displayed on a screen or printed on paper. The appearance is closely related to the choice of format: the size of font, margins and line spacing, how headings are represented, the location of figures, and the display of mathematics or other specialized notation. In a printed book, decisions about the appearance extend to the choice of paper and the type of binding. **Page-description languages** are used to store and render documents in a way that precisely describe their appearance. This chapter looks at three, rather different, approaches to page description: TeX, PostScript, and PDF.

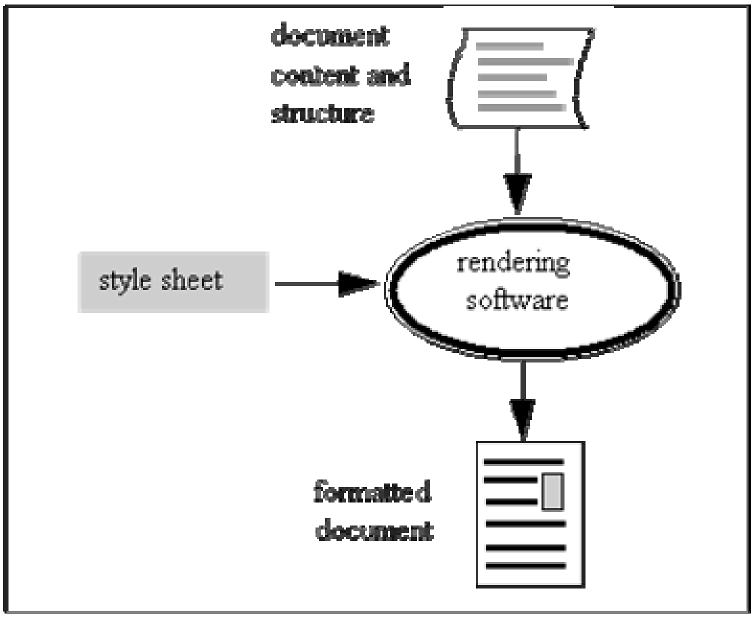


Figure 9.1. The relationship between structure and appearance

Structure and appearance are related by the design of the document. In conventional publishing, the designer creates a design specification, which describes how each structural element is to appear, with comprehensive rules for every situation that can arise. The specification is known as a **style sheet**. It enables a compositor to take a manuscript, which has been marked-up by a copy editor, and create a well-formatted document. Figure 9.1 is an outline of the procedure that several journal publishers use to produce electronic journals. Articles received from authors are marked up with SGML tags, which describe the structure and content of the article. A style sheet specifies how each structural element should appear. The SGML mark-up and the style sheet are input to rendering software which creates the formatted document.

In the early days of digital libraries, a common question was whether digital libraries would replace printed books. The initial discussion concentrated on questions of readability. Under what circumstances would people read from a screen rather than printed paper? With experience, people have realized that computers and books are not directly equivalent. Each has strengths that the other can not approach. Computing has the advantage that powerful searching is possible, which no manual system can provide, while the human factors of a printed book are superb. It is portable, can be annotated, can be read anywhere, can be spread out on a desk top or carried in the hand. No special equipment is needed to read it.

Since digital texts and printed materials serve different roles, publishers create printed and online versions of the same materials. Figure 9.2 shows how a mark-up language can be used to manage texts that will be both printed and displayed on computer screens. By using separate style sheets, a single document, represented by structural mark-up, can be rendered in different ways for different purpose. Thus, journal articles can be displayed on screen or printed. The layout and graphic design might be different, but they are both derived from the same source and present the same content.

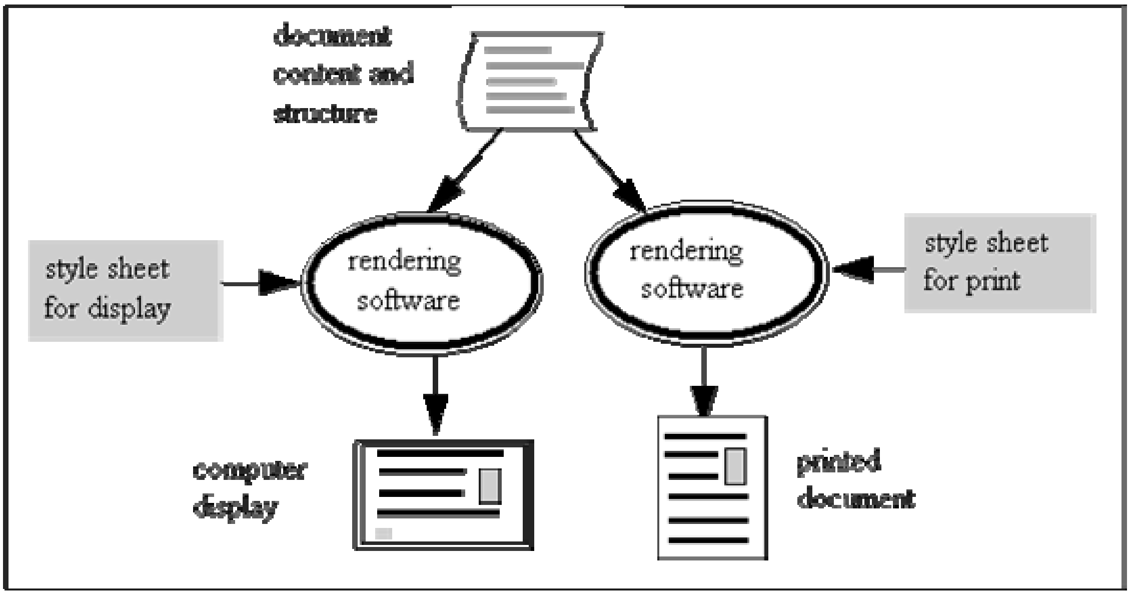


Figure 9.2. Alternative renderings of a single document

Strict control over the appearance of documents via SGML mark-up and style sheets proves to be very difficult. The ways that textual materials are organized, displayed, and interpreted have many subtleties. Mark-up languages can represent almost all structures, but the variety of structural elements that can be part of a document is huge, and the details of appearance that authors and designers could choose are equally varied. For example, many articles in journals published by the ACM contain mathematical notation. Authors often supply their articles in TeX format, which provides precise control over the layout of mathematics. During the production process it is converted to SGML mark-up. When rendered using a style sheet, the output may appear slightly different from the author's original, which can cause problems.

Panel 9.1
The Oxford English Dictionary

The new *Oxford English Dictionary* is a fine example of the use of a mark-up language to describe both structure and content, so that the same material can be used as the basis for a variety of products.

The first edition of the dictionary was created by James Murray and his colleagues over four decades. Nineteenth century photographs of Murray at work in his grandly named "Scriptorium" in Oxford show a primitive shack filled with slips of paper. The dictionary was a landmark in lexicography, but it existed only as static printed pages. Keeping it up to date by hand proved impossible.

To create a new edition of the dictionary, the first stage was to type into a database the entire text of the original edition, identifying all the typographic distinctions. The typography of the original dictionary is loaded with semantic information. Capital letters, bold and italic fonts, parentheses, smaller type sizes, and other formatting conventions convey semantics that are recorded nowhere else. A highly sophisticated computer program was written that extracted this buried semantic information and marked the textual elements with SGML tags.

The new dictionary is maintained as a computer database, in which the SGML mark-up is as important as the words of the text. Lexicographers update it continually. A wide variety of publications can be created with minimal effort, including printed books, CD-ROMs and other digital versions. This new *Oxford English Dictionary* required cooperation between the world's largest team of lexicographers, in Oxford, a team of computational linguists at Waterloo University in Ontario, and corporate support from IBM.

Converting text

Today, most documents are created digitally, beginning life on a word processor, but libraries are full of valuable documents that exist only on paper. Consequently, there is demand to convert printed documents to computer formats. For important documents, conversion projects capture the appearance and also identify the structure of the original.

The basic conversion technique is scanning. A document is scanned by sampling its image on a grid of points. Each point is represented by a brightness code. In the simplest form, only black and white are distinguished. With a resolution of 300 dots per inch, horizontally and vertically, a good image can be made of most printed pages. If the resolution is increased to 600 dots per inch, or by coding for 8 levels of gray, the clarity becomes excellent, and halftone illustrations can be represented. High-quality artwork requires at least 24 bits per dot to represent color combinations. This creates very large files. The files are compressed for convenience in storage and processing, but even simple black-on-white text files need at least 50,000 bytes to store a single page.

A scanned page reproduces the appearance of the printed page, but represents text simply as an image. In many applications, this is a poor substitute for marked-up text or even simple ASCII characters. In particular, it is not possible to search a page image for specific words. The serious scholar sometimes needs to work from the original and frequently needs to know its appearance. On other occasions, an electronic version that identifies the structure is superior; a marked-up electronic text is more convenient than an image of the original for creating a concordance or for textual analysis, since the text can be tagged to indicate its linguistic structure or its historical antecedents. Therefore the next stage of conversion is to provide an electronic text from the page image.

Optical character recognition is the technique of converting scanned images of characters to their equivalent characters. The basic technique is for a computer program to separate out the individual characters, and then to compare each character to mathematical templates. Despite decades of research, optical character recognition remains an inexact process. The error rate varies with the legibility of the original. If the original document is clear and legible, the error rate is less than one percent. With poor quality materials the error rate can be much higher. For many purposes, an error rate of even a fraction of one percent is too high. It corresponds to many incorrect characters on every page.

Various processes have been devised to get around these errors. One technique is to use several different character recognition programs on the same materials. Hopefully, characters that cause one program difficulty may be resolved by the others. Another approach is to use a dictionary to check the results. However, all high-quality conversion requires human proof reading. In some systems, a computer program displays the converted text on a screen, and highlights doubtful words with suggestions which an editor can accept or correct. One organization that has developed efficient processes of this type is UMI, which converts huge numbers of theses every year. Since most theses received by UMI are new, clean copy, UMI achieves low error rates from optical character recognition combined with manual resolution of doubtful words.

When the individual words have been recognized, the next stage of conversion is to identify the structure of a document and to tag key structural elements, such as headings. This is another aspect where, despite steady progress by researchers, high-quality conversion requires human proof reading and editing.

There is an alternative method of conversion that is widely used in practice: to retype the document from scratch and add mark-up tags manually. This approach is often cheaper than a combination of automatic and human processing. Since the work is labor intensive, it is usually carried out in countries where labor costs are low. One of the largest conversion projects is the Library of Congress's American Memory program. The documents to be converted are selected from the library's historic collections and are often less clear than a recently printed document. The conversion is carried out by contractors who guarantee a specified level of accuracy, but are at liberty to carry out the conversion by any convenient method. All the early contractors decided that the most economical method was to convert the documents by retyping them.

Encoding characters

ASCII

When representing text in a computer, the most basic elements are characters. It is important to distinguish between the concept of a character as a structural element and the various representations of that character, stored within a computer or displayed for reading. A character, such as the capital letter "A", is an abstract concept, which is independent of the encoding used for storage in a computer or the format used to display it.

Computers store a character, such as "A" or "5", as a sequence of bits, in which each distinct character is encoded as a different sequence. Early computers had codes for the twenty six letters of English (upper and lower case), the ten digits, and a small number of punctuation marks and special symbols. The internal storage representation within most computers is still derived from this limited character set. Most modern computers represent characters using the ASCII code. (ASCII stands for American Standard Code for Information Interchange, but the full form of the name is seldom used.)

Originally, ASCII represented each character by seven bits. This 7-bit encoding is known as standard ASCII. For example, the character "A" is encoded as the 7-bit sequence 1000001. Considered as a binary number, this sequence is the number 65. Hence, it is conventional to state that, in the standard ASCII code, the number 65 represents the character "A". There are 128 different pattern that can be made with seven bits. Standard ASCII associates a specific character with each number between 0 and 127. Of these, the characters 0 to 31 represent control characters, e.g., "carriage return". Table 9.1 shows the ASCII codes 32 to 127, known as the printable ASCII character set. (The space character is considered to be a printable character.)

Printable ASCII

32	Space	64	@	96	`
33	!	65	A	97	a
34	"	66	B	98	b
35	#	67	C	99	c
36	$	68	D	100	d
37	%	69	E	101	e
38	&	70	F	102	f
39	'	71	G	103	g
40	(	72	H	104	h
41	)	73	I	105	i
42	*	74	J	106	j
43	+	75	K	107	k
44	,	76	L	108	l
45	-	77	M	109	m

46	.	78	N	110	n
47	/	79	O	111	o
48	0	80	P	112	p
49	1	81	Q	113	q
50	2	82	R	114	r
51	3	83	S	115	s
52	4	84	T	116	t
53	5	85	U	117	u
54	6	86	V	118	v
55	7	87	W	119	w
56	8	88	X	120	x
57	9	89	Y	121	y
58	:	90	Z	122	z
59	;	91	[	123	{
60	<	92	\	124	|

61	=	93	]	125	}
62	>	94	^	126	~
63	?	95	_	127	□

Table 9.1. The printable character set from 7-bit ASCII

The printable ASCII character set is truly a standard. The same codes are used in a very wide range of computers and applications. Therefore, the ninety six printable ASCII characters are used in applications where interoperability has high priority. They are the only characters allowed in HTML and in many electronic mail systems. Almost every computer keyboard, display, and software program interprets these codes in the same way. There is also an extended version of ASCII that uses eight bits. It provides additional character encodings for the numbers 128 to 255, but it is not as universally accepted as 7-bit ASCII.

Unicode

Textual materials use a much wider range of characters than the printable ASCII set, with its basis in the English language. Other languages have scripts that require different character sets. Some European languages have additional letters or use diacritics. Even Old English requires extra characters. Other languages, such as Greek or Russian, have different alphabets. The Chinese, Japanese, and Korean writing does not use an alphabet. These languages use the Han characters, which represent complete words or syllables with a single character. Even for texts written in current English, the printable ASCII characters are inadequate. Disciplines such as mathematics, music, or chemistry use highly refined notation that require large numbers of characters. In each of these fields, comprehension often depends critically on the use of accepted conventions of notation.

The computer industry sells its products world-wide and recognizes the need to support the characters used by their customers around the world. This is an area in which the much-maligned Microsoft Corporation has been a leader. Since it is impossible to represent all languages using the 256 possibilities represented by an eight-bit byte, there have been several attempts to represent a greater range of character sets using a larger number of bits. Recently, one of these approaches has emerged as the standard that most computer manufacturers and software houses are supporting. It is called **Unicode**.

In strict Unicode, each character is represented by sixteen bits, allowing for a up to 65,536 distinct characters. Through the painstaking efforts of a number of dedicated specialists, the scripts used in a wide range of languages can now be represented in Unicode. Panel 9.3 lists the scripts that were completed by late 1997.

Panel 9.3
Scripts represented in Unicode

Version 2.0 of the Unicode standard contains 16-bit codes for 38,885 distinct characters. The encoding is organized by scripts rather than languages. Where several languages use a closely related set of characters, the set of symbols that covers the group of languages is identified as a single script. The Latin script contains all the characters used by English, French, Spanish, German, and related languages. Unicode 2.0 supports the following scripts:

Arabic Armenian Bengali

Bopomofo Cyrillic Devanagari

Georgian Greek Gujarati

Gurmkhi Han Hangul

Hebrew Hiragana Kannada

Katakana Latin Lao

Malayalam Oriya Phonetic

Tamil Telugu Thai

Tibetan

In addition to the above primary scripts, a number of other collections of symbols are also encoded by Unicode. They include the following:

Numbers

General Diacritics

General Punctuation

General Symbols

Mathematical Symbols

Technical Symbols

Dingbats

Arrows, Blocks, Box Drawing Forms, & Geometric Shapes

Miscellaneous Symbols

Presentation Forms

Unicode does not cover everything. Several modern languages, such as Ethiopic and Sinhala, will eventually be included. There are also numerous archaic scripts, such as Aramaic, Etruscan, and runes, which will probably be included at some date.

One important aspect of Unicode is support for the Han characters used by Chinese, Japanese, and Korean. Unicode supports the Unihan database, which is the result of an earlier project to reconcile the encoding systems previously used for these languages.

The acceptance of Unicode is not simply through the efforts of linguistic scholars in supporting a wide range of languages. The developers have thought carefully about the relationship between Unicode and existing software. If every computer program had to be changed, Unicode would never be adopted. Therefore, there is a special representation of the Unicode characters, known as **UTF-8**, that allows gradual transformation of ASCII-based applications to the full range of Unicode scripts.

UTF-8 is an encoding that uses from one to six bytes to represent each Unicode character. The most commonly used characters are represented by a single byte, the next most common by two bytes, the least common by six bytes. The crucial part of the design is that each printable ASCII character is represented by a single byte, which is identical to the corresponding ASCII character. Thus the same sequence of bytes can be interpreted as either Unicode characters (in UTF-8 representation) or as printable ASCII. For example, a page of HTML text, which has been created using printable ASCII characters, requires no modification to be used with a program that expects its data to be in UTF-8 encoding.

Transliteration

Unicode is not the only method used to represent a wide range of characters in computers. One approach is transliteration. This is a systematic way to convert characters in one alphabet into another set of characters. For example, the German ö is sometimes transliterated oe. A phonetic system of transliteration known as pinyin is frequently used to represent Chinese, especially Mandarin, in the English alphabet. Transliteration may have been acceptable in the days when typewriters were manual devices with a physically constrained character set. With today's computers, transliteration should not be needed and hopefully will soon become ancient history.

Libraries have a more immediate problem. They were using a wide range of alphabets long before the computing industry paid attention to the problem. In fact, libraries were well-advanced in this area in the days when most computers supported only upper case letters. As a result, MARC catalogs and other library systems contain huge volumes of material that are encoded in systems other than Unicode, including pinyin, and face a daunting task of conversion or coexistence.

SGML

Mark-up languages have been used since the early days of computing to describe the structure of the text and the format with which the text is to be displayed. Today, the most commonly used are the SGML family of languages.

SGML is not a single mark-up language. It is a system to define mark-up specifications. An individual specification defined within the SGML framework is called a **document type definition** (DTD). Many publishers and other agencies have developed their own private DTDs. Examples of DTDs that are of particular importance to scholarly libraries are the Text Encoding Initiative and the Encoded Archival Description, which are described in Panel 9.4.

Panel 9.4
Document type definitions (DTDs) for scholarship

Text Encoding Initiative

The Text Encoding Initiative (TEI) was an early and very thorough effort to represent existing texts in digital form, with an emphasis on use by humanities scholars. Documents from the past include almost every approach that has ever been used to represent text on paper, velum, papyrus, or even stone. They include print, handwriting, typewriting, and more. The documents include annotation, deletions, and damaged portions that can not be read. Every character set the world has ever known may be present.

SGML has proved effective in describing these materials, but designing the DTD provided a challenge. It would have been easy to have created a DTD that was so complex as to be completed unwieldy. The solution has been to create a family of DTDs, built up from components. A core tag set defines elements likely to be needed by all documents and therefore is part of all DTDs. Each DTD is required to add a base set of tags, selected from a choice of tags for prose, verse, drama, dictionaries, or data files. Usually, only one base tag set is appropriate for a given document. Finally, a variety of additional tag sets are available for specialized purposes. (The authors call this the Chicago Pizza model: every pizza has cheese and tomato sauce, one base, and optional extra toppings.)

The Encoded Archival Description (EAD)

The term **finding aid** covers a wide range of lists, inventories, indexes, and other textual documents created by archives, libraries, and museums to describe their holdings. Some finding aids provide fuller information than is normally contained within cataloging records; others are less specific and do not necessarily have detailed records for every item in a collection. Some are short; others run to hundreds of pages.

The Encoded Archival Description (EAD) is a DTD used to encode electronic versions of archival aids. The first version of the DTD was developed by a team from the University of California at Berkeley. The effort drew on experience from the Text Encoding Initiative. It reflects the conventions and practices established by archivists, and uses the heavily structured nature of finding aids. Much of the information is derived from hierarchical relationships and there are many other interrelationships that must be explicitly recognized when encoding a finding aid for use in a digital library.

The EAD has been widely embraced by archivists, who have worked collaboratively to refine the DTD, test it against existing finding aids, and provide extensive documentation. The EAD has become a highly specialized tool, tailored for the needs of a specialized community. For use by that community, it is an important tool that allows them to exchange and share information.

A DTD is built up from the general concepts of entities and elements. A DTD defines what entities and elements are allowable in a particular class of documents and declares the base character set encoding used for the document. **Entities** are specified by identifiers that begin with the letter "&" and end with ";". Here are some examples:

&alpha;
&logo;

In a typical DTD, the allowable entities include most of the ASCII character set and other characters, known as **character entities**, but any symbol or group of symbols can be defined as a entity. The name of the entity is simply a name. In standard character sets, "&alpha;", is the entity used to encode the first letter of the Greek alphabet, but a DTD could use this code for some totally different purpose. The DTDs used by scientific publishers define as many as 4,000 separate entities to represent all the special symbols and the variants used in different scientific disciplines.

Entities provide a stream of symbols that can be grouped together into **elements**. A DTD can define any string as the name of an element. The element is bracketed by two tags in angle brackets, with "/" used to denote the end tag. Thus the Text Encoding Initiative uses the tags <del> and </del> to bracket text that has been crossed-out in a manuscript. To mark the words "men and women" as crossed-out, they would be tagged as:

```
<del>men and women</del>
```

Examples of elements include the various types of headings, footnotes, expressions, references, and so on. Elements can be nested in hierarchical relationships. Each DTD has a grammar that specifies the allowable relationships as a set of rules that can be processed by a computer program.

Simplified versions of SGML

SGML is firmly established as a flexible approach for recording and storing high-quality texts. Its flexibility permits creators of textual materials to generate DTDs that are tailored to their particular needs. Panels 9.1 and 9.3 describe some examples. Publishers of scientific journals have developed their own DTDs, which they use in-house to mark-up journal articles as they are created and to store them on computers. Digital library projects, such as JSTOR and American Memory, use simple DTDs that are derived from the work of the Text Encoding Initiative.

The disadvantage of SGML's flexibility is the complexity of the software needed to process it. While a program to parse and render a simple DTD is not hard to write, it is a forbidding task to create a general-purpose package that can parse any DTD, combine information from any style sheet and render the document either on a computer screen or printer. The market for such software is quite small. Only one company has persevered and created a general package for rendering SGML. Even this package is not perfect; it does not implement all of SGML, runs only on some types of computers, and uses its own private form of style sheets. Hence, full SGML is unsuitable for use in digital libraries that emphasize interoperability with other systems.

HTML

The web has stimulated the development of simplified versions of SGML. HTML, the mark-up language used by the web, can be considered an unorthodox DTD. In many ways, however, it diverges from the philosophy of SGML, because it mixes structural information with format. HTML may have begun life as structural mark-up, relying on browsers to determine how to format the text for display, but its subsequent development has added a large number of features that give the designer of web pages control over how their material appears when rendered for screen display or printing. Panel 9.5 illustrates how far HTML has diverged from purely structural mark-up.

Panel 9.5
Features of HTML

HTML provides a variety of tags that are inserted into documents. Usually tags are in pairs; e.g., the pair of tags below indicates that the enclosed text is a main heading.

<h1>This is a main heading</h1>

A few tags are self-contained and do not bracket any text. An example is <hr>, which indicates a horizontal rule. Elements are often embedded or nested within each other. For example, a list can contain many paragraphs and tables can contain other tables.

The following examples are typical of the features provided by HTML. There are many more and several of these features have a wide selection of options. They illustrate how HTML combines structural mark-up with formatting and support for online applications.

Structural elements

Many of the tags in HTML are used to describe the structural elements of a document. They include the following.

<body>	the body of the document
<p>	a paragraph
<h1>, <h2>, ..., <h6>	headings (six standard levels)
<em>	emphasis
<ul>, <ol>,< dl>	lists (unordered, ordered, & definition)
<table>	a table

Formatting

Other tags in HTML define the appearance of the document when it is rendered for display on a computer or printed. They include the following.

<br>	line break
<i>	italic
<font>	details about the font to be used
<center>	center text
<pre>	pre-formatted text

Online services

A final group of HTML tags are designed for online applications. They include the following.

<a>	anchor (used with hyperlinks)
<img>	location of an image
<form>	an online form

HTML has grown from its simple beginnings and is continuing to grow. Depending on the viewpoint, these additions can be described as making HTML more powerful, or as adding complexity and inconvenience. Nobody would argue the great value of some additions, such as being able to embed images, through the <img> tag, which was introduced in Mosaic. The value of other additions is a matter of opinion. Simple formatting commands, such as <center> introduced by Netscape, can do little harm, but other features have added a great deal of complexity. Perhaps the most notable are the use of tables and frames. These two additions, more than any other, have changed HTML from a simple mark-up language. No longer can an author learn HTML in a morning or a programmer write a program to render HTML in a week.

The tension between structural mark-up in HTML and formatting to control appearance has become a serious problem. As discussed in the last chapter, many creators of web pages want to control what the user sees. They have difficulty accepting that they do not determine the appearance of a document, as seen by the user, and use every opportunity that HTML provides to control the appearance. An unfortunate trend is for designers to use structural elements to manipulate the design; many web pages are laid out as a single huge table, so that the designer can control the margins seen by the user. Skilled designers construct elegant web pages using such tricks. Less skilled designers impose on the users pages that are awkward to use. The worst abuses occur when a designer imposes a page layout that will not fit in the window size that the user chooses or over-rides a user preference, perhaps preventing a user with poor eyesight from using a large type size.

Almost all new features that have been added to HTML have come from the developers of browsers adding features, to enhance their products or to keep pace with their competitors. Some were indeed improvements, but others were unnecessary. The World Wide Web Consortium and the Internet Engineering Task Force provide valuable coordination and attempt to provide standards, but the fundamental control of HTML is exercised by the two principal browser manufacturers, Netscape and Microsoft.

XML

XML is a variant of SGML that attempts to bridge the gap between the simplicity of HTML and the power of full SGML. Simplicity is the key to the success of HTML, but simplicity is also its weakness. Every time a new feature is added to HTML it becomes less elegant, harder to use, and less of a standard shared by all browsers. SGML is the opposite. It is so flexible that almost any text description is possible, but the flexibility comes at the cost of complexity. Even after many years, only a few specialists are really comfortable with SGML and general-purpose software is still scarce.

XML is a subset of SGML that has been designed explicitly for use with the web. The design is based on two important criteria. The first is that it is simple to write computer programs that manipulate XML. The second is that it builds on the familiar world of HTML, so that people and systems can migrate to XML with a minimum of pain.

The underlying character set for XML is 16-bit Unicode, and in particular the UTF-8 stream encoding. This permits documents to be written in standard ASCII, but supports a wide range of languages and character sets. For convenience, some character entities are pre-defined, such as &lt; and &gt; for the less-than and greater-than symbols. XML does not specify particular methods for representing mathematics, but a there is a separate effort, MathML, tackling the problems of mathematics.

Standard HTML is acceptable as XML with minimal modifications. One modification is that end tags are always needed. For example, in HTML, the tags <p> and </p> delimit the beginning and end of a paragraph, but the </p> is optional, when followed by another paragraph. In XML the end tag is always needed. The other major modification concerns HTML tags that do not delimit any content. For example, the tag <br> indicates a line break. Standard HTML does not use an </br> tag. With XML, a line break is tagged either with the pair <br></br> or with the special shortened tag <br/>.

Since XML is a subset of SGML, every document is based on a DTD, but the DTD does not have to be specified explicitly. If the file contains previously undefined pairs of tags, which delimit some section of a document, the parser automatically adds them to the DTD.

The developers of XML have worked hard to gain wide acceptance for their work. Their strategy follows the philosophy of the Internet mainstream. The design has been an open process hosted by the World Wide Web Consortium. From the start, members of the design team wrote demonstration software that they have distributed freely to interested parties. Leading corporations, notably Microsoft and Netscape, have lent their support. This deliberate building of consensus and support seems to have been successful, and XML looks likely to become widely adopted.

Style sheets

Mark-up languages describe the structure of a document. SGML and XML use tags to describe the semantics of a document and its component parts. It does not describe the appearance of a document. Thus, SGML tags might be used to identify a section of text as a chapter heading, but would not indicate that a chapter heading starts a new page and is printed with a specific font and alignment. A common need is to take a document that has SGML or XML mark-up and render it for according to a specific design specification. For example, a publisher who creates journal articles according to a DTD may wish to render them in two different ways: printed output for conventional publication, and a screen format for delivery over the Internet and display on a computer screen. The first will be rendered in a format that is input to a typesetting machine. The second will be rendered in one of the formats that are supported by web browsers, usually HTML or PDF.

The process requires that the structural tags in the mark-up be translated into formats that can be displayed either in print or on the screen. This uses a style sheet. For example, a DTD may define an element as a heading of level two, denoted by <h2> and </h2> tags. The style sheet may state that this should be displayed as 13 points Times Roman font, bold, left aligned. It will also specify important characteristics such as the appropriate line spacing and how to treat a heading that falls near the bottom of the page. A style sheet provides detailed instructions for how to render every conceivable valid document that has been marked-up according to a specific DTD.

In creating style sheets, the power of SGML is a disadvantage. Readers are accustomed to beautifully designed books. Much of this beauty comes from the skill of the craftsmen who carry out the composition and page layout. Much of their work is art. The human eye is very sensitive; the details of how statistical tables are formatted, or the layout and pagination of an art book have never been reduced to mechanical rules. Style sheets can easily be very complex, yet still fail to be satisfactory when rendering complex documents.

Since SGML is a general framework, people have been working on the general problem of specifying style sheets for any DTD. This work is called Document Style Semantics and Specification Language (DSSSL). The developers of DSSSL are faced with a forbidding task. To date, DSSSL rendering programs have been written for some simple DTDs, but the general task appears too ambitious. It is much easier to create a satisfactory style sheet for a single DTD, used in a well-understood context. Many books and journals are printed from SGML mark-up, with special-purpose style sheets, and the results can be excellent.

With HTML, there has been no formal concept of style sheets. The mark-up combines structural elements, such as various types of list, with formatting tags, such as those that specify bold or italic fonts. The tags provide general guidance about appearance, which individual browsers interpret and adapt to the computer display in use. The appearance that the user sees comes from a combination of the mark-up provided by the designer of a web page, the formatting conventions built into a browser, and options chosen by the user. Authors of HTML documents wishing for greater control can embed various forms of scripts, applets, and plug-ins.

Panel 9.6 describes **CSS** (Cascading Style Sheets) and **XLS** (Extensible Style Language) (XSL), methods for providing style sheets for HTML and XML. The developers of XML have realized that the greatest challenge to be faced by XML before it becomes widely accepted is how to control the appearance of documents and have been supportive of both CSS and XSL. It is still too early to know what will succeed, but the combination is promising. These methods have the important concept of laying down precise rules for the action to take when the styles specified by the designer and the user disagree.

Panel 9.6
Cascading Style Sheets (CSS) and Extensible Style Language (XSL)

Mark-up languages describe the structural elements of a document. Style sheets specify how the elements appear when rendered on a computer display or printed on paper. Cascading Style Sheets (CSS) were developed for use with HTML mark-up. The Extensible Style Language (XSL) is an extension for XML mark-up. Just as

XML is a simplified version of full SGML that provides a simple conversion from HTML, XSL is derived from DSSSL and any CSS style sheet can be converted to XSL by a purely mechanical process. The original hope was that XSL would be a subset of DSSSL, but there are divergences. Currently, XSL is only a specification, but there is every hope that, as XML becomes widely adopted, XSL will become equally important.

Rules

In CSS a rule defines styles to be applied to selected elements of a document. Here is a simple rule:

```
h1 {color: blue}
```

This rule states that for elements "h1", which is the HTML tag for top-level headings, the property "color" should have the value "blue". More formally, each rule consists of a selector, which selects certain elements of the document, and a declaration, enclosed in braces, which states a style to be applied to those elements. The declaration has two parts, a property and a value, separated by a colon.

A CSS style sheet is a list of rules. Various conventions are provided to simplify the writing of rule. For example, the following rule specifies that headings h1 and h2 are to be displayed in blue, using a sans-serif font.

```
h1, h2 {font-family: sans-serif; color: blue}
```

Inheritance

The mark-up for an HTML document defines a structure which can be represented as a hierarchy. Thus headings, paragraphs, and lists are elements of the HTML body; list items are elements within lists; lists can be nested within each other. The rules in CSS style sheets also inherit from each other. If no rule explicitly selects an element, it inherits the rules for the elements higher in the hierarchy. For example, consider the pair of rules:

```
body {font-family: serif}
h1, h2 {font-family: sans-serif}
```

Headings h1 and h2 are elements of the HTML body, but have an explicit rule; they will be displayed with a sans-serif font. In this example, there is no explicit rule for paragraphs or lists. Therefore they inherit the styles that apply to body, which is higher up the hierarchy, and will be displayed with a serif typeface.

Cascading

Style sheets must be associated with an HTML page. The designer has several options, including embedding the style at the head of a page, or providing a link to an external file that contains the style sheet. Every browser has its own style sheet, which may be modified by the user, and a user may have a private style sheet.

Since several style sheets may apply to the same page, conflicts can occur, where rules conflict. A series of mechanisms have been developed to handle these situations, based on some simple principles. The most fundamental principle is that, when rules conflict, one is selected and the others are ignored. Rules that explicit select elements have priority over rules that are inherited. The most controversial convention is that when the designer's rule conflicts directly with the user's, the designer's has precedence. A user who wishes to over-ride this rule can mark a rule with the flag "!important", but this is awkward. However, it does permit special style sheets to be developed, for instance for users who have poor eyesight and wish to specify large font sizes.

Page-description languages

Since creators and readers both give high priority to the appearance of documents, it is natural to have methods that specify the appearance of a document directly, without going via structural mark-up. The methods differ greatly in details, but the underlying objective is the same, to render textual materials with the same graphic quality and control as the best documents printed by traditional methods. This is not easy. Few things in life are as pleasant to use as a well-printed book. Over the years, typography, page layout, paper manufacturing, and book binding have been refined to a high level of usability. Early text formatting methods were designed for printed output, but display on computer screens has become equally important. This section looks at three page-description languages: TeX, PostScript, and PDF. These three languages have different objectives and address them differently, but they are all practical, pragmatic approaches that perform well in production systems.

TeX

TeX is the earliest. It was developed by Donald Knuth and is aimed at high-quality printing, with a special emphasis on mathematics. The problems of encoding mathematics are complex. In addition to the numerous special characters, the notation relies on complex expressions that are not naturally represented as a single sequence of characters. TeX provides rules for encoding mathematics as a sequence of ASCII characters for input, storage, and manipulation by computer, with tags to indicate the format for display.

Most users make use of one of two TeX packages, plainTeX or LaTex. These packages define a set of formatting tags that cover the majority of situations that are normally encountered in typesetting. Closely allied with TeX is a system for designing fonts called Metafont. Knuth has taken great pains to produce versions of his standard font for a wide range of computer systems.

TeX remains unsurpassed for the preparation of printed mathematical papers. It is widely used by authors and journals in mathematics, physics, and related fields.

PostScript

PostScript was the first product of the Adobe corporation, which spun off from Xerox in 1984. PostScript is a programming language, to create graphical output for printing. Few people ever write programs in PostScript, but many computers have printer routines that will take text or graphics and create an equivalent PostScript program. The program can then be sent to a printer controller that executes the PostScript program and creates control sequences to drive the printer.

Explicit support for fonts is one of PostScript's strengths. Much of the early success of the Apple Macintosh computer came from the combination of bit-mapped designs on the screen with PostScript printers that provided a quality of output previously available only on very expensive computers. With both laser printing and screen display, characters are built up from small dots. Simple laser printers use 300 dots per inch and typesetting machines may have a resolution of 1,200 dots per inch or more, but most computer screens are about 75 dots per inch. The fonts that appear attractive on a computer screen are not quite the same as the ones used for printing, and the displays of text on a screen must be fitted to the coarse resolution. Usually, the operating system functions that display text on a screen are different from the

PostScript commands used for printing. Unless great care is taken, this can lead to different page breaks and other unacceptable variations.

Although PostScript is primarily a graphical output language, which had its initial impact representing computer output to laser printers, PostScript programs are used in other applications, as a format to store and exchange representations of any text or graphical output. PostScript is not ideal for this purpose, since the language has many variations and the programs make assumptions about the capabilities of the computer that they will be executed on. PDF, which built on Adobe's experience with PostScript, is a better format for storing page images in a portable format that is independent of any particular computer.

Portable Document Format (PDF)

The most popular page description language in use today is Adobe's Portable Document Format (PDF), which is described in Panel 9.7. Adobe has built on its experience with PostScript to create a powerful format and a set of tools to create, store, and display documents in it.

Panel 9.7
Portable Document Format (PDF)

Adobe's Portable Document Format (PDF) is an important format. It is also interesting as an example of how a corporation can create a standard, make it available to the world, and yet still generate good business.

PDF is a file format used to represent a document that is independent of applications and computer system. A PDF document consists of pages, each made up of text, graphics, and images, with supporting data. However, PDF pages can go beyond the static print view of a page, by supporting features that are only possible electronically, such as hyperlinks and searching.

PDF is an evolution of the PostScript programming language, which was also created by Adobe. One way to generate PDF is by diverting a stream of data that would usually go to a printer and storing it as a PDF file. Alternatively, the file can be converted from PostScript or other formats. The file can then be stored, transmitted over a network, displayed on a computer, or printed.

Technical reasons for PDF's success

Technically, PDF has many strengths. When viewed on a computer screen, most PDF documents are very legible while retaining the design characteristics of print. Except when they include bit-mapped images, the files are of moderate size. If the computer that is displaying a PDF document does not have the fonts that were used to generate it, font descriptors enable the PDF viewer to generate an artificial font that is usually close to the original.

The electronic additions are simple to use and provide many of the features that users want. They include hyperlinks, either within the document or to external URLs. The viewers provide a tool for searching for words within the text, though searching across documents is a problem. Annotations are provided. Printer support is excellent. There is even a method that creators can use to prevent users from printing or otherwise using a document in ways that are not approved.

PDF is not perfect. It has problems distinguishing between certain types of file, or working with unusual fonts, but overall it has become a important format for online documents.

Business reasons for PDF's success

After a hesitant introduction, Adobe has been shrewd in its marketing of PDF. The company provides excellent PDF viewers for almost every known type of computer; they are freely available over the Internet. Adobe makes its money from a family of products that are used by people who create PDF files, not from the users of these files.

Adobe owns the copyright in the PDF specification, but, to promote its use as an interchange format, gives broad permissions to anyone to create PDF files, to write applications that produce PDF output, and to write software that reads and processes PDF files. Potentially, competitors could create products that undercut those that Adobe sells, but the company accepts that risk.

PDF is widely used in commercial document management systems, but some digital libraries have been reluctant to use PDF. One reason is technical. PDF is best suited for representing documents that were generated from computer originals. PDF can also store bit-mapped images, and Adobe provides optical character recognition software to create PDF files, but many of PDF's advantages are lost when the used to store image files. The file sizes can be unpleasantly large and much of the flexibility that digital libraries require is lost.

In addition, some digital libraries and archives reject PDF because the format is proprietary to a single company. There is a fear that the decision to use PDF in a digital library is more vulnerable to future events than using a format blessed by one of the standards bodies. This reasoning appears misguided. PDF had its first success with corporate America, which welcomes well-supported, commercial products. The academic and scholarly community can accept that a format maintained by a corporation may be more stable in the long term than official standards that are not backed by good products and a wide user base. The definition of the format is widely published and the broad use of PDF in commercial applications guarantees that programs will be available for PDF, even if Adobe went out of business or ceased to support it.

Structure versus appearance of documents

This chapter began with a discussion of two requirements for storing documents in digital libraries: representations of structure and of appearance. They should not be seen as alternatives or competitors, but as twin needs, both of which deserve attention. In some applications a single representation serves both purpose, but a large number of digital libraries are storing two versions of each document. Textual materials are at the heart of digital libraries and electronic publishing. Authors and readers are very demanding, but the methods exist to meet their demands.

Chapter 10
Information retrieval and descriptive metadata

Information discovery

A core service provided by digital libraries is to helping user find information. This chapter is the first of two on this subject. It begins with a discussion of catalogs, indexes, and other summary information used to describe objects in a digital library; the general name for this topic is **descriptive metadata**. This is followed by a section on the methods used to search bodies of text for specific information, the subject known as **information retrieval**. Chapter 11 extends the concepts to distributed searching, how to discover information that is spread across separate and different collections, or scattered over many computer systems.

These two chapters concentrate on methods used to search for specific information, but direct searching is just one of the strategies that people use to discover information. Browsing is the general term for the unstructured exploration of a body of information; it is a popular and effective method for discovering the unexpected. Most traditional libraries arrange their collections by subject classification to help browsing. Classification schemes, such as the Dewey Decimal Classification or the Library of Congress classification, provide both subject information and a hierarchical structure that can be used to organize collections. The most widely used web information service, Yahoo, is fundamentally a classification of web resources, augmented by searching. Digital libraries with their hyperlinks lend themselves to strategies that combine searching and browsing.

Here are some of the reasons why users seek for information in digital libraries. The range of these needs illustrates why information discovery is such a complex topic, and why no one approach satisfies all users or fits all materials.

- **Comprehensive search.** The objective of a comprehensive search is to find everything on a specified topic. Serious scholarly, scientific, or legal study typically begins with a comprehensive search to discover the prior work in the field.

- **Known item.** Known items searches occur when the user is looking for a specific item, such as, "Find me the opening chapter of Moby Dick." Frequently the reference is not perfect and a search is needed to find the item, but there is a specific item and the user will recognize it when it is found.

- **Facts.** Facts are specific items of information that may be found in many sources of information. "What is the capital of Barbados?" "What is the atomic weight of mercury?" When seeking for a fact, the search is complete when the fact is found. The source of the fact must be reliable, but there may be many possible sources.

- **Introduction or overview.** General information on a topic is a common need. "How do I organize a wedding?" "What is public key encryption?" With such requests, there may be many suitable sources of information. The user wants a small selection.

- **Related information.** Once one useful item has been found, a user may wish to know about related items. In research, a common question is, "Is there any later research that builds on the work reported in this article?"

Descriptive metadata

Many methods of information discovery do not search the actual objects in the collections, but work from descriptive metadata about the objects. The metadata typically consists of a catalog or indexing record, or an abstract, one record for each object. Usually it is stored separately from the objects that it describes, but sometimes it is embedded in the objects.

Descriptive metadata is usually expressed as text, but can be used to describe information that is in formats other than text, such as images, sound recording, maps, computer programs, and other non-text materials, as well as for textual documents. A single catalog can combine records for every variety of genre, media, and format. This enables users of digital libraries to discover materials in all media by searching textual records about the materials.

Descriptive metadata is usually created by professionals. Library catalogs and scientific indexes represent huge investments by skilled people, sustained over decades or even centuries. This economic fact is crucial to understanding current trends. On one hand, it is vital to build on the investments and the expertise behind them. On the other, there is great incentive to find cheaper and faster ways to create metadata, either by automatic indexing or with computer tools that enhance human expertise.

Catalogs

Catalog records are short records that provide summary information about a library object. The word catalog is applied to records that have a consistent structure, organized according to systematic rules. An abstract is a free text record that summarizes a longer document. Other types of indexing records are less formal than a catalog record, but have more structure than a simple abstract.

Library catalogs serve many functions, not only information retrieval. Some catalogs provide comprehensive bibliographic information that can not be derived directly from the objects. This includes information about authors or the provenance of museum artifacts. For managing collections, catalogs contain administrative information, such as where items are stored, either online or on library shelves. Catalogs are usually much smaller than the collections that they represent; in conventional libraries, materials that are stored on miles of shelving are described by records that can be contained in a group of card drawers at one location or an online database. Indexes to digital libraries can be mirrored for performance and reliability .

Information in catalog records is divided into **fields** and **sub-fields** with tags that identify them. Thus, there might be a field for an author, with a sub-field for a surname. Chapter 3 introduced cataloguing using the Anglo American Cataloguing Rules and the MARC format. MARC cataloguing is used for many types of material including monographs, serials, and archives. Because of the labor required to create a detailed catalog record, materials are catalogued once, often by a national library such as the Library of Congress, and the records distributed to other libraries through utilities such as OCLC. In digital libraries the role of MARC and the related cataloguing rules is a source of debate. How far can traditional methods of cataloguing migrate to support new formats, media types, and methods of publishing? Currently, MARC cataloguing retains its importance for conventional materials; librarians have extended it to some of the newer types of object found in digital libraries, but MARC has not been adopted by organizations other than traditional libraries.

Abstracting and indexing services.

The sciences and other technical fields rely on **abstracting and indexing services** more than catalogs. Each scientific discipline has a service to help users find information in journal articles. The services include *Medline* for medicine and biology, *Chemical Abstracts* for chemistry, and *Inspec* for physics, computing, and related fields. Each service indexes the articles from a large set of journals. The record for an article includes basic bibliographic information (authors, title, date, etc.), supplemented by subject information, organized for information retrieval. The details differ, but the services have many similarities. Since abstracting and indexing services emerged at a time when computers were slower and more expensive than today, the information is structured to support simple textual searches, but the records have proved to be useful in more flexible systems.

Scientific users frequently want information on a specific subject. Because of the subtleties of language, subject searching is unreliable unless there is indexing information that describes the subject of each object. The subject information can be an abstract, keywords, subject terms, or other information. Some services ask authors to provide keywords or an abstract, but this leads to gross inconsistencies. More effective methods have a professional indexer assign subject information to each item.

An effective but expensive approach is to use a controlled vocabulary. Where several terms could be used to describe a concept, one is used exclusively. Thus the indexer has a list of approved subject terms and rules for applying them. No other terms are permitted. This is the approach used in the Library of Congress subject headings and the National Library of Medicine's MeSH headings (see Panel 10.1).

Panel 10.1
MeSH - medical subject headings

The National Library of Medicine has provided information retrieval services for medicine and related fields since the 1960s. Medicine is a huge field with complex terminology. Since the same concept may be described by scientific terms or by a variety of terms in common use, the library has developed a controlled vocabulary thesaurus, known as MeSH. The library provides MeSH subject headings for each of the 400,000 articles that it indexes every year and every book acquired by the library. These subject terms can then be used for information retrieval.

MeSH is a set of subject terms, with about 18,000 primary headings. In addition, there is a thesaurus of about 80,000 chemical terms. The terms are organized in a hierarchy. At the top are general terms, such as *anatomy*, *organisms*, and *diseases*. Going down the hierarchy, *anatomy*, for example, is divided into sixteen topics, beginning with *body regions* and *musculoskeletal system*; body regions is further divided into sections, such as *abdomen*, *axilla*, *back*; some of these are sub-divided until the bottom of the hierarchy is reached. To help the user, MeSH provides thousands of cross-references.

The success of MeSH depends upon the professional staff who maintain the thesaurus and the indexers who assign subject terms to documents. It also requires users or reference librarians who understand the field and are able to formulate queries using MeSH terms and the MeSH structures.

Controlled vocabulary requires trained indexers. It also requires skilled users, with tools to assist the users, because the terms used in a search query must be consistent with the terms assigned by the indexer. Medicine in the United States is especially fortunate in having a cadre of reference librarians who can support the users. In digital libraries, the trend is to provide users with tools that permit them to find information directly without the help of a reference librarian. A thesaurus, such as MeSH or the Art and Architecture Thesaurus (Panel 10.2), can be used to relate the terminology that a user provides to the controlled terms that have been used for indexing.

Panel 10.2
The Art and Architecture Thesaurus

The Art and Architecture Thesaurus was developed by the J. Paul Getty Trust as a controlled vocabulary for describing and retrieving information on fine art, architecture, decorative art, and material culture. It has almost 120,000 terms for objects, textual materials, images, architecture and culture from all periods and all cultures, with an emphasis on Western civilization. The thesaurus can be used by archives, museums, and libraries to describe items in their collections. It also be used to search for materials.

Serious work on the thesaurus began in the early 1980s, when the Internet was still an embryo, but the data was created in a flexible format which has allowed production of many versions, including an open-access version on the web, a printed book, and various computer formats. The Getty Trust has explicitly organized the thesaurus so that it can be used by computer programs, for information retrieval, and natural language processing.

The Art and Architecture Thesaurus is arranged into seven categories, each containing a hierarchies of terms. The categories are *associated concepts*, *physical attributes*, *styles and periods*, *agents*, *activities*, *materials*, and *objects*. A single concept is represented by a cluster of terms, one of which is established as the preferred term, or descriptor. The thesaurus provides not only the terminology for objects, but the vocabulary necessary to describe them, such as style, period, shape, color, construction, or use, and scholarly concepts, such as theories, or criticism.

The costs of developing and maintaining a large, specialized thesaurus are huge. Even in a mature field such as art and architecture terminology is changing continually, and the technical staff has to support new technology. The Getty Trust is extremely rich but, even so, developing the thesaurus was a major project spread over many years.

The Dublin Core

Since 1995, an international group, led by Stuart Weibel of OCLC, has been working to devise a set of simple metadata elements that can be applied to a wide variety of digital library materials. This is known as the **Dublin Core**. The name comes from Dublin, Ohio, the home of OCLC, where the first meeting was held. Several hundred people have participated in the Dublin Core workshops and discussed the design by electronic mail. Their spirit of cooperation is a model of how people with diverse interests can work together. They have selected fifteen elements, which are summarized in Panel 10.3.

Panel 10.3
Dublin Core elements

The following fifteen elements form the Dublin Core metadata set. All elements are optional and all can be repeated. The descriptions given below are condensed from the official Dublin Core definitions, with permission from the design team.

1. **Title.** The name given to the resource by the creator or publisher.
2. **Creator.** The person or organization primarily responsible for the intellectual content of the resource. For example, authors in the case of written documents, artists, photographers, or illustrators in the case of visual resources.
3. **Subject.** The topic of the resource. Typically, subject will be expressed as keywords or phrases that describe the subject or content of the resource. The use of controlled vocabularies and formal classification schemes is encouraged.
4. **Description.** A textual description of the content of the resource, including abstracts in the case of document-like objects or content descriptions in the case of visual resources.
5. **Publisher.** The entity responsible for making the resource available in its present form, such as a publishing house, a university department, or a corporate entity.
6. **Contributor.** A person or organization not specified in a creator element who has made significant intellectual contributions to the resource but whose contribution is secondary to any person or organization specified in a creator element (for example, editor, transcriber, and illustrator).
7. **Date.** A date associated with the creation or availability of the resource.
8. **Type.** The category of the resource, such as home page, novel, poem, working paper, preprint, technical report, essay, dictionary.
9. **Format.** The data format of the resource, used to identify the software and possibly hardware that might be needed to display or operate the resource.
10. **Identifier.** A string or number used to uniquely identify the resource. Examples for networked resources include URLs and URNs.
11. **Source.** Information about a second resource from which the present resource is derived.
12. **Language.** The language of the intellectual content of the resource.
13. **Relation.** An identifier of a second resource and its relationship to the present resource. This element permits links between related resources and resource descriptions to be indicated. Examples include an edition of a work (IsVersionOf), or a chapter of a book (IsPartOf).
14. **Coverage.** The spatial locations and temporal durations characteristic of the resource.
15. **Rights.** A rights management statement, an identifier that links to a rights management statement, or an identifier that links to a service providing information about rights management for the resource.

Simplicity is both the strength and the weakness of the Dublin Core. Whereas traditional cataloguing rules are long and complicated, requiring professional training to apply effectively, the Dublin Core can be described simply, but simplicity conflicts with precision. The team has struggled with this tension. Initially the aim was to create a single set of metadata elements, suitable for untrained people who publish electronic materials to describe their work. Some people continue to hold this minimalist view. They would like to see a simple set of rules that anybody can apply.

Other people prefer the benefits that come from more tightly controlled cataloguing rules and would accept the additional labor and cost. They point out that extra structure in the elements results in extra precision in the metadata records. For example, if entries in a subject field are drawn from the Dewey Decimal Classification, it is helpful to record that fact in the metadata. To further enhance the effectiveness of the metadata for information retrieval, several of the elements will have recommended lists of values. Thus, there might be a specified set of types and indexers would be recommended to select from the list.

The current strategy is to have two options, "minimalist" and "structuralist". The minimalist will meet the original criterion of being usable by people who have no formal training. The structured option will be more complex, requiring fuller guidelines and trained staff to apply them.

Automatic indexing

Cataloguing and indexing are expensive when carried out by skilled professionals. A rule of thumb is that each record costs about fifty dollars to create and distribute. In certain fields, such as medicine and chemistry, the demand for information is great enough to justify the expense of comprehensive indexing, but these disciplines are the exceptions. Even monograph cataloguing is usually restricted to an overall record of the monograph rather than detailed cataloguing of individual topics within a book. Most items in museums, archives, and library special collections are not catalogued or indexed individually.

In digital libraries, many items are worth collecting but the costs of cataloguing them individually can not be justified. The numbers of items in the collections can be very large, and the manner in which digital library objects change continually inhibits long-term investments in catalogs. Each item may go through several versions in quick succession. A single object may be composed of many other objects, each changing independently. New categories of object are being continually devised, while others are discarded. Frequently, the user's perception of an object is the result of executing a computer program and is different with each interaction. These factors increase the complexity and cost of cataloguing digital library materials.

For all these reasons, professional cataloguing and indexing is likely to be less central to digital libraries than it is in traditional libraries. The alternative is to use computer programs to create index records automatically. Records created by automatic indexing are normally of poor quality, but they are inexpensive. A powerful search system will go a long way towards compensating for the low quality of individual records. The web search programs prove this point. They build their indexes automatically. The records are not very good, but the success of the search services shows that the indexes are useful. At least, they are better than the alternative, which is to have nothing. Panel 10.4 gives two examples of records that were created by automatic indexing.

Panel 10.4
Examples of automatic indexing

The two following records are typical of the indexing records that are created automatically by web search programs. They are lightly edited versions of records that were created by the Altavista system in 1997.

> **Digital library concepts.** Key Concepts in the Architecture of the Digital Library. William Y. Arms Corporation for National Research Initiatives Reston, Virginia... http://www.dlib.org/dlib/July95/07arms.html - size 16K - 7-Oct-96 - English
>
> **Repository References.** Notice: HyperNews at union.ncsa.uiuc.edu will be moving to a new machine and domain very soon. Expect interruptions. Repository References. This is a page. http://union.ncsa.uiuc.edu/HyperNews/get/www/repo/references.html - size 5K - 12-May-95 - English

The first of these example shows automatic indexing at its best. It includes the author, title, date, and location of an article in an electronic journal. For many purposes, it is an adequate substitute for a record created by a professional indexer.

The second example shows some of the problems with automatic indexing. Nobody who understood the content would bother to index this web page. The information about location and date are probably all right, but the title is strange and the body of the record is simply the first few words of a the page.

Much of the development that led to automatic indexing came out of research in text skimming. A typical problem in this field is how to organize electronic mail. A user has a large volume of electronic mail messages and wants to file them by subject. A computer program is expected to read through them and assign them to subject areas. This is a difficult problem for people to carry out consistently and is a very difficult problem for a computer program, but steady progress has been made. The programs look for clues within the document. These clues may be structural elements, such as the subject field of an electronic mail message, they may be linguistic clues, or the program may simply recognize key words.

Automatic indexing also depends upon clues to be found in a document. The first of the examples in Panel 10.4 is a success, because the underlying web document provides useful clues. The Altavista indexing program was able to identify the title and author. For example, the page includes the tagged element:

```
<title>Digital library concepts</title>
```

The author inserted these tags to guide web browsers in displaying the article. They are equally useful in providing guidance to automatic indexing programs.

One of the potential uses of mark-up languages, such as SGML or XML, is that the structural tags can be used by automatic indexing programs to build records for information retrieval. Within the text of a document, the string, "Marie Celeste" might be the name of a person, a book, a song, a ship, a publisher, a play, or might not even be a name. With structural mark-up, the string can be identified and labeled for what it is. Thus, information provided by the mark-up can be used to distinguish specific categories of information, such as author, title, or date.

Automatic indexing is fast and cheap. The exact costs are commercial secrets, but they are a tiny fraction of one cent per record. For the cost of a single record created by a professional cataloguer or indexer, computer programs can generate a hundred thousand or more records. It is economically feasible to index huge numbers of items on the Internet and even to index them again at frequent intervals.

Creators of catalogs and indexes can balance costs against perceived benefits. The most expensive forms of descriptive metadata are the traditional methods used for library catalogs, and by indexing and abstracting services; structuralist Dublin Core will be moderately expensive, keeping most of the benefits while saving some costs; minimalist Dublin Core will be cheaper, but not free; automatic indexing has the poorest quality at a tiny cost.

Attaching metadata to content

Descriptive metadata needs to be associated with the material that it describes. In the past, descriptive metadata has usually been stored separately, as an external catalog or index. This has many advantages, but requires links between the metadata and the object it references. Some digital libraries are moving in the other direction, storing the metadata and the data together, either by embedding the metadata in the object itself or by having two tightly linked objects. This approach is convenient in distributed systems and for long-term archiving, since it guarantees that computer programs have access to both the data and the metadata at the same time.

Mechanisms for associating metadata with web pages have been a subject of considerable debate. For an HTML page, a simple approach is to embed the metadata in the page, using the special HTML tag , as in Table 10.1. These are the meta tags from an HTML description of the Dublin Core Element Set. Note that the choice of tags is a system design decision. The Dublin Core itself does not specify how the metadata is associated with the material.

Table 10.1
Metadata represented with HTML <meta> tags

```
<meta name="DC.subject"
content="dublin core metadata element set">

<meta name="DC.subject"
content="networked object description">

<meta name="DC.publisher"
content="OCLC Online Computer Library Center, Inc.">

<meta name="DC.creator"
content="Weibel, Stuart L., weibel@oclc.org.">

<meta name="DC.creator"
content="Miller, Eric J., emiller@oclc.org.">
```

```
<meta name="DC.title"
content="Dublin Core Element Set Reference Page">

<meta name="DC.date"
content="1996-05-28">

<meta name="DC.form" scheme="IMT"
content="text/html">

<meta name="DC.language" scheme="ISO639"
content="en">

<meta name="DC.identifier" scheme="URL"
content="http://purl.oclc.org/metadata/dublin_core">
```

Since meta tags can not be used with file types other than HTML and rapidly become cumbersome, a number of organizations working through the World Wide Web Consortium have developed a more general structure known as the Resource Description Framework (RDF). RDF is described in Panel 10.5.

Panel 10.5
The Resource Description Framework (RDF)

The Resource Description Framework (RDF) is a method that has been developed for the exchange of metadata. It has been developed by the World Wide Web Consortium, drawing concepts together from several other efforts, including the PICS format, which was developed to provide rating labels, to identify violence, pornography, and similar characteristics of web pages. The Dublin Core team is working closely with the RDF designers.

A metadata scheme, such as Dublin Core, can be considered as having three aspects: semantics, syntax, and structure. The semantics describes how to interpret concepts such as date or creator. The syntax specifies how the metadata is expressed. The structure defines the relationships between the metadata elements, such as the concepts of day, month and year as components of a date. RDF provides a simple but general structural model to express the syntax. It does not stipulate the semantics used by a metadata scheme. XML is used to describe a metadata scheme and for exchange of information between computer systems and among schemes.

The structural model consists of resources, property-types, and values. Consider the simple statement that Shakespeare is the author of the play Hamlet. In the Dublin Core metadata scheme, this can be represented as:

Resource		**Property-type**		**Value**
Hamlet	----------->	creator	------------->	Shakespeare
	----------->	type	------------->	play

A different metadata scheme, might use the term author in place of creator, and might use the term type with a completely different meaning. Therefore, the RDF mark-up would make explicit that this metadata is expressed in the Dublin core scheme:

```
<DC:creator> Shakespeare</DC:creator>
<DC:type> play</DC:type>
```

To complete this example, Hamlet needs to be identified more precisely. Suppose that it referenced by the (imaginary) URL, "http://hamlet.org/". Then the full RDF record, with XML mark-up, is:

```
<RDF:RDF>
  <RDF:description RDF:about = "http://hamlet.org/">
    <DC:creator> Shakespeare</DC:creator>
    <DC:type> play</DC:type>
  </RDF:description>
</RDF:RDF>
```

The mark-up in this record makes explicit that the terms description and about are defined in the RDF scheme, while creator and type are terms defined in the Dublin Core (DC). One more step is needed to complete this record: the schemes RDF and DC must be defined as XML namespaces.

The RDF structural model permits resources to have property-types that refer to other resources. For example, a database might include a record about Shakespeare with metadata about him, such as when and where he lived, and the various ways that he spelled his name. The DC:Creator property-type could reference this record as follows:

```
<DC:creator RDF:about = "http://people.net/WS/">
```

In this manner, arbitrarily complex metadata descriptions can be built up from simple components. By using the RDF framework for the syntax and structure, combined with XML representation, computer systems can associate metadata with digital objects and exchange metadata from different schemes.

Techniques of information retrieval

The rest of this chapter is a brief introduction to information retrieval. Information retrieval is a field in which computer scientists and information professionals have worked together for many years. It remains an active area of research and is one of the few areas of digital libraries to have a systematic methodology for measuring the performance of various methods.

Basic concepts and terminology

The various methods of information retrieval build on some simple concepts to search large bodies of information. A **query** is a string of text, describing the information that the user is seeking. Each word of the query is called a **search term**. A query can be a single search term, a string of terms, a phrase in natural language, or a stylized expression using special symbols.

Some methods of information retrieval compare the query with every word in the entire text, without distinguishing the function of the various words. This is called **full text searching**. Other methods identify bibliographic or structural fields, such as author or heading, and allow searching on specified field, such as "author = Gibbon". This is called **fielded searching**. Full text and fielded searching are both powerful tools, and modern methods of information retrieval often use the techniques in combination. Fielded searching requires some method of identifying the fields. Full text searching does not require such support. By taking advantage of the power of modern computers, full text searching can be effective even on unprocessed text, but heterogeneous texts of varying length, style, and content are difficult to search effectively and the results can be inconsistent. The legal information systems, Westlaw and Lexis, are based on full text searching; they are the exceptions. When descriptive metadata is available, most services prefer either fielded searching or free text searching of abstracts or other metadata.

Some words occur so frequently that they are of little value for retrieval. Examples include common pronouns, conjunctions, and auxiliary verbs, such as "he", "and", "be", and so on. Most systems have a list of common words which are ignored both in building inverted files and in queries. This is called a **stop list**. The selection of stop words is difficult. The choice clearly depends upon the language of the text and may also be related to the subject matter. For this reason, instead of have a predetermined stop list, some systems use statistical methods to identify the most commonly used words and reject them. Even then, no system is perfect. There is always the danger that some perfectly sensible queries might be rejected because every word is in the stop list, as with the quotation, "To be or not to be?"

Panel 10.6
Inverted files

An **inverted file** is a list of the words in a set of documents and their locations within those documents. Here is a small part of an inverted file.

Word	Document	Location
abacus	3	94
	19	7
	19	212
actor	2	66
	19	200
	29	45
aspen	5	43
atoll	11	3
	34	40

This inverted file shows that the word "abacus" is word 94 in document 3, and words 7 and 212 in document 19; the word "actor" is word 66 in document 2, word 200 in document 19, and word 45 in document 29; and so on. The list of locations for a given word is called an **inverted list**.

An inverted file can be used to search a set of documents to find every occurrence of a single search term. In the example above, a search for the word "actor" would look in the inverted file and find that the word appears in documents 2, 19, and 29. A simple reference to an inverted file is typically a fast operation for a computer.

Most inverted lists contain the location of the word within the document. This is important for displaying the result of searches, particularly with long documents. The section of the document can be displayed prominently with the search terms highlighted.

Since inverted files contain every word in a set of documents, except stop words, they are large. For typical digital library materials, the inverted file may approach half the total size of all the documents, even after compression. Thus, at the cost of storage space, an inverted file provides a fast way to find every occurrence of a single word in a collection of documents. Most methods of information retrieval use inverted files.

Boolean searching

Panel 10.6 describes inverted files, the basic computational method that is used to compare the search terms against a collection of textual documents. **Boolean queries** consist of two or more search terms, related by a logical operators, such as *and*, *or*, or *not*. Consider the query "abacus and actor" applied to the inverted file in Panel 10.6. The query includes two search terms separated by a Boolean operator. The first stage in carrying out this query is to read the inverted lists for "abacus" (documents 3 and 19) and for "actor" (documents 2, 19, and 29). The next stage is to compare the two lists for documents that are in both lists. Both words are in document 19, which is the only document that satisfies the query. When the inverted lists are short, Boolean searches with a few search terms are almost as fast as simple queries, but the computational requirements increase dramatically with large collections of information and complex queries.

Inverted files can be used to extend the basic concepts of Boolean searching. Since the location of words within documents are recorded in the inverted lists, they can be used for searches that specify the relative position of two words, such as a search for the word "West" followed by "Virginia". They can also be used for **truncation**, to search for words that begin with certain letters. In many search systems, a search for "comp?" will search for all words that begin the four letters "comp". This will find the related words "compute", "computing", "computer", "computers", and "computation". Unfortunately, this approach will not distinguish unrelated words that begin with the same letters, such as "company".

Ranking closeness of match

Boolean searching is a powerful tool, but it finds exact matches only. A search for "library" will miss "libraries"; "John Smith" and "J. Smith" are not treated as the same name. Yet everybody recognizes that these are similar. A range of techniques address such difficulties.

The modern approach is not to attempt to match documents exactly against a query but to define some measure of similarity between a query and each document. Suppose that the total number of different words in a set of documents is n. A given document can be represented by a vector in n-dimensional space. If the document contains a given word, the vector has value 1 in the corresponding dimension, otherwise 0. A query can also be represented by a vector in the same space. The closeness with which a document matches a query is measured by how close these two vectors are to each other. This might be measured by the angle between these two vectors in n-dimensional space. Once these measures have been calculated for every document, the results can be ranked from the best match to the least good. Several ranking technique are variants of the same general concepts. A variety of probabilistic methods make use of the statistical distribution of the words in the collection. These methods derive from the observation that the exact words chosen by an author to describe a topic or by a user to express a query were chosen from a set of possibilities, but that other words might be equally appropriate.

Natural language processing and computational linguistics

The words in a document are not simply random strings of characters. They are words in a language, such as English, arranged into phrases, sentences, and paragraphs. **Natural language processing** is the branch of computer science that uses computers to interpret and manipulate words as part of a language. The spelling checkers that are used with word processors are a well-known application. They use methods of natural language processing to suggest alternative spellings for words that they do not recognize.

Computational linguistics deals with grammar and linguistics. One of the achievements of computational linguistics has been to develop computer programs that can parse almost any sentence with good accuracy. A parser analyzes the structure of sentences. It categorizes words by part of speech (verb, noun, adjective, etc.), groups them into phrases and clauses, and identifies the structural elements (subject, verb, object, etc.). For this purpose, linguists have been required to refine their understanding of grammar, recognizing far more subtleties than were contained in traditional grammars. Considerable research in information retrieval has been carried out using noun phrases. In many contexts, the content of a sentence can be found by extracting the nouns and noun phrases and searching on them. This work has not been restricted to English, but has been carried out for many languages.

Parsing requires an understanding of the **morphology** of words, that is variants derived from the same stem, such as plurals (library, libraries), and verb forms (look, looks, looked). For information retrieval, it is often effective to reduce morphological variants to a common stem and to use the stem as a search term. This is called **stemming**. Stemming is more effective than truncation since it separates words with totally different meanings, such as "computer" from "company", while recognizing that "computer" and "computing" are morphological variants from the same stem. In English, where the stem is almost always at the beginning of the word, stemming can be carried out by truncating words and perhaps making adjustments to the final few letters. In other language, such as German, it is also necessary to trim at the beginning of words.

Computational linguists have developed a range of dictionaries and other tools, such as lexicons and thesauruses, that are designed for natural language processing. A **lexicon** contains information about words, their morphological variants, and their grammatical usage. A **thesaurus** relates words by meaning. Some of these tools are general purpose; others are tied to specific disciplines. Two were described earlier in this chapter; the Art and Architecture Thesaurus and the MeSH headings for medicine. Linguistic can greatly augment information retrieval. By recognizing words as more than random strings of characters, they can recognize synonyms ("car" and "automobile"), relate a general term and a particular instance ("science" and "chemistry"), or a technical term and the vernacular ("cranium" and "brain"). The creation of a lexicon or thesaurus is a major undertaking and is never complete. Languages change continually, especially the terminology of fields in which there is active research.

User interfaces and information retrieval systems

Information retrieval systems depend for their effectiveness on the user making best use of the tools provided. When the user is a trained medical librarian or a lawyer whose legal education included training in search systems, these objectives are usually met. Untrained users typically do much less well at formulating queries and understanding the results.

A feature of the vector-space and probabilistic methods of information retrieval is that they are most effective with long queries. An interesting experiment is to use a very long query, perhaps an abstract from a document. Using this as a query is equivalent to asking the system to find documents that match the abstract. Many modern search systems are remarkably effective when given such an opportunity, but methods that are based on vector space or linguistic techniques require a worthwhile query to display their full power.

Statistics of the queries that are used in practice show that most queries consist of a single word, which is a disappointment to the developers of powerful retrieval systems. One reasons for these short queries is that many users made their first searches on Boolean systems, where the only results found are exact matches, so that a long query usually finds no matches. These early systems had another characteristic that encouraged short queries. When faced with a long or complex query their performance deteriorated terribly. Users learned to keep their queries short. Habits that were developed with these systems have been retained even with more modern systems.

However, the tendency of users to supply short queries is more entrenched than can be explained by these historical factors, or by the tiny input boxes sometimes provided. The pattern is repeated almost everywhere. People appear to be inhibited from using long queries. Another unfortunate characteristic of users, which is widely observed, is that few people read even the most simple instructions. Digital libraries are failing to train their users in effective searching and users do not take advantage of the potential of the systems that they use.

Evaluation

Information retrieval has a long tradition of performance evaluation. Two long-standing criteria are **precision** and **recall**. Each refers to the results from carrying out a single search on a given body of information. The result of such a search is a set of hits. Ideally every hit would be relevant to the original query, and every relevant item in the body of information would be found. In practice, it usually happens that some of the hits are irrelevant and that some relevant items are missed by the search.

- **Precision** is the percentage of the hits that are relevant, the extent to which the set of hits retrieved by a query satisfies the requirement that generated the query.

- **Recall** is the percentage of the relevant items that are found by the query, the extent to which the query found all the items that satisfy the requirement.

Suppose that, in a collection of 10,000 documents, 50 are on a specific topic. An ideal search would find these 50 documents and reject all others. An actual search identifies 25 documents of which 20 prove to be relevant but 5 were on other topics. In this instance, the precision is 20 out of 25, or 0.8. The recall is 20 out of 50, or 0.4.

Precision is much easier to measure than recall. To calculate precision, a knowledgeable person looks at each document that is identified and decides whether it is relevant. In the example, only the 25 documents that are found need to be examined. Recall is difficult to measure, since there is no way to know all the items in a collection that satisfy a specific query other than to go systematically through the entire collection, looking at every object to decide whether it fits the criteria. In this example, all 10,000 documents must be examined. With large numbers of documents, this is an imposing task.

Tipster and TREC

Tipster

Tipster was a long-running project sponsored by DARPA to improve the quality of text processing methods. The focus was on several problems, all of which are important in digital libraries. Research on document detection combines both information retrieval on stored documents and identifying relevant documents from a stream of new text. Information extraction is the capability to locate specified information within a text, and summarization the capability to condense the size of a document or collection.

Over the years, Tipster has moved its emphasis from standard information retrieval to the development of components that can tackle specific tasks and architectures for sharing these components. The architecture is an ambitious effort, based on concepts from object-oriented programming, to define standard classes for the basic components of textual materials, notably documents, collections, attributes, and annotations.

The TREC conferences

TREC is the acronym for the annual series of Text Retrieval Conferences, where researchers can demonstrate the effectiveness of their methods on standard bodies of text. The TREC conferences are an outstanding example of quantitative research in digital libraries. They are the creation of the National Institute of Standards and Technology, with the help of many other organizations. The organizers of the conferences have created a corpus of several million textual documents, a total of more than five gigabytes of data. Researchers evaluate their work by attempting a standard set of tasks. One task is to search the corpus for topics provided by a group of twenty surrogate users. Another task evaluates systems that match a stream of incoming documents against standard queries. Participants at TREC conferences include large commercial companies, small information retrieval vendors, and university research groups. By evaluating their different methods on the same large collection they are able to gauge their strengths and enhance their systems.

The TREC conferences provides an opportunity to compare the performance of different techniques, including methods using automatic thesauri, sophisticated term weighting, natural language techniques, relevance feedback, and advanced machine learning. In later years, TREC has introduced a number of smaller tracks to evaluate other aspects of information retrieval, including Chinese texts, spoken documents, and cross-language retrieval. There is also a track that is experimenting with methods for evaluating interactive retrieval.

The TREC conferences have had enormous impact on research in information retrieval. This is an impressive program and a model for all areas of digital libraries research.

The criteria of precision and recall have been of fundamental importance in the development of information retrieval since they permit comparison of different methods, but they were devised in days when computers were slower and more expensive than today. Information retrieval then consisted of a single search of a large set of data. Success or failure was a one-time event. Today, searching is usually interactive. A user will formulate a query, carry out an initial search, examine the results, and repeat the process with a modified query. The effective precision and recall should be judged by the overall result of a session, not of a single search. In the jargon of the field, this is called searching "with a human in the loop".

Performance criteria, such as precision and recall, measure technical properties of aspects of computer systems. They do not measure how a user interacts with a system, or what constitutes an adequate result of a search. Many newer search programs have a strategy of ranking all possible hits. This creates a high level of recall at the expense of many irrelevant hits. Hopefully, the higher ranking hits will have high precision, with a long tail of spurious hits. Criteria are needed that measure the effectiveness of the ranking in giving high ranks to the most relevant items. This chapter began by recognizing that users look for information for many different reasons and use many strategies to seek for information. Sometimes they are looking for specific facts; sometimes they are exploring a topic; only rarely are they faced with the standard information retrieval problem, which is to find every item relevant to a well-defined topic, with the minimal number of extraneous items. With interactive computing, users do not carry out a single search. They iterate through a series of steps, combining searching, browsing, interpretation, and filtering of results. The effectiveness of information discovery depends upon the users' objectives and how well the digital library meets them.

Chapter 11
Distributed information discovery

Distributed computing and interoperability

This chapter is about distributed information retrieval, seeking for information that is spread across many computer systems. This is part of the broad challenge of interoperability, which is an underlying theme of this entire book. Therefore, the chapter begins with a discussion of the general issues of interoperability.

The digital libraries of the world are managed by many organizations, with different management styles, and different attitudes to collections and technology. Few libraries have more than a small percentage of the materials that users might want. Hence, users need to draw from collections and services provided by many different sources. How does a user discover and have access to information when it could be drawn from so many sources?

The technical aspects of coordinating separate computers so that they provide coherent service is called **distributed computing**. Distributed computing requires that the various computers share some technical standards. With distributed searching, for example, a user might want to search many independent collections with a single query, compare the results, choose the most promising, and retrieve selected materials from the collections. Beyond the underlying networking standards, this requires some method of identifying the collections, conventions for formulating the query, techniques for submitting it, means to return results so that they can be compared, and methods for obtaining the items that are discovered. The standards may be formal standards, blessed by official standards bodies, or they may be local standards developed by a small group of collaborators, or agreements to use specific commercial products. However, distributed computing is not possible without some shared standards.

An ideal approach would be to develop a comprehensive set of standards that all digital libraries would adopt. This concept fails to recognize the costs of adopting standards, especially during times of rapid change. Digital libraries are in a state of flux. Every library is striving to improve its collections, services, and systems, but no two libraries are the same. Altering part of a system to support a new standard is time-consuming. By the time the alteration is completed there may be a new version of the standard, or the community may be pursuing some other direction. Comprehensive standardization is a mirage.

The **interoperability** problem of digital libraries is to develop distributed computing systems in a world where the various computers are independently operated and technically dissimilar. Technically, this requires formats, protocols, and security systems so that messages can be exchanged. It also requires semantic agreements on the interpretation of the messages. These technical aspects are hard, but the central challenge is to find approaches that independent digital libraries have incentives to incorporate. Adoption of shared methods provides digital libraries with extra functionality, but shared methods also bring costs. Sometimes the costs are directly financial: the purchase of equipment and software, hiring and training staff. More often the major costs are organizational. Rarely can one aspect of a digital library be changed in isolation. Introducing a new standard requires inter-related changes to existing systems, altered work flow, changed relationships with suppliers, and so on.

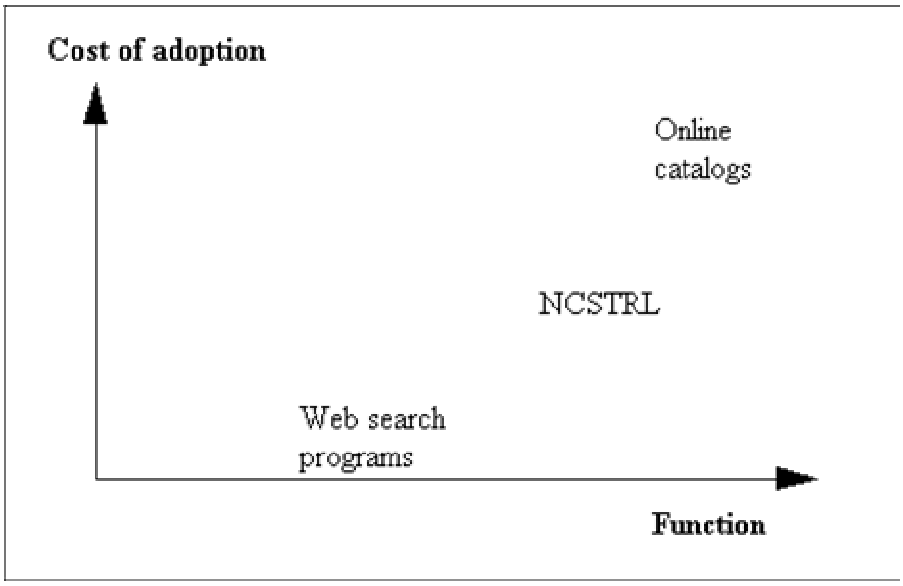


Figure 11.1. Strategies for distributed searching: function versus cost of adoption

Figure 11.1 shows a conceptual model that is useful in thinking about interoperability; in this instance it is used to compare three methods of distributed searching. The horizontal axis of the figure indicates the functionality provided by various methods. The vertical axis indicates the costs of adopting them. The ideal methods would be at the bottom right of the graph, high functionality at low cost. The figure shows three particular methods of distributed searching, each of which is discussed later in this chapter. The web search programs have moderate functionality; they are widely used because they have low costs of adoption. Online catalogs based on MARC cataloguing and the Z39.50 protocol have much more function, but, because the standards are complex, they are less widely adopted. The NCSTRL system lies between them in both function and cost of adoption.

More generally. it is possible to distinguish three broad classes of methods that might be used for interoperability.

- Most of the methods that are in widespread use for interoperability today have moderate function and low cost of acceptance. The main web standards, HTML, HTTP, and URLs, have these characteristics. Their simplicity has led to wide adoption, but limits the functions that they can provide.

- Some high-end services provide great functionality, but are costly to adopt. Z39.50 and SGML are examples. Such methods are popular in restricted communities, where the functionality is valued, but have difficulty in penetrating into broader communities, where the cost of adoption becomes a barrier.

- Many current developments in digital libraries are attempts to find the middle ground, substantial functionality with moderate costs of adoption. Examples include the Dublin Core, XML, and Unicode. In each instance, the designers have paid attention to providing a moderate cost route for adoption. Dublin Core allows every field to be optional. Unicode provides UTF-8, which accepts existing ASCII data. XML reduces the cost of adoption by its close relationships with both HTML and SGML.

The figure has no scale and the dimensions are only conceptual, but it helps to understand the fundamental principle that the costs of adopting new technology are a factor in every aspect of interoperability. Technology can never be considered by itself, without studying the organizational impact. When the objective is to interoperate with others, the creators of a digital library are often faced with the choice between the methods that are best for their particular community and adopting more generally accepted standards, even though they offer lesser functionality.

New versions of software illustrate this tension. A new version will often provide the digital library with more function but fewer users will have access to it. The creator of a web site can use the most basic HTML tags, a few well-established formats, and the services provided by every version of HTTP; this results in a simple site that can be accessed by every browser in the world. Alternatively, the creator can choose the latest version of web technology, with Java applets, HTML frames, built-in security, style sheets, with audio and video inserts. This will provide superior service to those users who have high-speed networks and the latest browsers, but may be unusable by others.

Web search programs

The most widely used systems for distributed searching are the web search programs, such as Infoseek, Lycos, Altavista, and Excite. These are automated systems that provide an index to materials on the Internet. On the graph in Figure 11.1, they provide moderate function with low barriers to use: web sites need take no special action to be indexed by the search programs, and the only cost to the user is the tedium of looking at the advertisements. The combination of respectable function with almost no barriers to use makes the web search programs extremely popular.

Most web search program have the same basic architecture, though with many differences in their details. The notable exception is Yahoo, which has its roots in a classification system. The other systems have two major parts: a web crawler which builds an index of material on the Internet, and a retrieval engine which allows users on the Internet to search the index.

Web crawlers

The basic way to discover information on the web is to follow hyperlinks from page to page. A web indexing programs follows hyperlinks continuously and assembles a list of the pages that it finds. Because of the manner in which the indexing programs traverse the Internet, they are often called **web crawlers**.

A web crawler builds an ever-increasing index of web pages by repeating a few basic steps. Internally, the program maintains a list of the URLs known to the system, whether or not the corresponding pages have yet been indexed. From this list, the crawler selects the URL of an HTML page that has not been indexed. The program retrieves this page and brings it back to a central computer system for analysis. An automatic indexing program examines the page and creates an index record for it which is added to the overall index. Hyerlinks from the page to other pages are extracted; those that are new are added to the list of URLs for future exploration.

Behind this simple framework lie many variations and some deep technical problems. One problem is deciding which URL to visit next. At any given moment, the web crawler has millions of unexplored URLs, but has little information to know which to select. Possible criteria for choice might include currency, how many other URLs link to the page, whether it is a home page or a page deep within a hierarchy, whether it references a CGI script, and so on.

The biggest challenges concern indexing. Web crawlers rely on automatic indexing methods to build their indexes and create records to present to users. This was a topic discussed in Chapter 10. The programs are faced with automatic indexing at its most basic: millions of pages, created by thousands of people, with different concepts of how information should be structured. Typical web pages provide meager clues for automatic indexing. Some creators and publishers are even deliberately misleading; they fill their pages with terms that are likely to be requested by users, hoping that their pages will be highly ranked against common search queries. Without better structured pages or systematic metadata, the quality of the indexing records will never be high, but they are adequate for simple retrieval.

Searching the index

The web search programs allow users to search the index, using information retrieval methods of the kind described in Chapter 10. The indexes are organized for efficient searching by large numbers of simultaneous users. Since the index records themselves are of low quality and the users likely to be untrained, the search programs follow the strategy of identifying all records that vaguely match the query and supplying them to the user in some ranked order.

Most users of web search programs would agree that they are remarkable programs, but have several significant difficulties. The ranking algorithms have little information to base their decisions on. As a result, the programs may give high ranks to pages of marginal value; important materials may be far down the list and trivial items at the top. The index programs have difficulty recognizing items that are duplicates, though they attempt to group similar items; since similar items tend to rank together, the programs often return long lists of almost identical items. One interesting approach to ranking is to use link counts. Panel 11.1 describes Google, a search system that has used this approach. It is particularly effective in finding introductory or overview material on a topic.

Panel 11.1
Page ranks and Google

Citation analysis is a commonly used tool in scientific literature. Groups of articles that cite each other are probably about related topics; articles that are heavily cited are likely to be more important than articles that are never cited. Lawrence Page, Sergey Brin, and colleagues at Stanford University have applied this concept to the web. They have used the patterns of hyperlinks among web pages as a basis for ranking pages and incorporated their ideas into an experimental web search program known as Google.

As an example, consider a search for the query "Stanford University" using various web search programs. There are more than 200,000 web pages at Stanford; most search programs have difficulty in separating those of purely local or ephemeral interest from those with broader readership. All web search programs find enormous numbers of web pages that match this query, but, in most cases, the order in which they are ranked fails to identify the sites that most people would consider to be most important. When this search was submitted to Google, the top ten results were the following. Most people would agree that this is a good list of high-ranking pages that refer to Stanford University.

Stanford University Homepage
(www.stanford.edu/)

Stanford University Medical Center
(www-med.stanford.edu/)

Stanford University Libraries & Information Resources
(www-sul.stanford.edu/)

Stanford Law School
(www-leland.stanford.edu/group/law/)

Stanford Graduate School of Business
(www-gsb.stanford.edu/)

Stanford University School of Earth Sciences
(pangea.stanford.edu/)

SUL: Copyright & Fair Use
(fairuse.stanford.edu/)

Computer Graphics at Stanford University
(www-graphics.stanford.edu/

SUMMIT (Stanford University) Home Page
(summit.stanford.edu/)

Stanford Medical Informatics
(camis.stanford.edu/)

The basic method used by Google is simple. A web page to which many other pages provide links is given a higher rank than a page with fewer links. Moreover, links from high-ranking pages are given greater weight than links from other pages. Since web pages around the world have links to the home page of the Stanford Law School, this page has a high rank. In turn, it links to about a dozen other pages, such as the university's home page, which gain rank from being referenced by a high-ranking page.

Calculating the page ranks is an elegant computational challenge. To understand the basic concept, imagine a huge matrix listing every page on the web and identifying every page that links to it. Initially, every page is ranked equally. New ranks are then calculated, based on the number of links to each page, weighted according to the rank of the linking pages and proportional to the number of links from each. These ranks are used for another iteration and the process continued until the calculation converges.

The actual computation is a refinement of this approach. In 1998, Google had a set of about 25 million pages, selected by a process derived from the ranks of pages that link to them. The program has weighting factors to account for pages with no links, or groups of pages that link only to each other. It rejects pages that are generated dynamically by CGI scripts. A sidelight on the power of modern computers is that the system was able to gather, index, and rank these pages in five days using only standard workstation computers.

The use of links to generate page ranks is clearly a powerful tool. It helps solve two problems that bedevil web search programs: since they can not index every page on the web simultaneously, which should they index first, and how should they rank the pages found from simple queries to give priority to the most useful.

Web search programs have other weaknesses. Currency is one. The crawlers are continually exploring the web. Eventually, almost everything will be found, but important materials may not be indexed until months after they are published on the web. Conversely, the programs do an indifferent job at going back to see if materials have been withdrawn, so that many of the index entries refer to items that no longer exist or have moved.

Another threat to the effectiveness of web indexing is that a web crawler can index material only if it can access it directly. If a web page is protected by some form of authentication, or if a web page is an interface to a database or a digital library collection, the indexer will know nothing about the resources behind the interface. As more and more web pages are interfaces controlled by Java programs or other scripts, this results in high-quality information being missed by the indexes.

These problems are significant but should not be over-emphasized. The proof lies in the practice. Experienced users of the web can usually find the information that they want. They use a combination of tools, guided by experience, often trying several web search services. The programs are far from perfect but they are remarkably good - and use of them is free.

Business issues

A fascinating aspect of the web search services is their business model. Most of the programs had roots in research groups, but they rapidly became commercial companies. Chapter 6 noted that, initially, some of these organizations tried to require users to pay a subscription, but Lycos, which was developed by a researcher at Carnegie Mellon University, was determined to provide public, no-cost searching. The others were forced to follow. Not charging for the basic service has had a profound impact on the Internet and on the companies. Their search for revenue has led to aggressive attempts to build advertising. They have rapidly moved into related markets, such as licensing their software to other organizations so that they can build indexes to their own web sites.

A less desirable aspect of this business model is that the companies have limited incentive to have a comprehensive index. At first, the indexing programs aimed to index the entire web. As the web has grown larger and the management of the search programs has become a commercial venture, comprehensiveness has become secondary to improvements in interfaces and ancillary services. To build a really high quality index of the Internet, and to keep it up to date, requires a considerable investment. Most of the companies are content to do a reasonable job, but with more incentives, their indexes would be better.

Federated digital libraries

The tension that Figure 11.1 illustrates between functionality and cost of adoption has no single correct answer. Sometimes the appropriate decision for digital libraries is to select simple technology and strive for broad but shallow interoperability. At other time, the wise decision is to select technology from the top right of the figure, with great functionality but associated costs; since the costs are high, only highly motivated libraries will adopt the methods, but they will see higher functionality.

The term **federated digital library** describes a group of organizations working together, formally or informally, who agree to support some set of common services and standards, thus providing interoperability among their members. In a federation, the partners may have very different systems, so long as they support an agreed set of services. They will need to agree both on technical standards and on policies, including financial agreements, intellectual property, security, and privacy.

Research at the University of Illinois, Urbana Champaign provides a revealing example of the difficulties of interoperability. During 1994-98, as part of the Digital Libraries Initiative, a team based at the Grainger Engineering Library set out to build a federated library of journal articles from several leading science publishers. Since each publisher planned to make its journals available with SGML mark-up, this appeared to be an opportunity to build a federation; the university would provide central services, such as searching, while the collections would be maintained by the publishers. This turned out to be difficult. A basic problem was incompatibility in the way that the publishers use SGML. Each has its own Document Type Definition (DTD). The university was forced to enormous lengths to reconcile the semantics of the DTDs, both to extract indexing information and to build a coherent user interface. This problem proved to be so complex that the university resorted to copying the information from the publishers' computers onto a single system and converting it to a common DTD. If a respected university research group encountered such difficulties with a relatively coherent body of information, it is not surprising that others face the same problems. Panel 11.2 describes this work in more detail.

Panel 11.2
The University of Illinois federated library of scientific literature

The Grainger Engineering Library at the University of Illinois is the center of a prototype library that is a federation of collections of scientific journal articles. The work began as part of the Digital Libraries Initiative under the leadership of Bruce Schatz and William Mischo. By 1998, the testbed collection had 50,000 journal articles from the following publishers. Each of the publishers provides journal articles with SGML mark-up at the same time as their printed journals are published.

The IEEE Computer Society

The Institute of Electrical and Electronic Engineers (IEEE)

The American Physical Society

The American Society of Civil Engineers

The American Institute of Aeronautics and Astronautics

The prototype has put into practice concepts of information retrieval from marked-up text that have been frequently discussed, but little used in practice. The first phase, which proved to be extremely tough, was to reconcile the DTDs (Document Type Definitions) used by the different publishers. Each publisher has its own DTD to represent the structural elements of its documents. Some of the difference are syntax; <author> , <aut> , or <au> are alternatives for an author tag. Other variations, however, reflect significant semantic differences. For indexing and retrieval, the project has written software that maps each DTD's tags into a canonical set. The interfaces to the digital library use these tags, so that users can search for text in a given context, such as looking for a phrase within the captions on figures.

The use of this set of tags lies to the top right of Figure 11.1. Converting the mark-up for all the collections to this set imposes a high cost whenever a new collection is added to the federation, but provides powerful functionality.

Because of technical difficulties, the first implementation loaded all the documents into a single repository at the University of Illinois. Future plans call for the federation to use repositories maintained by individual publishers. There is also interest in expanding the collections to embrace bibliographic databases, catalogs, and other indexes.

Even the first implementation proved to be a fertile ground for studying users and their wishes. Giving users more powerful methods of searching was welcomed, but also stimulated requests. Users pointed out that figures or mathematical expressions are often more revealing of content than abstracts of conclusions. The experiments have demonstrated, once again, that users have great difficulty finding the right words to include in search queries when there is no control over the vocabulary used in the papers, their abstracts and the search system.

Online catalogs and Z39.50

Many libraries have online catalogs of their holding that are openly accessible over the Internet. These catalogs can be considered to form a federation. As described in Chapter 3, the catalog records follow the Anglo American Cataloguing Rules, using the MARC format, and libraries share records to reduce the costs. The library community developed the Z39.50 protocol to meets its needs for sharing records and distributed searching; Z39.50 is described in Panel 11.3. In the United States, the Library of Congress, OCLC and the Research Libraries Group, have been active in developing and promulgating these standards; there have been numerous independent implementations at academic sites and by commercial vendors. The costs of belonging to this federation are high, but they have been absorbed over decades, and are balanced by the cost savings from shared cataloguing.

Panel 11.3
Z39.50

Z39.50 is a protocol, developed by the library community, that permits one computer, the client, to search and retrieve information on another, the database server. Z39.50 is important both technically and for its wide use in library systems. In concept, Z39.50 is not tied to any particular category of information or type of database, but much of the development has concentrated on bibliographic data. Most implementations emphasize searches that use a bibliographic set of attributes to search databases of MARC catalog records and present them to the client.

Z39.50 is built around a abstract view of database searching. It assumes that the server stores a set of databases with searchable indexes. Interactions are based on the concept of a session. The client opens a connection with the server, carries out a sequence of interactions and then closes the connection. During the course of the session, both the server and the client remember the state of their interaction. It is important to understand that the client is a computer. End-user applications of Z39.50 need a user interface for communication with the user. The protocol makes no statements about the form of that user interface or how it connects to the Z39.50 client.

A typical session begins with the client connecting to the server and exchanging initial information, using the *init* facility. This initial exchange establishes agreement on basics, such as the preferred message size; it can include authentication, but the actual form of the authentication is outside the scope of the standard. The client might then use the explain service to inquire of the server what databases are available for searching, the fields that are available, the syntax and formats supported, and other options.

The *search* service allows a client to present a query to a database, such as:

> In the database named "Books" find all records for which the access point title contains the value "evangeline" and the access point author contains the value "longfellow."

The standard provides several choices of syntax for specifying searches, but only Boolean queries are widely implemented. The server carries out the search and builds a *results set*. A distinctive feature of Z39.50 is that the server saves the results set. A subsequent message from the client can reference the result set. Thus the client can modify a large set by increasingly precise requests, or can request a presentation of any record in the set, without searching the entire database.

Depending on parameters of the search request, one or more records may be returned to the client. The standard provides a variety of ways that clients can manipulate results sets, including services to *sort* or *delete* them. When the searching is complete, the next step is likely to be that the client sends a *present* request. This requests the server to send specified records from the results set to the client in a specified format. The present service has a wide range of options for controlling content and formats, and for managing large records or large results sets.

This is a large and flexible standard. In addition to these basic services, Z39.50 has facilities for browsing indexes, for access control and resource management, and supports extended services that allow a wide range of extensions.

One of the principal applications of Z39.50 is for communication between servers. A catalog system at a large library can use the protocol to search a group of peers to see if they have either a copy of a work or a catalog record. End users can use a single Z39.50 client to search several catalogs, sequentially, or in parallel. Libraries and their patrons gain considerable benefits from sharing catalogs in these ways, yet interoperability among public access catalogs is still patchy. Some Z39.50 implementations have features that others lack, but the underlying cause is that the individual catalogs are maintained by people whose first loyalty is to their local communities. Support for other institutions is never the first priority. Even though they share compatible versions of Z39.50, differences in how the catalogs are organized and presented to the outside world remain.

NCSTRL and Dienst

A **union catalog** is a single catalog that contains records about the materials in several libraries. Union catalogs were used by libraries long before computers. They solve the problem of distributed searching by consolidating the information to be searched into a single catalog. Web search services can be considered to be a union catalogs for the web, albeit with crude catalog records. An alternative method of distributed searching is not to build a union catalog, but for each collection to have its own searchable index. A search program sends queries to these separate indexes and combines the results for presentation to the user.

Panel 11.4 describes an interesting example. The Networked Computer Science Technical Reference Library (**NCSTRL**) is a federation of digital library collections that are important to computer science researchers. It uses a protocol called **Dienst**. To minimize the costs of acceptance, Dienst builds on a variety of technical standards that are familiar to computer scientists, who are typically heavy users of Unix, the

Internet, and the web. The first version of Dienst sent search requests to all servers. As the number of servers grew this approach broke down; Dienst now uses a search strategy that makes use of a master index, which is a type of union catalogs. For reasons of performance and reliability, this master index is replicated at regional centers.

Panel 11.4
NCSTRL and the Dienst model of distributed searching

The Networked Computer Science Technical Reference Library (NCSTRL) is a distributed library of research materials in computer science, notably technical reports. Cooperating organizations mount their collections on locally maintained servers. Access to these servers uses either FTP or Dienst, a distributed library protocol developed by Jim Davis of Xerox and Carl Lagoze of Cornell University. Dienst was developed as part of the Computer Science Technical Reports project, mentioned in Chapter 4. Initially there were five cooperating universities, but, by 1998, the total number was more than 100 organizations around the world, of whom forty three were operating Dienst servers.

Dienst is an architecture that divides digital library services into four basic categories: repositories, indexes, collections, and user interfaces. It provides an open protocol that defines these services. The protocol supports distributed searching of independently managed collections. Each server has an index of the materials that it stores. In early versions of Dienst, to search the collections, a user interface program sent a query to all the Dienst sites, seeking for objects that match the query. The user interface then waited until it received replies from all the servers. This is distributed searching at its most basic and it ran into problems when the number of servers grew large. The fundamental problem was that the quality of service seen by a user was determined by the service level provided by the worst of the Dienst sites. At one period, the server at Carnegie Mellon University was undependable. If it failed to respond, the user interface waited until an automatic time-out was triggered. Even when all servers were operational, since the performance of the Internet is highly variable, there were often long delays caused by slow connections to a few servers.

A slow search is tedious for the user; for a search to miss some of the collections is more serious. When a user carries out research, failure to search all the indexes means that the researcher might miss important information, simply because of a server problem.

To address these problems, Dienst was redesigned. NCSTRL is now divided into regions. Initially there were two regional centers in the United States and four in Europe. With this regional model, a master index is maintained at a central location, Cornell University, and searchable copies of this index are stored at the regional centers. Everything that a user needs to search and locate information is at each regional site. The user contacts the individual collection sites only to retrieve materials stored there. The decision which regional site to use is left to the individual user; because of the vagaries of the Internet, a user may get best performance by using a regional site that is not the closest geographically.

NCSTRL and Dienst combine day-to-day services with a testbed for distributed information management. It is one of only a small number of digital libraries where a research group operates an operational service.

Research on alternative approaches to distributed searching

In a federation, success or failure is at least as much an organizational challenge as a technical one. Inevitably, some members provide better service than others and their levels of support differ widely, but the quality of service must not be tied to the worst performing organization. Panel 11.4 describes how the Dienst system, used by NCSTRL, was redesigned to address this problem.

Every information service makes some implicit assumptions about the scenarios that it supports, the queries that it accepts, and the kinds of answers that it provides. These are implemented as facilities, such as search, browse, filter, extract, etc. The user desires coherent, personalized information, but, in the distributed world, the information sources may individually be coherent, but collectively differ. How can a diverse set of organizations provide access to their information in a manner that is effective for the users? How can services that are designed around different implied scenarios provide effective resource discovery without draconian standardization? There are tough technical problems for a single, centrally managed service. They become really difficult when the information sources are controlled by independent organizations.

The organizational challenges are so great that they constrain the technical options. Except within tight federations, the only hope for progress is to establish technical frameworks that organizations can accept incrementally, with parallel acceptance strategies for them. For each of methods there must be a low-level alternative, usually the status quo, so that services are not held back from doing anything worthwhile because of a few systems. Thus, in NCSTRL, although Dienst is the preferred protocol, more than half the sites mount their collections on simple servers using the FTP protocol.

Much of the research in distributed searching sets out to build union catalogs from metadata provided by the creator or publisher. This was one of the motivations behind the Dublin Core and the Resource Description Framework (RDF), which were described in Chapter 10. Computers systems are needed to assemble this metadata and consolidate it into a searchable index. Panel 11.5 describes the Harvest architecture for building indexes of many collections.

Panel 11.5
The Harvest architecture

Harvest was a research project in distributed searching, led by Michael Schwartz who was then at the University of Colorado. Although the project ended in 1996, the architectural ideas that it developed remain highly relevant. The underlying concept is to take the principal functions that are found in a centralized search system and divide them into separate subsystems. The project defined formats and protocols for communication among these subsystems, and implemented software to demonstrate their use.

A central concept of Harvest is a **gatherer**. This is a program that collects indexing information from digital library collections. Gatherers are most effective when they are installed on the same system as the collections. Each gatherer extracts indexing information from the collections and transmits it in a standard format and protocol to programs called brokers. A broker builds a combined index with information about many collections.

The Harvest architecture is much more efficient in its use of network resources than indexing methods that rely on web crawlers, and the team developed caches and methods of replication for added efficiency, but the real benefit is better searching and information discovery. All gatherers transmit information in a specified protocol, called the Summary Object Interchange Format (SOIF), but how they gather the information can be tailored to the individual collections. While web crawlers operate only on open access information, gatherers can be given access privileges to index restricted collections. They can be configured for specific databases and need not be restricted to web pages or any specific format. They can incorporate dictionaries or lexicons for specialized topic areas. In combination, these are major advantages.

Many benefits of the Harvest architecture are lost if the gatherer is not installed locally, with the digital library collections. For this reason, the Harvest architecture is particularly effective for federated digital libraries. In a federation, each library can run its own gatherer and transmit indexing information to brokers that build consolidated indexes for the entire library, combining the benefits of local indexing with a central index for users.

Another area of research is to develop methods for restricting searches to the most promising collections. Users rarely want to search every source of information on the Internet. They want to search specific categories, such as monograph catalogs, or indexes to medical research. Therefore, some means is needed for collections to provide summaries of their contents. This is particularly important where access is limited by authentication or payment mechanisms. If open access is provided to a source, an external program can, at least in theory, generate a statistical profile of the types of material and the vocabulary used. When an external user has access only through a search interface such analysis is not possible.

Luis Gravano, while at Stanford University, studied how a client can combine results from separate search services. He developed a protocol, known as STARTS, for this purpose. This was a joint project between Stanford University and several leading Internet companies. The willingness with which the companies joined in the effort shows that they see the area as fundamentally important to broad searching across the Internet. A small amount of standardization would lead to greatly improved searching.

In his analysis, Gravano viewed the information on the Internet as a large number of collection of materials, each organized differently and each with its own search engine. The fundamental concept is to enable clients to discover broad characteristics of the search engines and the collections that they maintain. The challenge is that the search engines are different and the collections have different characteristics. The difficulty is not simply that the interfaces have varying syntaxes, so that a query has to be re-formulated to be submitted to different systems. The underlying algorithms are fundamentally different. Some use Boolean methods; others have methods of ranking results. Search engines that return a ranked list give little indication how the ranks were calculated. Indeed, the ranking algorithm is often a trade secret. As a result it is impossible to merge ranking lists from several sources into a single, overall list with sensible ranking. The ranking are strongly affected by the words used in a collection, so that, even when two sources use the same ranking algorithm, merging them is fraught with difficulty. The STARTS protocol enables the search engines to report characteristics of their collections and the ranks that they generate, so that a client program can attempt to combine results from many sources.

Beyond searching

Information discovery is more than searching. Most individuals use some combination of browsing and systematic searching. Chapter 10 discussed the range of requirements that users have when looking for information and the difficulty of evaluating the effectiveness of information retrieval in an interactive session with the user in the loop. All these problems are aggravated with distributed digital libraries.

Browsing has always been an important way to discover information in libraries. It can be as simple as going to the library shelves to see what books are stored together. A more systematic approach is to begin with one item and then move to the items that it refers to. Most journal articles and some other materials include list of references to other materials. Following these citations is an essentially part of research, but is a tedious task when the materials are physical objects that must be retrieved one at a time. With hyperlinks, following references becomes straightforward. A gross generalization is that following links and references is easier in digital libraries, but the quality of catalogs and indexes is higher in traditional libraries. Therefore, browsing is likely to be relatively more important in digital libraries.

If people follow a heuristic combination of browsing and searching, using a variety of sources and search engines, what confidence can they have in the results? This chapter has already seen the difficulties of comparing results obtained from searching different sets of information and deciding whether two items found in different sources are duplicates of the same information. For the serious user of a digital library there is a more subtle but potentially more serious problem. It is often difficult to know how comprehensive a search is being carried out. A user who searches a central database, such as the National Library of Medicine's Medline system, can be confident of searching every record indexed in that system. Contrast this with a distributed search of a large number of datasets. What is the chance of missing important information because one dataset is behind the others in supplying indexing information, or fails to reply to a search request?

Overall, distributed searching epitomizes the current state of digital libraries. From one viewpoint, every technique has serious weaknesses, the technical standards have not emerged, the understanding of user needs is embryonic, and organizational difficulties are pervasive. Yet, at the same time, enormous volumes of material are accessible on the Internet, web search programs are freely available, federations and commercial services are expanding rapidly. By intelligent combination of searching and browsing, motivated users can usually find the information they seek.

Chapter 12
Object models, identifiers, and structural metadata

Materials in digital library collections

Information comes in many forms and formats, each of which must be captured, stored, retrieved, and manipulated. Much of the early development of digital libraries concentrated on material that has a direct analog to some physical format. These materials can usually be represented as simply structured computer files. Digital libraries can go far beyond such simple digital objects; they include anything that can be represented in a digital format. The digital medium allows for new types of library objects such as software, simulations, animations, movies, slide shows, and sound tracks, with new ways to structure material. Computing has introduced its own types of object: spread sheets, databases, symbolic mathematics, hypertext, and more. Increasingly, computers and networks support continuous streams of digital information, notably speech, music, and video. Even the simplest digital objects may come in many versions and be replicated many times.

Methods for managing this complexity fall into several categories: **identifiers** for digital objects, **data types** which specify what the data represents, and **structural metadata** to represent the relationship between digital objects and their component parts. In combination, these techniques create an **object model**, a description of some category of information that enables computer systems to store and provide access to complex information. In looking at these topics, interoperability and long term persistence are constant themes. Information in today's digital libraries must be usable many years from now, using computer systems that have not yet been imagined.

Works, expressions, manifestations, and items

Users of a digital library usually want to refer to items at a higher level of abstraction than a file. Common English terms, such as "report", "computer program", or "musical work" often refer to many digital objects that can be grouped together. The individual objects may have different formats, minor differences of content, different usage restrictions, and so on, but usually the user considers them as equivalent. This requires a conceptual model that it able to describe content at various levels of abstraction.

Chapter 1 mentioned the importance of distinguishing between the underlying intellectual work and the individual items in a digital library, and the challenges of describing such differences in a manner that makes sense for all types of content. In a 1998 report, an IFLA study of the requirements for bibliographic records proposed the following four levels for describing content.

- **Work.** A work is the underlying abstraction, such as *The Iliad*, Beethoven's Fifth Symphony, or the Unix operating system.

- **Expression.** A work is realized through an expression. Thus, *The Iliad* was first expressed orally, then it was written down as a fixed sequence of words. A musical work can be expressed as a printed score or by any one of many performances. Computer software, such as Unix, has separate expressions as source code and machine code.

- **Manifestation.** A expression is given form in one or more manifestations. The text of *The Iliad* has been manifest in numerous manuscripts and printed books. A musical performance can be distributed on CD, or broadcast on television. Software is manifest as files, which may be stored or transmitted in any digital medium.

- **Item.** When many copies are made of a manifestation, each is a separate item, such as a specific copy of a book or computer file.

Clearly, there are many subtleties buried in these four levels. The distinctions between versions, editions, translations, and other variants are a matter for judgment, not for hard rules. Some works are hard to fit into the model, such as jazz music where each performance is a new creation, or paintings where the item is the work. Overall, however, the model holds great promise as a framework for describing this complicated subject.

Expressions

Multimedia

Chapters 9, 10, and 11 paid particular attention to works that are expressed as texts. Textual materials have special importance in digital libraries, but object models have to support all media and all types of object. Some non-textual objects are files of fixed information. They include the digital equivalents of familiar objects, such as maps, audio recordings and video, and other objects where the user is provided with a direct rendering of the stored form of a digital object. Even such apparently straightforward materials have many subtleties when the information is some medium other than text. This sections looks at three examples. The first is Panel 12.1which describes the Alexandria Digital Library of geospatial information.

Panel 12.1
Geospatial collections: the Alexandria library

The Alexandria Digital Library at the University of California, Santa Barbara is led by Terence Smith. It was one of the six projects funded by the Digital Libraries Initiative from 1994 to 1998. The collections in Alexandria cover any data that is referenced by a geographical footprint. This includes all classes of terrestrial maps, aerial and satellite photographs, astronomical maps, databases, and related textual information. The project combines a broad program of research with practical implementation at the university's map library.

These collections have proved to be a fertile area for digital libraries research. Geospatial information is of interest in many fields: cartography, urban planning, navigation, environmental studies, agriculture, astronomy, and more. The data comes from many sources: survey data, digital photographs, printed maps, and analog images. A digital library of these collections faces many of the same issues as a digital library of textual documents, but forces the researchers to examine every topic again to see which standard techniques can be used and which need to be modified.

Information retrieval and metadata

With geospatial data, information retrieval concentrates on coverage or scope, in contrast with other categories of material, where the focus is on subject matter or bibliographic information such as author or title. Coverage defines the geographical area covered, such as the city of Santa Barbara or the Pacific Ocean. Scope describes the varieties of information, such as topographical features, political boundaries, or population density. Alexandria provides several methods for capturing such information and using it for retrieval.

Coordinates of latitude and longitude provide basic metadata for maps and for geographical features. Systematic procedures have been developed for capturing such data from existing maps. A large gazetteer has been created which contains millions of place names from around the world; it forms a database and a set of procedures that translate between different representations of geospatial references, e.g., place names, geographic features, coordinates, postal codes, and census tracts. The Alexandria search engine is tailored to the peculiarities of searching for place names. Researchers are making steady progress at the difficult task of feature extraction, using automatic programs to identify objects in aerial photographs or printed maps, but this is a topic for long-term research.

Computer systems and user interfaces

Digitized maps and other geospatial information create large files of data. Alexandria has carried out research in applying methods of high-performance computing to digital libraries. Wavelets are a method that the library has exploited for both storage and in user interfaces. They provide a multi-level decomposition of an image, in which the first level is a small coarse image that can be used as a thumbnail. Extra levels provide greater detail at the expense of larger volumes of data.

Alexandria has developed several user interfaces, each building on earlier experience. Common themes have been the constraints of the small size of computer displays and the slow performance of the Internet in delivering large files to users. Good design helps to mitigate the first, though it is impossible to fit a large image and a comprehensive search interface onto a single screen. To enhance performance, every attempt is made never to transmit the same information twice. The user interfaces retain state throughout a session, so that user can leave the system and return to the same place without having to repeat any steps.

The next example of a format other than text looks at searching **video collections**. Panel 12.2 describes how the Informedia digital library has combined several methods of information retrieval to build indexes automatically and search video segments. Individually, each method is imprecise, but in combination results are achieved for indexing and retrieval that are substantially better than could be obtained from any single method. The team use the term "multi-modal" to describe this combination of methods.

Panel 12.2
Informedia: multi-modal information retrieval

Chapter 8 introduced the Informedia digital library of segments of digitized video and described some of the user interface concepts. Much of the research work of the project aims at indexing and searching video segments, with no assistance from human indexers or catalogers.

The multi-modal approach to information retrieval

The key word in Informedia is "multi-modal". Many of the techniques used, such as identifying changes of scene, use computer programs to analyze the video materials for clues. They analyze the video track, the sound track, the closed captioning if present, and any other information. Individually, the analysis of each mode gives imperfect information but combining the evidence from all can be surprisingly effective.

Informedia builds on a number of methods from artificial intelligence, such as speech recognition, natural language processing, and image recognition. Research in these fields has been carried out by separate research projects; Informedia brings them together to create something that is greater than the sum of its parts.

Adding material to the library

The first stage in adding new materials to the Informedia collection is to take the incoming video material and to divide it into segments by topics. The computer program uses a variety of techniques of image and sound processing to look for clues as to when one topic ends and another begins. For example, with materials from broadcast television, the gap intended for advertisements often coincides with a change of topic.

The next stage is to identify any text associated with the segment. This is obtained by speech recognition on the sound track, by identifying any captions within the video stream, and from closed captioning if present. Each of these inputs is prone to error. The next phase is to process the raw text with a variety of tools from natural language processing to create an approximate index record that is loaded into the search system.

Speech recognition

The methods of information discovery discussed in Chapters 10 and 11 can be applied to audio material, such as audio tapes and the sound track of video, if the spoken word can be converted to computer text. This conversion proves to be a tough computing problem, but steady progress has been made over the years, helped by ever increasing computer power.

Informedia has to tackle some of the hardest problems in speech recognition, including speaker independence, indistinct speech, noise, unlimited vocabulary, and accented speech. The computer program has to be independent of who is speaking. The speech on the sound track may be indistinct, perhaps because of background noise of music. It may contain any word in the English language including proper nouns, slang, and even foreign words. Under these circumstances, even a human listener misses some words. Informedia successfully recognizes about 50 to 80 percent of the words, depending on the characteristics of the specific video segment.

Searching the Informedia collection

To search the Informedia collection, a user provides a query either by typing or by speaking it aloud to be processed by the speech recognition system. Since there may be errors in the recognition of a spoken query and since the index is known to be built from inexact data, the information retrieval uses a ranking method that identifies the best apparent matches. The actual algorithm is based on the same research as the Lycos web search program and the index uses the same underlying retrieval system.

The final example in this section looks at the problems of delivering **real-time information**, such as sound recordings, to users. Digitized sound recordings are an example of continuous streams of data, requiring a special method of dissemination, so that the data is presented to the user at the appropriate pace. Sound recordings are on the boundary of what can reasonably be transmitted over the Internet as it exists today. User interfaces have a choice between real-time transmission, usually of indifferent quality, and batch delivery, requiring the user to wait for higher quality sound to be transmitted more slowly. Panel 12.3 describes RealAudio, one way to disseminate low-quality sound recordings within the constraints of today's Internet.

Panel 12.3
RealAudio

One hour of digitized sound of CD quality requires 635 megabytes of storage if uncompressed. This poses problems for digital libraries. The storage requirements for any substantial collection are huge and transmission needs high-speed networks. Uncompressed sound of this quality challenges even links that run at 155 megabits/second. Since most local area networks share Ethernets that run at less than a tenth of this speed and dial-up links are much slower, some form of compression is needed.

RealAudio is a method of compression and an associated protocol for transmitting digitized sound. In RealAudio format, one hour of sound requires about 5 megabytes of storage. Transmission uses a streaming protocol between the repository where the information is stored and a program running on the user's computer. When the user is ready, the repository transmits a steady sequence of sound packets. As they arrive at the user's computer, they are converted to sound and played by the computer. This is carried out at strict time intervals. There is no error checking. If a packet has not arrived when the time to play it is reached, it is ignored and the user hears a short gap in the sound.

This process seems crude, but, if the network connection is reasonably clear, the transmission of spoken sounds in RealAudio is quite acceptable when transmitted over dial-up lines at 28.8 thousand bits per second. An early experiment with RealAudio was to provide a collection of broadcast segments from the programs of National Public Radio.

The service uses completely standard web methods, except in two particulars, both of which are needed to transmit audio signals over the Internet in real time. The first is that the user's browser must accept a stream of audio data in RealAudio format. This requires adding a special player to the browser, which can be downloaded over the Internet. The second is that, to achieve a steady flow of data, the library sends data using the UDP protocol instead of TCP. Since some network security systems do not accept UDP data packets, RealAudio can not be delivered everywhere.

Dynamic and complex objects

Many of the digital objects that are now being considered for digital library collections can not be represented as static files of data.

- **Dynamic objects.** Dynamic or active library objects include computer programs, Java applets, simulations, data from scientific sensors, or video games. With these types of object, what is presented to the user depends upon the execution of computer programs or other external activities, so that the user gets different results every time the object is accessed.

- **Complex objects.** Library objects can be made up from many inter-related elements. These elements can have various relationships to each other. They can be complementary elements of content, such as the audio and picture channels of a video recording. They can be alternative manifestations, such as a high-resolution or low-resolution satellite image, or they can be surrogates, such as data and metadata. In practice these distinctions are often blurred. Is a thumbnail photograph an alternative manifestation, or is it metadata about a larger image?

- **Alternate disseminations.** Digital objects may offer the user a choice of access methods. Thus a library object might provide the weather conditions at San Francisco Airport. When the user accesses this object, the information returned might be data, such as the temperature, precipitation, wind speed and direction, and humidity, or it might be a photograph to show cloud cover. Notice that this information might be read directly from sensors, when requested, or from tables that are updated at regular intervals.
- **Databases.** A database comprises many alternative records, with different individuals selected each time the database is accessed. Some databases can be best thought of as complete digital library collections, with the individual records as digital objects within the collections. Other databases, such as directories, are library objects in their own right.

The methods for managing these more general objects are still subjects for debate. Whereas the web provides a unifying framework that most people use for static files, there is no widely accepted framework for general objects. Even the terminology is rife with dispute. A set of conventions that relate the intellectual view of library materials to the internal structure is sometimes called a "document model", but, since it applies to all aspects of digital libraries, "object model" seems a better term.

Identification

The first stage in building an object model is to have a method to identify the materials. The identifier is used to refer to objects in catalogs and citations, to store and access them, to provide access management, and to archive them for the long term. This sounds simple, but identifiers must meet requirements that overlap and frequently contradict each other. Few topics in digital libraries cause as much heated discussion as names and identifiers.

One controversy is whether semantic information should be embedded in a name. Some people advocate completely semantic names. An example is the Serial Item and Contribution Identifier standard (**SICI**). By a precisely defined set of rules, a SICI identifies either an issue of a serial or an article contained within a serial. It is possible to derive the SICI directly from a journal article or citation. This is a daunting objective and the SICI succeeds only because there is already a standard for identifying serial titles uniquely. The following is a typical SICI; it identifies a journal article published by John Wiley & Sons:

0002-8231(199601)47:1<23:TDOMII>2.0.TX;2-2

Fully semantic names, such as SICIs, are inevitably restricted to narrow classes of information; they tend to be long and ugly because of the complexity of the rules that are used to generate them. Because of the difficulties in creating semantic identifiers for more general classes of objects, compounded by arguments over trademarks and other names, some people advocate the opposite: random identifiers that contain no semantic information about who assigned the name and what it references. Random strings used as identifiers can be shorter, but without any embedded information they are hard for people to remember and may be difficult for computers to process.

In practice, many names are mnemonic; they contain information that makes them easy to remember. Consider the name "www.apple.com". At first glance this appears to be a semantic name, the web site of a commercial organization called Apple, but this is just an informed guess. The prefix "www" is conventionally used for web sites, but this is merely a convention. There are several commercial organizations called Apple and the name gives no hint whether this web site is managed by Apple Computer or some other company.

Another difficulty is to decide what a name refers to: work, expression, manifestation, or item. As an example, consider the International Standard Book Number (ISBN). This was developed by publishers and the book trade for their own use. Therefore ISBNs distinguish separate products that can be bought or sold; a hard back book will usually have a different ISBN from a paper back version, even if the contents are identical. Libraries, however, may find this distinction to be unhelpful. For bibliographic purposes, the natural distinction is between versions where the content differs, not the format. For managing a collection or in a rare book library, each individual copy is distinct and needs its own identifier. There is no universal approach to naming that satisfies every need.

Domain names and Uniform Resource Locators (URL)

The most widely used identifiers on the Internet are domain names and Uniform Resource Locators (URLs). They were introduced in Chapter 2. Panel 12.4 gives more information about domain names and how they are allocated.

Panel 12.4
Domain names

The basic purpose of domain names is to identify computers on the Internet by name, rather than all references being by IP address. An advantage of this approach is that, if a computer system is changed, the name need not change. Thus the domain name "library.dartmouth.edu" was assigned to a series of computers over the years, with different IP addresses, but the users were not aware of the changes.

Over the years, additional flexibility has been added to domain names. A domain name need no longer refer to a specific computer. Several domain names can refer to the same computer, or one domain name can refer to a service that is spread over a set of computers.

The allocation of domain names forms a hierarchy. At the top are the root domain names. One set of root names are based on types of organization, such as:

.com	commercial
.edu	educational
.gov	government
.net	network services
.org	other organizations

There is a second series of root domains, based on geography. Typical examples are:

ca	Canada
.jp	Japan
.nl	Netherlands

Organizations are assigned domain names under one of these root domains. Typical examples are:

cmu.edu	Carnegie Mellon University
elsevier.nl	Elsevier Science
loc.gov	Library of Congress
dlib.org	D-Lib Magazine

Historically, in the United States, domain names have been assigned on a first-come, first-served basis. There is a small annual fee. Minimal controls were placed on who could receive a domain name, and what the name might be. Thus anybody could register the name "pittsburgh.net", without any affiliation to the City of Pittsburgh. This lack of control has led to a proliferation of inappropriate names, resulting in trademark disputes and other arguments.

URLs extend the concept of domain names in several directions, but all are expansions of the basic concept of providing a name for a location on the Internet. Panel 12.5 describes some of the information that can be contained in a URL.

Panel 12.5
Information contained in a Uniform Resource Locator (URL)

The string of characters that comprises a URL is highly structured. A URL combines the specification of a protocol, a file name, and options that will be used to access the file. It can contain the following.

Protocols. The first part of a full URL is the name of a protocol or service ending with a colon. Typical examples are *http:*, *mailto:*, and *ftp:*.

Absolute and relative URLs. A URL can refer to a file by its domain name or its location relative to another file. If the protocol is followed by "//", the URL contains a full domain name, e.g.,

http://www.dlib.org/figure.jpg

Otherwise, the address is relative to the current directory. For example, within an HTML page, the anchor,

<a href="figure.jpg">

refers to a file "figure.jpg" in the same directory.

Files. A URL identifies a specific file on a specified computer system. Thus, in the URL,

http://www.dlib.org/contents.html

"www.dlib.org" is a domain name that identifies a computer on the Internet; "contents.html" is a file on that computer.

Port. A server on the Internet may provide several services running concurrently. The TCP protocol provides a "port" which identifies which service to use. The port is specified as a colon followed by a number at the end of the domain name. Thus, the URL,

http://www.dlib.org:80/index.html

references the port number 80. Port 80 is the default port for the HTTP protocol and therefore could be omitted from this particular URL.

Parameters. A variety of parameters can be appended to a URL, following either a "#" or "?" sign. These are passed to the server when the file is accessed.

URLs have proved extremely successful. They permit any number of versatile applications to be built on the Internet, but they pose a long-term problem for digital libraries. The problem is the need for persistence. Users of digital libraries wish to be able to access material consistently over long periods of time. URLs identify resources by a location derived from a domain name. If the domain name no longer exists, or if the resource moves to a different location, the URL is no longer valid.

A famous example of the problem comes from the early days of the web. At the beginning, the definitive documentation was maintained at CERN in Geneva. When the World Wide Web Consortium was established at M.I.T. in 1994, this documentation was transferred. Every hyperlink that pointed to CERN was made invalid. In such instances, the convention is to leave behind a web page stating that the site has moved, but forwarding addresses tend to disappear with time or become long chains. If a domain name is canceled, perhaps because a company goes out of business, all URLs based on that domain name are broken for ever. There are various forms of aliases that can be used with domain names and URLs to ameliorate this problem, but they are tricks not solutions.

Persistent Names and Uniform Resource Names (URN)

To address this problem, the digital library community and publishers of electronic materials have become interested in persistent names. These are sometimes called Uniform Resource Names (URNs). The idea is simple. Names should be globally unique and persist for all time. The objective is to have names that can last longer than any software system that exists today, longer even than the Internet itself.

A persistent name should be able to reference any Internet resource or set of resources. One application of URNs is to reference the current locations of copies of an object, defined by a list of URLs. Another application is to provide electronic mail addresses that do not need to be changed when a person changes jobs or moves to a different Internet service provider. Another possibility is to provide the public keys of named services. In each of these applications, the URN is linked to data, which needs to have an associated data type, so that computer systems can interpret and process the data automatically. Panel 12.6 describes the handle system, which is a system to create and manage persistent names, and its use by publishers for digital object identifiers (**DOI**s).

Panel 12.6
Handles and Digital Object Identifiers

Handles are a naming system developed at CNRI as part of a general framework proposed by Robert Kahn of CNRI and Robert Wilensky of the University of California, Berkeley. Although developed independently from the ideas of URNs, the two concepts are compatible and handles can be considered the first URN system to be used in digital libraries. The handle system has three parts:

- A name scheme that allows independent authorities to create handle names with confidence that they are unique.

- A distributed computer system that stores handles along with data that they reference, e.g., the locations where material is stored. A handle is resolved by sending it to the computer system and receiving back the stored data.
- Administrative procedures to ensure high-quality information over long periods of time.

Syntax

Here are two typical handles:

hdl:cnri.dlib/magazine
hdl:loc.music/musdi.139

These strings have three parts. The first indicates that the string is of type hdl:. The next, "cnri.dlib" or "loc.music", is a naming authority. The naming authority is assigned hierarchically. The first part of the name, "cnri" or "loc", is assigned by the central authority. Subsequent naming authorities, such as "cnri.dlib" are assigned locally. The final part of the handle, following the "/" separator, is any string of characters that are unique to the naming authority.

Computer system

The handle system offers a central computer system, known as the global handle registry, or permits an organization to set up a local handle service on its own computers, to maintain handles and provide resolution services. The only requirements are that the top-level naming authorities must be assigned centrally and that all naming authorities must be registered in the central service. For performance and reliability, each of these services can be spread over several computers and the data can be automatically replicated. A variety of caching services are provided as are plug-ins for web browsers, so that they can resolve handles.

Digital Object Identifiers

In 1996, an initiative of the Association of American Publishers adopted the handle system to identify materials that are published electronically. These handles are called Digital Object Identifiers (DOI). This led to the creation of an international foundation which is developing DOIs further. Numeric naming authorities are assigned to publishers, such as "10.1006" which is assigned to Academic Press. The following is the DOI of a book published by Academic Press:

doi:10.1006/0121585328

The use of numbers for naming authorities reflects a wish to minimize the semantic information. Publishers frequently reorganize, merge, or transfer works to other publishers. Since the DOIs persist through such changes, they should not contain the name of the original publisher in a manner that might be confusing.

Computer systems for resolving names

Whatever system of identifiers is used, there must be a fast and efficient method for a computer on the Internet to discover what the name refers to. This is known as resolving the name. Resolving a domain name provides the IP address or addresses of the computer system with that name. Resolving a URN provides the data associated with it.

Since almost every computer on the Internet has the software needed to resolve domain names and to manipulate URLs, several groups have attempted to build systems for identifying materials in digital libraries that use these existing mechanisms. One approach is OCLC's PURL system. A **PURL** is a URL, such as:

http://purl.oclc.org/catalog/item1

In this identifier, "purl.oclc.org" is the domain name of a computer that is expected to be persistent. On this computer, the file "catalog/item1" holds a URL to the location where the item is currently stored. If the item is moved, this URL must be changed, but the PURL, which is the external name, is unaltered.

PURLs add an interesting twist to how names are managed. Other naming systems set out to have a single coordinated set of names for a large community, perhaps the entire world. This can be considered a top-down approach. PURLs are bottom-up. Since each PURL server is separate, there is no need to coordinate the allocation of names among them. Names can be repeated or used with different semantics, depending entirely on the decisions made locally. This contrasts with the Digital Object Identifiers, where the publishers are building a single set of names that are guaranteed to be unique.

Structural metadata and object models

Data types

Data types are structural metadata which is used to describe the different types of object in a digital library. The web object model consist of hyperlinked files of data, each with a data type which tells the user interface how to render the file for presentation to the user. The standard method of rendering is to copy the entire object and render it on the user's computer. Chapter 2 introduced the concept of data type and discussed the importance of MIME as a standard for defining the type of files that are exchanged by electronic mail or used in the web. As described in Panel 12.7, MIME is a brilliant example of a standard that is flexible enough to cover a wide range of applications, yet simple enough to be easily incorporated into computer systems.

Panel 12.7. MIME

The name MIME was originally an abbreviation for "Multipurpose Internet Mail Extensions". It was developed by Nathaniel Borenstein and Ned Freed explicitly for electronic mail, but the approach that they developed has proved to be useful in a wide range of Internet applications. In particular, it is one of the simple but flexible building blocks that led to the success of the web.

The full MIME specification is complicated by the need to fit with a wide variety of electronic mail systems, but for digital libraries the core is the concept that MIME calls "Content-Type". MIME describes a data type as three parts, as in the following example:

Content-Type: text/plain; charset = "US-ASCII"

The structure of the data type is a *type* ("text"), a *subtype* ("plain"), and one or more optional parameters. This example defines plain text using the ASCII character set. Here are some commonly used types:

text/plain
text/html

image/gif

image/jpeg
image/tiff

audio/basic
audio/wav

video/mpeg
video/quickdraw

The *application* type provides a data type for information that is to be used with some application program. Here are the MIME types for files in PDF, Microsoft Word, and PowerPoint formats:

application/pdf
application/msword
application/ppt

Notice that *application/msword* is not considered to be a text format, since a Microsoft Word file may contain information other than text and requires a specific computer program to interpret it.

These examples are official MIME types, approval by the formal process for registering MIME types. It is also possible to create unofficial types and subtypes with names beginning "x-", such as "audio/x-pn-realaudio".

Lessons from MIME

MIME's success is a lesson on how to turn a good concept into a widely adopted system. The specification goes to great lengths to be compatible with the systems that preceded it. Existing Internet mail systems needed no alterations to handle MIME messages. The processes for checking MIME versions, for registering new types and subtypes, and for changing the system were designed to fit naturally within standard Internet procedures. Most importantly, MIME does not attempt to solve all problems of data types and is not tightly integrated into any particular applications. It provides a flexible set of services that can be used in many different contexts. Thus a specification that was designed for electronic mail has had its greatest triumph in the web, an application that did not exist when MIME was first introduced.

Complex objects

Materials in digital libraries are frequently more complex than files that can be represented by simple MIME types. They may be made of several elements with different data types, such as images within a text, or separate audio and video tracks; they may be related to other materials by relationships such as part/whole, sequence, and so on. For example, a digitized text may consist of pages, chapters, front matter, an index, illustrations, and so on. An article in an online periodical might be stored on a computer system as several files containing text and images, with complex links among them. Because digital materials are easy to change, different versions are created continually.

A single item may be stored in several alternate digital formats. When existing material is converted to digital form, the same physical item may be converted several times. For example, a scanned photograph may have a high-resolution archival version, a medium quality version, and a thumbnail. Sometimes, these formats are exactly equivalent and it is possible to convert from one to the other (e.g., an uncompressed image and the same image stored with a lossless compression). At other times, the different formats contain different information (e.g., differing representations of a page of text in SGML and PostScript formats).

Structural types

To the user, an item appears as a single entity and the internal representation is unimportant. A bibliography or index will normally refer to it as a single object in a digital library. Yet, the internal representation as stored within the digital library may be complex. Structural metadata is used to represent the various components and the relationships among them. The choice of structural metadata for a specific category of material creates an object model.

Different categories of object need different object models: e.g., text with SGML mark-up, web objects, computer programs, or digitized sound recordings. Within each category, rules and conventions describe how to organize the information as sets of digital objects. For example, specific rules describe how to represent a digitized sound recording. For each category, the rules describe how to represent material in the library, how the components are grouped as a set of digital objects, the internal structure of each, the associated metadata, and the conventions for naming the digital objects. The categories are distinguished by a **structural type**.

Object models for computer programs have been a standard part of computing for many years. A large computer program consists of many files of programs and data, with complex structure and inter-relations. This relationship is described in a separate data structure that is used to compile and build the computer system. For a Unix program, this is called a "make file".

Structural types need to be distinguished from **genres**. For information retrieval, it is convenient to provide descriptive metadata that describes the genre. This is the category of material considered as an intellectual work. Thus, genres of popular music include jazz, blues, rap, rock, and so on. Genre is a natural and useful way to describe materials for searching and other bibliographic purposes, but another object model is required for managing distributed digital libraries. While feature films, documentaries, and training videos are clearly different genres, their digitized equivalents may be encoded in precisely the same manner and processed identically; they are the same structural type. Conversely, two texts might be the same genre - perhaps they are both exhibition catalogs - but, if one is represented with SGML mark-up, and the other in PDF format, then they have different structural types and object models.

Panel 12.8 describes an object model for a scanned image. This model was developed to represent digitized photographs, but the same structural type can be used for any bit-mapped image, including maps, posters, playbills, technical diagrams, or even baseball cards. They represent different content, but are stored and manipulated in a computer with the same structure. Current thinking suggests that even complicated digital library collections can be represented by a small number of structural types. Less than ten structural types have been suggested as adequate for all the categories of material being converted by the Library of Congress. They include: digitized image, a set of pages images, a set of page images with associated SGML text, digitized sound recording, and digitized video recording.

Panel 12.9
An object model for scanned images

In joint work with the Library of Congress, a group from CNRI developed object models for some types of material digitized by the National Digital Library Program. The first model was for digitized images, such as scanned photographs. When each image is converted, several digital versions are made. Typically, there is a high-resolution archival version, a medium resolution version for access via the web, and one or more thumbnails. Each of these versions has its own metadata and there is other metadata that is shared by all the versions.

The first decision made in developing the object model was to keep separate the bibliographic metadata that describes the content. This is stored in separate catalog records. Other metadata is stored with the scanned images and is part of the object model. This includes structural metadata, which relates the various versions, and administrative metadata, which is used for access management.

Since there are occasions on which the individual versions may be used separately, each version is stored as a separate digital object. Each of these digital objects is identified by a handle and has two elements: a digitized image and metadata. Another digital object, known as a meta-object, is used to bring the versions together. The main purpose of the meta-object is to provide a list of the versions. It also has a data type field that describes the function of each image, which can be for reference, access, or as a thumbnail. The meta-object has its own identifier, a handle, and contains metadata that is shared by all the versions of the image.

Object models for interoperability

Object models are evolving slowly. How the various parts should be grouped together can rarely be specified in a few dogmatic rules. The decision depends upon the context, the specific objects, their type of content and sometimes the actual content. After the introduction of printing, the structure of a book took decades to develop to the form that we know today, with its front matter, chapters, figures, and index. Not surprisingly, few conventions for object models and structural metadata have yet emerged in digital libraries. Structural metadata is still at the stage where every digital library is experimenting with its own specifications, making frequent modifications as experience suggests improvements.

This might appear to pose a problem for interoperability, since a client program will not know the structural metadata used to store a digital object in an independent repository. However, clients do not need to know the internal details of how repositories store objects; they need to know the functions that the repository can provide.

Consider the storage of printed materials that have been converted to digital formats. A single item in a digital library collection consists of a set of page images. If the material is also converted to SGML, there will be the SGML mark-up, DTD, style sheets, and related information. Within the repository, structural metadata defines the relationship between the components, but a client program does not need to know these details. Chapter 8 looked at a user interface program that manipulates a sequence of page images. The functions include display a specified page, or go to the page with a specific page number, and so forth. A user interface that is aware of the functions supported for this structural type of information is able to present the material to the user, without knowing how it is stored on the repository.

Disseminations

In a digital library, the stored form of information is rarely the same as the form that is delivered to the user. In the web model of access, information is copied from the server to the user's computer where it is rendered for use by the user. This rendering typically takes the form of converting the data in the file into a screen image, using suitable fonts and colors, and embedding that image in a display with windows, menus, and icons.

Getting stored information and presenting it to the user of a digital library can be much more complicated than the web model. At the very least, a general architecture must allow that the stored information will be processed by computer programs on the server before being sent to the client or on the client before being presented to the user. The data that is rendered need not have been stored explicitly on the server as a file or files. Many servers run a computer program on a collection of stored data, extract certain information and provide it to the user. This information may be fully formatted by the server, but is often transmitted as a file that has been formatted in HTML, or some other intermediate format that can be recognized by the user's computer.

Two important types of dissemination are direct interaction between the client and the stored digital objects, and continuous streams of data. When the user interacts with the information, access by a user is not a single discrete event. A well-known example is a video game, but the same situation applies with any interactive material, such as a simulation of some physical situation. Access to such information consists of a series of interactions, which are guided by the user's control and the structural metadata of the individual objects.

Disseminations controlled by the client

Frequently, a client program may have a choice of disseminations. The selection might be determined by the equipment that a user has; when connected over a low-speed network, users sometimes choose to turn off images and receive only the text of web pages. The selection might be determined by the software on the user's computer; if a text is available in both HTML and PostScript versions, a user whose computer does not have software to view PostScript will want the HTML dissemination. The selection may be determined on non-technical grounds, by the user's wishes or convenience; a user may prefer to see a shortened version of a video program, rather than the full length version.

One of the purposes of object models is to provide the user with a variety of dissemination options. The aim of several research projects is to provide methods so that the client computer can automatically discover the range of disseminations that are available and select the one most useful at any given time. Typically, one option will be a default dissemination, what the user receives unless a special request is made. Another option is a short summary. This is often a single line of text, perhaps an author and title, but it can be fancier, such as a thumbnail image or a short segment of video. Whatever method is chosen, the aim of the summary is to provide the user with a small amount of information that helps to identify and describe the object. This topic is definitely the subject of research and there are few good examples of such systems in practical use.

Chapter 13
Repositories and archives

Repositories

This chapter looks at methods for storing digital materials in repositories and archiving them for the long term. It also examines the protocols that provide access to materials stored in repositories. It may seem strange that such important topics should be so late in the book, since long-term storage is central to digital libraries, but there is a reason. Throughout the book, the emphasis has been on what actually exists today. Research topics have been introduced where appropriate, but most of the discussion has been of systems that are used in libraries today. The topics in this chapter are less well established. Beyond the ubiquitous web server, there is little consensus about repositories for digital libraries and the field of digital archiving is new. The problems are beginning to be understood, but, particularly in the field of archiving, the methods are still embryonic.

A **repository** is any computer system whose primary function is to store digital material for use in a library. Repositories are the book shelves of digital libraries. They can be huge or tiny, storing millions of digital objects or just a single object. In some contexts a mobile agent that contains a few digital objects can be considered a repository, but most repositories are straightforward computer systems that store information in a file system or database and present it to the world through a well-defined interface.

Web servers

Currently, by far the most common form of repository is a web server. Panel 13.1 describes how they function. Several companies provide excellent web servers. The main differences between them are in the associated programs that are linked to the web servers, such as electronic mail, indexing programs, security systems, electronic payment mechanisms, and other network services.

Panel 13.1
Web servers

A **web server** is a computer program whose task is store files and respond to requests in HTTP and associated protocols. It runs on a computer connected to the Internet. This computer can be a dedicated web server, a shared computer which also runs other applications, or a personal computer that provides a small web site.

At the heart of a web server is a process called httpd. The letter "d" stands for "demon". A demon is a program that runs continuously, but spends most of its time idling until a message arrives for it to process. The HTTP protocol runs on top of TCP, the Internet transport protocol. TCP provides several addresses for every computer, known as ports. The web server is associated with one of these ports, usually port 80 but others can be specified. When a message arrives at this port, it is passed to the demon. The demon starts up a process to handle this particular message, and continues to listen for more messages to arrive. In this way, several messages can be processed at the same time, without tying up the demon in the details of their processing.

The actual processing that a web server carries out is tightly controlled by the HTTP protocol. Early web servers did little more than implement the get command. This command receives a message containing a URL from a client; the URL specifies a file which is stored on the server. The server retrieves this file and returns it to the client, together with its data type. The HTTP connection for this specific message then terminates.

As HTTP has added features and the size of web sites has grown, web servers have become more complicated than this simple description. They have to support the full set of HTTP commands, and extensions, such as CGI scripts. One of the requirements of web servers (and also of web browsers) is to continue to support older versions of the HTTP protocol. They have to be prepared for messages in any version of the protocol and to handle them appropriately. Web servers have steadily added extra security features which add complexity. Version 1.1 of the protocol also includes persistent connections, which permit several HTTP commands to be processed over a single TCP connection.

High-volume web servers

The biggest web sites are so busy that they need more than one computer. Several methods are used to share the load. One straightforward method is simply to replicate the data on several identical servers. This is convenient when the number of request is high, but the volume of data is moderate, so that replication is feasible. A technique called "DNS round robin" is used to balance the load. It uses an extension of the domain name system that allows a domain name to refer to a group of computers with different IP addresses. For example, the domain name "www.cnn.com" refers to a set of computers, each of which has a copy of the CNN web site. When a user accesses this site, the domain name system chooses one of the computers to service the request.

Replication of a web site is inconvenient if the volume of data is huge or if it is changing rapidly. Web search services provide an example. One possible strategy is to divide the processing across several computers. Some web search system use separate computers to carry out the search, assemble the page that will be returned to the user, and insert the advertisements.

For digital libraries, web servers provide moderate functionality with low costs. These attributes have led to broad acceptance and a basic level of interoperability. The web owes much of its success to its simplicity, and web servers are part of that success, but some of their simplifying assumptions cause problems for the implementers of digital libraries. Web servers support only one object model, a hierarchical file system where information is organized into separate files. Their processing is inherently stateless; each message is received, processed, and forgotten.

Advanced repositories

Although web servers are widely used, other types of storage systems are used as repositories in digital libraries. In business data processing, relational databases are the standard way to manage large volumes of data. Relational databases are based on an object model that consist of data tables and relations between them. These relations allow data from different tables to be joined or viewed in various ways. The tables and the data fields within a relational database are defined by a schema and a data dictionary. Relational databases are excellent at managing large amounts of data with a well-defined structure. Many of the large publishers mount collections on relational databases, with a web server providing the interface between the collections and the user.

Catalogs and indexes for digital libraries are usually mounted on commercial search systems. These systems have a set of indexes that refer to the digital objects. Typically, they have a flexible and sophisticated model for indexing information, but only a primitive model for the actual content. Many began as full-text systems and their greatest strength lies in providing information retrieval for large bodies of text. Some systems have added relevance feedback, fielded searching, and other features that they hope will increase their functionality and hence their sales.

Relational databases and commercial search systems both provide good tools for loading data, validating it, manipulating it, and protecting it over long terms. Access control is precise and they provide services, such as audit trails, that are important in business applications. There is an industry-wide trend for database systems to add full-text searching, and for search systems to provide some parts of the relational database model. These extra features can be useful, but no company has yet created a system that combines the best of both approaches.

Although some digital libraries have used relational databases with success, the relational model of data, while working well with simple data structures, lacks flexibility for the richness of object models that are emerging. The consensus among the leading digital libraries appears to be that more advanced repositories are needed. A possible set of requirements for such a repository are as follows.

- **Information hiding.** The internal organization of the repository should be hidden from client computers. It should be possible to reorganize a collection, change its internal representation, or move it to a different computer without any external effect.

- **Object models.** Repositories need to support a flexible range of object models, with few restrictions on data, metadata, external links, and internal relationships. New categories of information should not require fundamental changes to other aspects of the digital library.

- **Open protocols and formats.** Clients should communicate with the repository through well-defined protocols, data types, and formats. The repository architecture must allow incremental changes of protocols as they are enhanced over time. This applies, in particular, to access management. The repository must allow a broad set of policies to be implemented at all levels of granularity and be prepared for future developments.

- **Reliability and performance.** The repository should be able to store very large volumes of data, should be absolutely reliable, and should perform well.

Metadata in repositories

Repositories store both data and metadata. The metadata can be considered as falling into the general classes of descriptive, structural, and administrative metadata. Identifiers may need to distinguish elements of digital objects as well as the objects themselves. Storage of metadata in a repository requires flexibility since there is a range of storage possibilities:

- Descriptive metadata is frequently stored in catalogs and indexes that are managed outside the repository. They may be held in separate repositories and cover material in many independent digital libraries. Identifiers are used to associate the metadata with the corresponding data.

- Structural and administrative metadata is often stored with each digital object. Such metadata can be actually embedded within the object.
- Some metadata refers to a group of objects. Administrative metadata used for access management may apply to an entire repository or a collection within a repository. Finding aids apply to many objects.
- Metadata may be stored as separate digital objects with links from the digital objects to which they apply. Some metadata is not stored explictly but is generated when required.

One of the uses of metadata is for interoperability, yet every digital library has its own ideas about the selection and specification of metadata. The Warwick Framework, described in Panel 13.2, is a conceptual framework that offers some semblance of order to this potentially chaotic situation.

Panel 13.2
The Warwick Framework

The Warwick Framework is an attempt at a general model that can represent the various parts of a complex object in a digital library. The genesis of the framework was some ideas that came out of a 1996 workshop at the University of Warwick in England.

The basic concept is aimed at organizing metadata. A vast array of metadata can apply to a single digital object, including descriptive metadata such as MARC cataloguing, access management metadata, structural metadata, and identifiers. The members of the workshop suggested that the metadata might be organized into packages. A typical package might be a Dublin Core metadata package or a package for geospatial data. This separation has obvious advantages for simplifying interoperability. If a client and a repository are both able to process a specific package type they are able to reach some level of interoperation, even if the other metadata packages that they support are not shared.

Subsequently, Carl Lagoze of Cornell University and Ron Daniel then at the Los Alamos National Laboratory took this simple idea and developed an elegant way of looking at all the components of a digital object. Their first observation is that the distinction between data and metadata is frequently far from clear. To use a familiar analog, is the contents page of a book part of the content or is it metadata about the content of the book? In the Warwick Framework, such distinctions are unimportant. Everything is divided into packages and no distinction is made between data and metadata packages.

The next observation is that not all packages need to be stored explicitly as part of a digital object. Descriptive metadata is often stored in a different repository as a record in a catalog or index. Terms and conditions that apply to many digital objects are often best stored in separate policy records, not embedded in each individual digital object. This separation can be achieved by allowing indirect packages. The package is stored wherever is most convenient, with a reference stored in the repository. The reference may be a simple pointer to a location, or it may invoke a computer program whose execution creates the package on demand.

Digital objects of the same structural type will usually be composed of a specific group of packages. This forms an object model for interoperability between clients and repositories for that type of digital object.

The Warwick Framework has not been implemented explicitly in any large-scale system, but the ideas are appealing. The approach of dividing information into well-defined packages simplifies the specification of digital objects and provides flexibility for interoperability.

Protocols for interoperability

Interoperability requires protocols that clients use to send messages to repositories and repositories use to return information to clients. At the most basic level, functions are needed that deposit information in a repository and provide access. The implementation of effective systems requires that the client is able to discover the structure of the digital objects, different types of objects require different access methods, and access management may require authentication or negotiation between client and repository. In addition, clients may wish to search indexes within the repository.

Currently, the most commonly used protocol in digital libraries is HTTP, the access protocol of the web, which is discussed in Panel 13.3. Another widely used protocol is Z39.50; because of its importance in information retrieval, it was described in Chapter 11.

Panel 13.3. HTTP

Chapter 2 introduced the HTTP protocol and described the get message type. A *get* message is an instruction from the client to the server to return whatever information is identified by the URL included in the message. If the URL refers to a process that generates data, it is the data produced by the process that is returned.

The response to a *get* command has several parts. It begins with a status, which is a three digit code. Some of these codes are familiar to users of the web because they are error conditions, such as 404, the error code returned when the resource addressed by the URL is not found. Successful status codes are followed by technical information, which is used primarily to support proxies and caches. This is followed by metadata about the body of the response. The metadata provides information to the client about the data type, its length, language and encoding, a hash, and date information. The client used this metadata to process the final part of the message, the response body, which is usually the file referenced by the URL.

Two other HTTP message types are closely related to *get*. A *head* message requests the same data as a get message except that the message body itself is not sent. This is useful for testing hypertext links for validity, accessibility, or recent modification without the need to transfer large files. The *post* message is used to extend the amount of information that a client sends to the server. A common use is to provide a block of data, such as for a client to submit an HTML form. This can then be processed by a CGI script or other application at the server.

The primary use of HTTP is to retrieve information from a server, but the protocol can also be used to change information on a server. A *put* message is used to store specified information at a given URL and a *delete* message is used to delete information. These are rarely used. The normal way to add information to a web server is by separate programs that manipulate data on the server directly, not by HTTP messages sent from outside.

Many of the changes that have been made to HTTP since its inception are to allow different versions to coexist and to enhance performance over the Internet. HTTP recognizes that many messages are processed by proxies or by caches. Later versions include a variety of data and services to support such intermediaries. There are also special message types: options which allows a client to request information about the communications options that are available, and trace which is used for diagnostic and testing.

Over the years, HTTP has become more elaborate, but it is still a simple protocol. The designers have done a good job in resisting pressures to add more and more features, while making some practical enhancements to improve its performance. No two people will agree exactly what services a protocol should provide, but HTTP is clearly one of the Internet's success stories.

Object-oriented programming and distributed objects

One line of research is to develop the simplest possible repository protocol that supports the necessary functions. If the repository protocol is simple, information about complex object types must be contained in the digital objects. (This has been called "SODA" for "smart object, dumb archives".)

Several advanced projects are developing architectures that use the computing concept of **distributed objects**. The word "object" in this context has a precise technical meaning, which is different from the terms "digital object" and "library object" used in this book. In modern computing, an object is an independent piece of computer code with its data, that can be used and reused in many contexts. The information within an object is encapsulated, so that the internals of the object are hidden. All that the outside world knows about a class of objects is a public interface, consisting of **methods**, which are operations on the object, and **instance data**. The effect of a particular method may vary from class to class; in a digital library a "render" method might have different interpretations for different classes of object.

After decades of development, object-oriented programming languages, such as C++ and Java, have become accepted as the most productive way to build computer systems. The driving force behind object-oriented programming is the complexity of modern computing. Object-oriented programming allows components to be developed and tested independently, and not need to be revised for subsequently versions of a system. Microsoft is a heavy user of object-oriented programming to develop its own software. Versions of its object-oriented environment are known variously as OLE, COM, DCOM, or Active-X. They are all variants of the same key concepts.

Distributed objects generalize the idea of objects to a networked environment. The basic concept is that an object executing on one computer should be able to interact with an object on another, through its published interface, defined in terms of methods and instance data. The leading computer software companies - with the notable exception of Microsoft - have developed a standard for distributed objects known as **CORBA**. CORBA provides the developers on distributed computing systems with many of the same programming amenities that object-oriented programming provides within a single computer.

The key notion in CORBA is an Object Request Broker (ORB). When an ORB is added to an application program, it establishes a client-server relationships between objects. Using an ORB, a client can transparently invoke a method on a server object, which might be on the same machine or across a network. The ORB intercepts the call; it finds an object that can implement the request, passes it the parameters, invokes its method, and returns the results. The client does not have to be aware of where the object is located, its programming language, its operating system, or any other system aspects that are not part of the object's interface. Thus, the ORB provides interoperability between applications on different machines in heterogeneous distributed environments.

Data hiding

Objects in the computing sense and digital objects in the library sense are different concepts, but they have features in common. The term **data hiding** comes from object-oriented programming, but applies equally to digital objects in libraries. When a client accesses information in a repository, it needs to known the interface that the repository presents to the outside world, but it does not need to know how it is stored in the repository. With a web server, the interface is a protocol (HTTP), an address scheme (URL), and a set of formats and data types. With other repositories it can expressed in the terminology of object-oriented programming. What the user perceives as a single digital object might be stored in the repository as a complex set of files, records in tables, or as active objects that are executed on demand.

Hiding the internal structure and providing all access through a clearly defined interface clearly simplifies interoperability. Clients benefit because they do not need to know the internal organization of repositories. Two repositories may choose to organize similar information in different manners. One may store the sound track and pictures of a digitized film as two separate digital objects, the other as a single object. A client program should be able to send a request to begin playback, unaware of these internal differences. Repositories benefit because internal reorganization is entirely a local concern. What the user sees as a single digital object may in fact be a page in HTML format with linked images and Java applets. With data hiding, it is possible to move the images to a different location or change to a new version of Java, invisibly to the outside.

Chapter 3 introduced two major projects that both offer users thumbnail images of larger images, JSTOR and the National Digital Library Program at the Library of Congress. The Library of Congress has decided to derive the thumbnail in advance and to store them as separate data. JSTOR does not store thumbnails. They are computed on demand from the stored form of the larger images. Each approach is reasonable. These are internal decisions that should be hidden from the user interface and could be changed later. External systems need to know that the repository can supply a thumbnail. They do not need to know how it is created.

Legacy systems

Conventional attempts at interoperability have followed a path based on agreed standards. The strategy is to persuade all the different organizations to decide on a set of technical and organizational standards. For complex system, such as digital libraries, this is an enormous undertaking. Standards are needed for networking, for data types, methods of identification, security, searching and retrieval, reporting errors, and exchanging payment. Each of these standards will have several parts describing syntax, semantics, error procedures, extensions, and so forth. If a complete set of standards could be agreed and every organization implemented them in full, then a wonderful level of interoperability might be achieved. In practice, the pace of standardization is slower than the rate of change of technology, and no organization ever completely integrates all the standards into its systems before it begins to change those systems to take advantage of new opportunities or to avoid new difficulties.

Hence, a fundamental challenge of interoperability is how can different generations of system work together. Older systems are sometime given the disparaging name **legacy systems**, but many older systems do a fine job. Future plans always need to accommodate the existing systems and commitments. Panel 13.4 describes research at Stanford University that accepts legacy systems for what they are and builds simple programs, known as proxies, that translate the external view that a system has into a neutral view. They call this approach the InfoBus.

Panel 13.4
The Stanford InfoBus

One of the projects funded by the Digital Libraries Initiative was at Stanford University, under the leadership of Hector Garcia-Molina, Terry Winograd, and Andreas Paepcke. They tackled the extremely challenging problem of interoperability between existing systems. Rather than define new standards and attempt to modify older systems, they accept the systems as they find them.

The basic approach is to construct Library Service Proxies which are CORBA objects representing online services. At the back-end, these proxies communicate with the services via whatever communication channel they provide. At the front, the proxies provide interfaces defined in terms of CORBA methods. For example, a client with a Z39.50 search interface might wish to search an online search service, such as Dialog. This requires two proxies. One translate between the Z39.50 search protocol and the InfoBus model. The second translates between the Dialog interface and the InfoBus model. By using this pair of proxies, the client can search Dialog despite their different interfaces.

Perhaps the most interesting InfoBus tools are those that support Z39.50. Stanford has developed a proxy that allows Z39.50 clients to interact with search services that do not support Z39.50. Users can submit searches to this proxy through any user interface that was designed to communicate with a Z39.50 server. The proxy forwards the search requests through the InfoBus to any of the InfoBus-accessible sources, even sources that do not support Z39.50. The proxy converts the results into a format that Z39.50 clients can understand. In a parallel activity, researchers at the University of Michigan implemented another proxy that makes all Z39.50 servers accessible on the InfoBus. The Stanford project also constructed proxies for HTTP and for web search systems, including, Lycos, WebCrawler, and Altavista, and for other web services, such as ConText, which is Oracle's document summarization tool.

Archiving

Archives are the raw material of history. In the United States, the National Archives and Records Administration has the task to keep records "until the end of the republic". Hopefully, this will be a long time. At the very least, archives must be prepared to keep library information longer than any computer system that exists today, and longer than any electronic or magnetic medium has ever been tested. Digital archiving is difficult. It is easier to state the issues that to resolve them and there are few good examples to use as exemplars. The foundation of modern work on digital archiving was the report of the Task Force on Archiving of Digital Information described in Panel 13.5.

Panel 13.5
Task Force on Archiving of Digital Information

The Task Force on Archiving of Digital Information was established in late 1994 by the Commission on Preservation and Access and the Research Libraries Group to study the problems of digital preservation. It was chaired by John Garrett, then at CNRI, and Donald Waters, of Yale University. The report was the first comprehensive account to look at these issues from a legal, economic, organizational, and technical viewpoint.

The report illustrates the dangers of technological obsolescence with some frightening examples. For example, the report noted that computer programs no longer exist to analyze the data collected by the New York Land Use and Natural Resources Inventory Project in the 1960s.

The task force explored the possibility of establishing a national system of archives as a basis for long term preservation. While it considers such a system desirable, perhaps even essential, it recognizes that it is not a panacea. The fundamental responsibility for maintaining information lies with the people and organizations who manage it. They need to consider themselves part of a larger system.

Technically, the report is important in stressing that archiving is more than simply copying bits from one aging storage medium to another; the methods to interpret and manipulate the bits must also be preserved. For this reason, the report considers that migration of information from one format to another and from one type of computer system to another is likely to be the dominant form of digital archiving.

The report is particularly interesting in its discussions of the legal issues associated with archiving. Much of the information that should be archived is owned by organizations who value the intellectual property, for financial or other reasons. They are naturally hesitant about providing copies for archives. Archives, on the other hand, may have to be assertive. When information is in danger of being lost, perhaps because a company is going out of business, they need to be able to acquire information for the public good.

The task force carried out its work in 1995. Its report remains the most complete study of the field. Although its financial calculations now appear dated, the analysis of the options and the choices available have not been seriously questioned.

Conventional archiving distinguishes between conservation which looks after individual artifacts, and preservation which retains the content even if the original artifact decays or is destroyed. The corresponding techniques in digital archiving are **refreshing**, which aims to preserve precise sequences of bits, and **migration**, which preserves the content at a semantic level, but not the specific sequences of bits. This distinction was first articulated by the Task Force on Archiving of Digital Information, which recommended a focus on migration as the basic technique of digital archiving.

Both refreshing and migration require periodic effort. Business records are maintained over long periods of time because a team of people is paid to maintain that specific data. It is their job and they pay attention to the inter-related issues of security, back-up, and long term availability of data. Publishers have also come to realize that their digital information is an asset that can generate revenue for decades and are looking after it carefully, but in many digital collections, nobody is responsible for preserving the information beyond its current usefulness. Some of the data may be prized centuries from now, but today it looks of no consequence. Archiving the data is low on everybody's priority and is the first thing to be cut when budgets are tight.

Storage

In the past, whether a physical artifact survived depended primarily of the longevity of it materials. Whether the artifacts were simple records from churches and governments, or treasures such as the Rosetta Stone, the Dead Sea Scrolls, the Domesday Book, and the Gutenberg Bibles, the survivors have been those that were created with material that did not perish, notably high-quality paper.

Of today's digital media, none can be guaranteed to last for long periods. Some, such as magnetic tape, have a frighteningly short life span before they deteriorate. Others, such as CDs are more stable, but nobody will predict their ultimate life. Therefore, unless somebody pays attention, all digital information will be lost within a few decades. Panel 13.6 describes some of the methods that are used to store digital materials today. Notice that the emphasis is on minimizing the cost of equipment and fast retrieval times, not on longevity.

Panel 13.6
Storing digital information

Storage media

The ideal storage medium for digital libraries would allow vast amounts of data to be stored at low cost, would be fast to store and read information, and would be exceptionally reliable and long lasting.

Rotating **magnetic disks** are the standard storage medium in modern computer systems. Sizes range from a fraction of a gigabyte to arrays of thousands of gigabytes. (A gigabyte is a thousand million bytes.) Disks are fast enough for most digital library applications, since data can be read from disks faster than it can be transmitted over networks. When data is read from a disk, there is a slight delay (about 15 milliseconds) while the disk heads are aligned to begin reading; then data is read in large blocks (typically about 25 megabytes per second). These performance characteristics suit digital library applications, which typically read large blocks of data at a time.

The decline in the cost of disks is one of the marvels of technology. Magnetic disks are coming down in price even faster than semi-conductors. In 1998, the price of disks was a few hundred dollars per gigabyte. The technology is advancing so rapidly that digital libraries can confidently plan that in ten years time the cost will be no more than five percent and very likely will be less than one percent of today's.

Disks have a weakness: reliability. The data on them is easily lost, either because of a hardware failure or because a program over-writes it. To guard against such failures, standard practice is for the data to be regularly copied onto other media, usually magnetic tape. It is also common to have some redundancy in the disk stores so that simple errors can be corrected automatically. Neither disks nor magnetic tape can be relied on for long term storage. The data is encoded on a thin magnetic film deposited on some surface. Sooner or later this film decays. Disks are excellent for current operations, but not for archiving.

Hierarchical stores

Large digital library collections are sometimes stored in hierarchical stores. A typical store has three levels, magnetic disks, optical disks, and magnetic tapes. The magnetic disks are permanently online. Information can be read in a fraction of a second. The optical disks provide a cheaper way of storing huge amounts of data, but the disk platters are stored in an automated silo. Before an optical disk can be used, a robot must move it from the silo to a disk reader, which is a slow process. The magnetic tapes are also stored in a silo with a robot to load them.

To the computers that use it, a hierarchical store appears to be a single coherent file system. Less frequently used data migrates from the higher speed, but more expensive magnetic disks to the slower but cheaper media. As the cost and capacity of magnetic disks continue to fall dramatically, the need for an intermediate level of storage becomes questionable. The magnetic disks and tapes serve distinctive functions and are both necessary, but the intermediate storage level may not by needed in future.

Compression

Digital libraries use huge amounts of storage. A page of ASCII text may be only a few thousand characters, but a scanned color image, only one inch square, requires more than a megabyte (one million bytes). An hour of digitized sound, as stored on a compact disk is over 600 megabytes and a minute of video can have more than one gigabyte of data before compression.

To reduce these storage requirements, large items are compressed, including almost all images, sound, and video. The basic idea of compression is simple, though the mathematics are complex. Digitized information contains redundancy. A page image has areas of white space; it is not necessary to encode every single pixel separately. Successive frames of video differ only slightly from each other; rather than encode each frame separately, it is simpler to record the differences between them.

Compression methods can be divided into two main categories. **Lossless** compression removes redundant information in a manner that is completely reversible; the original data can be reconstructed exactly as it was. **Lossy** compression can not be reversed; approximations are made during the compression that lose some of the information. In some applications, compression must be lossless. In a physics experiment, a single dot on an image may be crucial evidence; any modification of the image might undermine the validity of the experiment. In most applications, however, some losses are acceptable. The JPEG compression used for images and MPEG used for video are lossy methods that are calibrated to provide images that are very satisfactory to a human eye.

Compression methods reduce the size of data considerably, but the files are still large. After compression, a monochrome page of scanned text is more than 50,000 bytes. MPEG compression reduces digitized video from 20 or 30 megabytes per second to 10 megabytes per minute. Since digital libraries may store millions of these items, storage capacity is an important factor.

Replication and refreshing

Replication is a basic technique of data processing. Important data that exists only as a single copy on one computer is highly vulnerable. The hardware can fail; the data can be obliterated by faulty software; an incompetent or dishonest employee can remove the data; the computer building may be destroyed by fire, flooding, or other disaster. For these reasons, computer centers make routine copies of all data for back-up and store this data in safe locations. Good organizations go one step further and periodically consolidate important records for long term storage. One approach is to retain copies of financial and legal records on microform, since archival quality microform is exceptionally durable.

Because all types of physical media on which digital information is stored have short lives, methods of preservation require that the data is copied periodically onto new media. Digital libraries must plan to **refresh** their collections in the same manner. Every few years the data must be moved onto new storage media. From a financial viewpoint this is not a vast challenge. For the next few decades, computing equipment will continue to tumble in price while increasing in capacity. The equipment that will be needed to migrate today's data ten years from now will cost a few percent of the cost today and robots can minimize the labor involved.

Preserving content by migration

Assume (a big assumption) that the bits are systematically refreshed from media to media, so that the technical problem of preserving the raw date is resolved. The problems are just beginning. Digital information is useless unless the formats, protocols, and metadata can be recognized and processed. Ancient manuscripts can still be read because languages and writing have changed slowly over the years. Considerable expertise is needed to interpret old documents, but expertise has been passed down through generations and scholars can decipher old materials through persistence and inspiration.

Computing formats change continually. File formats of ten years ago may be hard to read. There is no computer in the world that can run programs for computers that were widespread a short time ago. Some formats are fairly simple; if, at some future date, an archeologist were to stumble onto a file of ASCII text, even if all knowledge of ASCII had been lost, the code is sufficiently simple that the text could probably be interpreted, but ASCII is an exception. Other formats are highly complex; it is hard to believe that anybody could ever decipher MPEG compression without a record of the underlying mathematics, or understand a large computer program from its machine code.

Therefore, in addition to the raw data, digital archiving must preserve ways to interpret the data, to understand its type, its structure, and its formats. If a computer program is needed to interpret the data, then the program must be preserved with some device that can execute it, or the data migrated to a different form. In the near term, it is possible to keep old computer systems for these purposes, but computers have a short life span. Sooner or later the computer will break down, spare parts will no longer be available, and any program that depends upon the computer will be useless. Therefore migration of the content becomes necessary.

Migration has been standard practice in data processing for decades. Businesses, such as pension funds, maintain records of financial transactions over many years. In the United States, the Social Security Administration keeps a record of payroll taxes paid on behalf of all workers throughout their careers. These records are kept on computers, but the computer systems are changed periodically. Hardware is replaced and software systems are revised. When these changes take place, the data is migrated from computer to computer, and from database to database. The basic principle of migration is that the formats and structure of the data may be changed, but the semantics of the underlying content is preserved.

Another method that is sometimes suggested by people with limited knowledge of computing is **emulation**; the idea is to specify in complete detail the computing environment that is required to execute a program. Then, at any time in the future, an emulator can be built that will behave just like the original computing environment. In a few, specialized circumstances this is a sensible suggestion. For example, it is possible to provide such a specification for a program that renders a simple image format, such as JPEG. In all other circumstances, emulation is a chimera. Even simple computing environments are much too complex to specify exactly. The exact combination of syntax, semantics, and special rules is beyond comprehension, yet subtle, esoteric aspects of a system are often crucial to correct execution.

Digital archeology

Societies go through periods of stress, including recessions, wars, political upheavals, and other times when migrating archival material is of low priority. Physical artifacts can lie forgotten for centuries in attics and storerooms, yet still be recovered. Digital information is less forgiving. Panel 13.7 describes how one such period of stress, the collapse of East Germany, came close to losing the archives of the state. It is an example of **digital archeology**, the process of retrieving information from damaged, fragmentary, and archaic data sources.

Panel 13.7
Digital archeology in Germany

An article in the New York Times in March 1998 provided an illustration of the challenges faced by archivists in a digital world unless data is maintained continuously from its initial creation.

In 1989, when the Berlin Wall was torn down and Germany was reunited, the digital records of East Germany were in disarray. The German Federal Archives acquired a mass of punched cards, magnetic disks, and computer tapes that represented the records of the Communist state. Much of the media was in poor condition, the data was in undocumented formats, and the computer centers that had maintained it were hastily shut down or privatized. Since then, a small team of German archivists has been attempting to reconstruct the records of East Germany. They call themselves "digital archeologists" and the term is appropriate.

The first problem faced by the digital archeologists was to retrieve the data from the storage media. Data on even the best quality magnetic tape has a short life and this tape was in poor condition, so that the archeologists could only read it once. In many cases the data was stored on tape following a Russian system that is not supported by other computers. Over the years, the archivists have obtained several of these computers, but some 30 percent of the data is unreadable.

When the data had been copied onto other media, the problems were far from solved. To save space, much of the data had been compressed in obscure and undocumented ways. An important database of Communist officials illustrates some of the difficulties. Since the computers on which the data had been written and the database programs were clones of IBM systems, recovering the database itself was not too difficult, but interpreting the data was extremely difficult, without documentation. The archivists had one advantage. They have been able to interview some of the people who built these databases, and have used their expertise to interpret much of the information and preserve it for history.

Dr. Michael Wettengel, the head of the German group of archivists, summarizes the situation clearly, "Computer technology is made for information processing, not for long-term storage."

Creating digital libraries with archiving in mind

Since digital archiving has so many risks, what can we do today to enhance the likelihood that the digital archeologist will be able to unscramble the bits? Some simple steps are likely to make a big difference. The first is to create the information in formats that are widely adopted today. This increases the chance that, when a format becomes obsolete, conversion programs to new formats will be available. For example, HTML and PDF are so widely used in industry that viewers will surely be available many years from now.

One interesting suggestion is to create an archive that contains the definition of formats, metadata standards, protocols and the other building blocks of digital libraries. This archive should be on the most persistent media that is known, e.g., paper or microfilm, and everything should be described in simple text. If the formats and encoding schemes are preserved, the information can still be recovered. Future digital archeologist may have a tough job, creating an interpreter that can resolve long-obsolete formats or instruction sets, but it can be done. Modern formats are complex. Whereas a digital archeologist might reverse engineer the entire architecture of an early IBM computer from a memory dump, the archeologist will be helpless with more complex materials, unless the underlying specification is preserved.

Perhaps the most important way that digital libraries can support archiving is through selection. Not everything needs to be preserved. Most information is intended to have a short life; much is ephemeral, or valueless. Publishers have always made decisions about what to publish and what to reject; even the biggest libraries acquire only a fraction of the world's output. Digital libraries are managed collections of information. A crucial part of that management is deciding what to collect, what to store, what to preserve for the future, and what to discard.

Chapter 14
Digital libraries and electronic publishing today

Digital libraries and electronic publishing are here. They are not an academic concept to debate, or a dream of utopia. In this last chapter, the temptation is to predict what lies ahead, but the history of computing shows how fruitless such predictions can be. Rather than attempt to predict the future, it is more profitable to celebrate the present.

This is not to ignore the future. The work that is in progress today will be in production soon. Not all new projects enter the mainstream, but understanding what is happening today helps comprehension of the potential that lies ahead. This book was written over a twelve month period in 1997 and 1998. Even during that period, digital libraries developed at a rapid pace. This final chapter looks at some of the current trends, not to forecast the future, but to understand the present.

A reason that predictions are so difficult is that, at present, the technology is more mature than the uses being made of it. In the past, forecasts of basic technology have proved to be reasonably accurate. Every year, semiconductors and magnetic devices are smaller, cheaper, faster, and with greater capacity. There are good engineering reasons to expect these trends to continue for the next five or ten years. Funds are already committed for the high-speed networks that will be available in a few years time.

Predictions about the new applications that will develop have always been much less accurate. Seminal applications such as spread sheets, desk-top publishing, and web browsers emerged from obscurity with little warning. Even if no such breakthrough takes place, forecasts about how the applications will be used and the social effect of new technology are even less dependable. The only reliable rule is that, when a pundit starts out with the words, "It is inevitable that," the statement is inevitably wrong.

A possible interpretation of the current situation is that digital libraries are at the end of an initial phase and about to begin a new one. The first phase can be thought of as a movement of traditional publications and library collections to digital networks. Online newspapers, electronic versions of scientific journals, and the conversion of historic materials all fall into this category. Fundamentally, they use new technology to enhance established types of information. If the thinking is correct, the next phase will see new types of collections and services that have no analog in traditional media. The forms that they will take are almost impossible to anticipate.

A myopic view of digital libraries

Looking back on any period of time, trends and key events become apparent that were not obvious at the time. Those of us who work in digital libraries can only have a myopic view of the field, but here are some observations about how the field of digital libraries appears to an insider.

A useful metaphor is the contrast between a rowing boat and a supertanker. A rowing boat can change direction quickly and accelerate to full speed, but has little momentum. The early digital library projects, such as our Mercury project at Carnegie Mellon University, were like rowing boats. They made fast progress, but when the initial enthusiasm ended or funding expired they lost their momentum. Established libraries, publishers, and commercial corporations are like supertankers. They move more deliberately and changing direction is a slow process, but once they head in a new direction they move steadily in that direction. In digital libraries and electronic publishing, success requires attention to the thousands of details that transform good ideas into practical services. As the Internet and the web mature, many organizations are making these long-term investments in library collections, electronic publications, and online services. The supertankers are changing direction.

Consider these two years, 1997 and 1998, as typical. During this short period, many developments matured that had been in progress for several years. Online versions of newspapers reached a high quality and the level of readership of some of them began to rival the readership of printed editions. In 1997 and 1998, major scientific publications first became available online, from both commercial and society publishers. It was also an important time for conversion projects in libraries, such as JSTOR and the Library of Congress, as the volume of materials available from these projects accelerated sharply. During 1997 and 1998, the Dublin Core approach to metadata and Digital Object Identifiers for electronic publications both gained momentum; they appear to have made the transition from the fringe to the mainstream.

On the technical front, products for automatic mirroring and caching reached the market, tools for web security became available, and the Java language at last became widely used. These developments were all begun in earlier years, none can be considered research, yet in aggregate they represent tremendous progress.

In the United States, electronic commerce on the Internet grew rapidly. It has become widely accepted to buy books, airline tickets, stock market investments, and automobiles through Internet transactions. The Internal Revenue Service now urges people to pay their income tax online. Congress passed a reasonable revision of the copyright law. Funds are available for entrepreneurial investments; markets for new products are open for the right idea. Internet stocks are the favorite speculation of the stock market. Low-cost personal computers, under $1,000, are selling vigorously, bringing online information to an ever broader range of people.

No year is perfect and pessimists can point out a few worrying events. During 1997, junk electronic mail reached an annoying level. Leading manufacturers released incompatible versions of the Java programming language. The United States policy on encryption continued to emulate an ostrich. These are short-term problems. Hopefully, by the time that this book is published, junk electronic mail will be controlled and Java will be standardized. (There appears to be less hope for the United States' encryption policy.)

For the next few years, incremental developments similar to those in 1997 in 1998 can be expected. They can be summarized succinctly. Large numbers of energetic people are exploiting the opportunities provided by the Internet to provide new products and services.

People

From a myopic viewpoint, it is easy to identify the individual activities. It is much harder to see the underlying changes of which they are part. In the long term, the most fundamental trend is perhaps the most difficult to measure in the short term. How are people's habits changing? Many people are writing for the web. Graphic designers are designing online materials. People are reading these materials. Who are these people? What would they be doing otherwise?

Habits clearly have changed. Visit a university campus or any organization that uses information intensively and it is obvious that people spend hours every week in front of computers, using online information. At home, there is evidence that part of the time people use the web is time that used to be spent watching television. At work, are people reading more, or are they substituting online information for traditional activities such as visits to the library? Ten years from now, when we look back at this period of change, the answers to these questions may be obvious. Today we can only hypothesize or extrapolate wildly from small amounts of data. Here are some guesses, based on personal observation and private hunches.

The excitement of online information has brought large numbers of new people into the field. People who would have considered librarianship dull and publishing too bookish, are enthralled by creating and designing online materials. The enthusiasm and energy that these newcomers bring is influencing the older professions more fundamentally than anybody might have expected. Although every group has its Luddites, many people are reveling in their new opportunities.

When the web began, Internet expertise was in such short supply that anybody with a modicum of skill could command a high salary. Now, although real experts are still in great demand, the aggregate level of skill is quite high. A sign of this change is the growth in program that help mid-career people to learn about the new fields. In the United States, every community college is running courses on the Internet and the web. Companies that provide computer training programs are booming. Specialist program in digital libraries are over-subscribed.

The new generation

In 1997, a Cornell student who was asked to find information in the library reportedly said, "Please can I use the web? I don't do libraries." More recently, a faculty member from the University of California at Berkeley mused that the term "digital library" is becoming a tautology. For the students that she sees, the Internet is the library. In the future, will they think of Berkeley's fine conventional libraries as physical substitutes for the real thing?

Are these the fringe opinions of a privileged minority or are they insights about the next generation of library users? Nobody knows. The data is fragmentary and often contradictory. The following statistics have been culled from a number of sources. They should be treated with healthy skepticism.

A survey in Pittsburgh found that 56 percent of people aged eighteen to twenty four used the Internet, but only 7 percent of those over fifty five are Internet users. Another poll, in 1997, found that 61 percent of American teenagers use the web. Although boys outnumber girls, the difference in usage is only 66 to 56 percent. In 1996, a third study found that 72 percent of children aged eight to twelve had spent time on a computer during the last month.

The network that these young people are beginning to use as their library had about 5.3 million computers in 1998. A careful study in late 1996 estimated that one million web site names were in common usage, on 450,000 unique host machines, of which 300,000 appear to be stable, with about 80 million HTML pages on public servers. Two years later, the number of web pages was estimated at about 320 million web pages. Whatever the exact numbers, everybody agrees that they are large and growing fast.

A pessimist would read these figures as a statement that, in the United States, the young people have embraced online information, the mid-career people are striving to convert to the new world, and the older people who make plans and control resources are obsolete. Observation, however, shows that this analysis is unduly gloomy. The fact that so many large organizations are investing heavily in digital information shows that at least some of the leaders embrace the new world.

Organizations

Although many organizations are investing in digital libraries and electronic publications, nobody can be sure what sort of organizations are likely to be most successful. In some circumstances, size may be an advantage, but small, nimble organizations are also thriving.

A 1995 article in The Economist described control of the Internet in terms of giants and ants. The giants are big companies, such as the telephones companies and the media giants. The ants are individuals; separately, each has tiny power, but in aggregate they have consistently succeeded in shaping the Internet, often in direct opposition to the perceived interests of the giants. In particular, the community of ants has succeeded in keeping the Internet and its processes open. During the past few years, both digital libraries and electronic publishing have seen consolidation in ever large organizational units. In libraries this is seen as the movement to consortia; in publishing it has been a number of huge corporate mergers. Yet, even as these giants have been formed, the energy of the ants has continued. At present, it appears that giants and ants can coexist and both are thriving. Meanwhile, some of the ants, such as Yahoo and Amazon.com, are becoming the new giants.

Collections and access

Two factors that will greatly influence the future of digital libraries are the rate at which well-managed collections become available on the Internet and the business models that emerge. Currently, materials are being mounted online at an enormous rate. The growth of the web shows no sign of slowing down.

The number of good, online sites is clearly growing. The sites run by newspapers and news agencies are fine examples of what is best and what is most vulnerable. There are many online news services, from the Sydney Morning Herald to the New York Times and CNN. These site provides up-to-date, well-presented news at no charge to the users. The readership probably exceeds any American newspaper. As a source of current information they are excellent, but ephemeral. The information is changed continually and at the end of day most of it disappears. Conventional libraries collect newspapers and store them for centuries, usually on microform. No library or archive is storing these web sites. Among standard library materials, it is hard to estimate which disciplines have the largest proportion of their material available through digital libraries. Large portions of the current scientific and technical literature are now available, as is much government information. Legal information has long been online, albeit at a steep price. Business and medical information are patchy. Public libraries play an important role in providing current information such as newspapers, travel timetables, job advertisements, and tax forms. These are mainly available online, usually with open access.

In many situations, current information is in digital form but not historic materials, though projects to convert traditional materials to digital format and mount them in digital libraries are flourishing. Libraries are converting their historic collections; publishers are converting their back-runs. Established projects are planning to increase their rate of conversion and new projects are springing up. Several projects have already converted more than a million pages. The first plan to convert a billion pages was made public in 1998.

Currently, the biggest gap is in commercial entertainment. Some of the major creators of entertainment - films, television, radio, novels, and magazines - have experimented with ways to use the Internet, but with little impact. Partly, their difficulty comes from the technical limitations of the Internet. Most people receive much better images from cable television than can be delivered over the network or rendered on moderate priced personal computers. Partly, the rate of change is dictated by business practices. Entertainment is big business and has not yet discovered how to use the Internet for its profit.

Open access appears to be a permanent part of digital libraries, but few services have yet discovered good business models for providing open access material. A few web sites make significant money from advertising, but most are supported by external funds. Digital libraries and electronic publishing require skilled professionals to create, organize, and manage information; they are expensive. Ultimately, an economic balance will emerge, with some collections open access and others paid for directly by their users, but it is not obvious yet what this balance will be.

Technology

The web technology, which has fueled so much recent growth, is maturing. During the middle years of the 1990s, the web developed so rapidly that people coined the term "web year" for a short period of time, packed with so much change that it seemed like a full calendar year, though in fact it was much shorter. As the web has matured, the pace of change in the technology has slowed down to the normal rate in computing. Every year brings incremental change and, in combination, over a few years these incremental changes are substantial, but the hectic pace has slowed down.

This does not mean that the growth in the web has ended, far from it. The number of web sites continues to grow rapidly. Busy sites report that the volume of usage is doubling every year. The quality of graphics and the standards of service improve steadily. For digital libraries and electronic publishing, several emerging technologies show promise: persistent names such as handles and Digital Object Identifiers, XML mark-up, the Resource Description Framework, and Unicode. The success of these technologies will depend upon the vagaries of the market place. Widespread acceptance of any or all would be highly beneficial to digital libraries.

Finally, the growth in performance of the underlying Internet remains spectacular. The past couple of years have seen a series of governmental and commercial initiatives that aim to provide leaps in performance, reliability, and coverage over the next few years. We can not predict how digital libraries will use this performance, but it provides remarkable opportunities.

Research and development

Digital libraries are now an established field of research, with the usual paraphernalia of workshops and conferences. There have even been attempts to establish printed journals about digital libraries. More importantly, at least one thousand people consider that their job is to carry out research in the field. The immediate impact of this work is difficult to evaluate, but it is clearly significant. Examples of projects funded by the NSF and DARPA appear throughout this book. Many of the more recent ones were funded explicitly as digital libraries research.

Panel 14.1
The Santa Fe workshop

In March 1997, the National Science Foundation sponsored a workshop in Santa Fe, New Mexico to discuss future research in digital libraries. It was part of a planning process that subsequently led to the announcement of a major new digital libraries research program, the Digital Libraries Initiative, Phase 2. The meeting was fascinating because it was an opportunity for the people who carry out research to describe how they saw the development of digital libraries and the research opportunities.

Many of the people at the meeting had been part of the first Digital Libraries Initiative or other federally funded research projects. Naturally, they were interested in continuing their research, but they did not simply recommend a continuation of the same programs. These early projects constructed digital library test collections and used them for research, mainly on technical topics. Some people at the workshop argued that a valuable use of government funds would be to build large digital libraries. Many people agreed that archiving is a vital topic, worthy of serious research.

Most of the discussion, however, was about making existing collections more usable. The people at the workshop are senior researchers. They need not only to find information, but also to escape from too much information. The discussions sometimes suggested that the central problem of digital libraries research is information overload. How can automatic methods be used to filter, extract, and consolidate information? The discussions embraced methods by which individuals can manage their private libraries or groups can carry out collaborative work over the Internet. Digital libraries are managed collections of information. How can they be managed for the convenience of the users?

Social, economic, and legal issues were also fully discussed. As ever, in these areas, the difficulty is how to articulate a coherent research strategy. While nobody denies the importance of these areas, there remains skepticism whether they can be tackled by large research projects.

Technical people at the meeting pointed out that digital libraries have become one of the major uses of supercomputers. In 1986, when the national supercomputing centers were established, the Internet backbone ran at 56 kbits/second, shared by all users. Today, this speed is provided by low-cost, dial-up modems used by individuals. Today's laptop computers have the performance of the supercomputers of twelve years ago. In twelve year's time ever smaller computers will have the performance of today's supercomputers.

After a few years, rapid incremental growth adds up to fundamental changes. At the Santa Fe workshop, we were asked to explore the assumptions about digital libraries that are so deeply rooted in our thinking that we take them for granted. By challenging such assumptions, the aim was to stimulate a creative agenda for the next generation of digital library research.

Footnote

Finally, here is a personal footnote. In writing this book, I looked at hundreds of sources. Most were primary material, descriptions written by researchers or the builders of digital libraries about their own work. One source was an exhibit at the U.S. Copyright Office; one was an out-of-print book; for a couple of topics, I sent electronic mail to friends. For everything else, the source of the material was the Internet. Many of the materials do not exist in conventional formats. In the field of digital libraries, the Internet is already the library.

A dream of future libraries combines everything that we most prize about traditional methods, with the best that online information can offer. Sometimes we have nightmares in which the worst aspects of each are combined. In the first years of this century, the philanthropy of Andrew Carnegie brought public libraries to the United States. Now a new form of library is emerging. Hopefully, digital libraries will attract the same passion and respect, and serve the same deep needs that have long been associated with the best of libraries and publishing.

Glossary

Digital libraries have absorbed terminology from many fields, including computing, libraries, publishing, law, and more. This glossary gives brief explanations of how some common terms are used in digital libraries today, which may not be the usage in other contexts. Often the use in digital libraries has diverged from or extended the original sense of a term.

AACR2 (Anglo-American Cataloguing Rules)

A set of rules that describe the content that is contained in library catalog records.

abstracting and indexing services

Secondary information services that provide searching of scholarly and scientific information, in particular of individual journal articles.

access management

Control of access to material in digital libraries. Sometimes called terms and conditions or rights management.

ACM Digital Library

A digital library of the journals and conference proceedings published by the Association for Computing Machinery.

Alexandria Digital Library

A digital library of geospatial information, based at the University of California, Santa Barbara.

American Memory and the National Digital Library Program

The Library of Congress's digital library of materials converted from its primary source materials related to American history.

applet

A small computer program that can be transmitted from a server to a client computer and executed on the client.

archives

Collections with related systems and services, organized to emphasize the long-term preservation of information.

Art and Architecture Thesaurus

A controlled vocabulary for fine art, architecture, decorative art, and material culture, a project of the J. Paul Getty Trust.

artifact

A physical object in a library, archive, or museum.

ASCII (American Standard Code for Information Interchange)

A coding scheme that represents individual characters as 7 or 8 bits; printable ASCII is a subset of ASCII.

authentication

Validation of a user, a computer, or some digital object to ensure that it is what is claims to be.

authorization

Giving permission to a user or client computer to access specific information and carry out approved actions.

automatic indexing

Creation of catalog or indexing records using computer programs, not human cataloguers.

Boolean searching

Methods of information retrieval where a query consists of a sequence of search terms, combined with operators, such as "and", "or", and "not".

browser

A general-purpose user interface, used with the web and other online information services. Also known as a web browser.

browsing

Exploration of a body of information, based on the organization of the collections or scanning lists, rather than by direct searching.

cache

A temporary store that is used to keep a readily available copy of recently used data or any data that is expected to be used frequently.

California Digital Library

A digital library that serves the nine campuses of the University of California.

catalog

A collection of bibliographic records created according to an established set of rules.

classification

An organization of library materials by a hierarchy of subject categories.

client

A computer that acts on behalf of a user, including a user's personal computer, or another computer that appears to a server to have that function.

CGI (Common Gateway Interface)

A programming interface that enables a web browser to be an interface to information services other than web sites.

Chemical Abstracts

A secondary information service for chemistry.

CNI (Coalition for Networked Information)

A partnership of the Association for Research Libraries and Educause to collaborate on academic networked information.

complex object

Library object that is made up from many inter-related elements or digital objects.

compression

Reduction in the size of digital materials by removing redundancy or by approximation; lossless compression can be reversed; lossy compression can not be reversed since information is lost by approximation.

computational linguistics

The branch of natural language processing that deals with grammar and linguistics.

controlled vocabulary

A set of subject terms, and rules for their use in assigning terms to materials for indexing and retrieval.

conversion

Transformation of information from one medium to another, including from paper to digital form.

CORBA

A standard for distributed computing where an object on one computer invokes an Object Request Broker (ORB) to interact with an object on another computer.

CORE

A project from 1991 to 1995 by Bellcore, Cornell University, OCLC, and the American Chemical Society to convert chemistry journals to digital form.

Cryptolope

Secure container used to buy and sell content securely over the Internet, developed by IBM.

CSS (Cascading Style Sheets)

System of style sheets for use with HTML, the basis of XLS.

CSTR (Computer Science Technical Reports project)

A DARPA-funded research project with CNRI and five universities, from 1992 to 1996.

DARPA (Defense Advanced Research Projects Agency)

A major sponsor of computer science research in the U.S., including digital libraries. Formerly ARPA.

data type

Structural metadata associated with digital data that indicates the digital format or the application used to process the data.

DES (Data Encryption Standard)

A method for private key encryption.

Dewey Decimal Classification

A classification scheme for library materials which uses a numeric code to indicate subject areas.

desktop metaphor

User interface concept on personal computers that represents information as files and folders on a desktop.

Dienst

An architecture for digital library services and an open protocol that provides those services, developed at Cornell University, used in NCSTRL.

digital archeology

The process of retrieving information from damaged, fragmentary, and archaic data sources.

Digital Libraries Initiative

A digital libraries research program. In Phase 1, from 1994 to 1998, NSF/DARPA/NASA funded six university projects
Phase 2 began in 1998/9.

digital object

An item as stored in a digital library, consisting of data, metadata, and an identifier.

digital signature

A cryptographic code consisting of a hash, to indicate that data has not changed, that can be decrypted with the public key of the creator of the signature.

dissemination

The transfer from the stored form of a digital object in a repository to a client.

distributed computing

Computing systems in which services to users are provided by teams of computers collaborating over a network.

D-Lib Magazine

A monthly, online publication about digital libraries research and innovation.

DLITE

An experimental user interface used with the Stanford University InfoBus.

document

Digital object that is the analog of a physical document, especially textual materials; a document model is an object model for documents.

domain name

The name of a computer on the Internet; the domain name service (DNS) converts domain names to IP addresses.

DOI (Digital Object Identifier)

An identifier used by publishers to identify materials published electronically, a form of handle.

DSSSL (Document Style Semantics and Specification Language)

A general purpose system of style sheets for SGML.

DTD (Document Type Definition)

A mark-up specification for a class of documents, defined within the SGML framework.

Dublin Core

A simple set of metadata elements used in digital libraries, primarily to describe digital objects and for collections management, and for exchange of metadata.

dynamic object

Digital object where the dissemination presented to the user depends upon the execution of a computer program, or other external activity.

EAD (Encoded Archival Description)

A DTD used to encode electronic versions of finding aids for archival materials.

electronic journal

A online publication that is organized like a traditional printed journal, either an online version of a printed journal or a journal that has only an online existence.

eLib

A British program of innovation, around the theme of electronic publication.

emulation

Replication of a computing system to process programs and data from an early system that is no longer available.

encryption

Techniques for encoding information for privacy or security, so that it appears to be random data; the reverse process, decryption, requires knowledge of a digital key.

entities and elements

In a mark-up language, entities are the basic unit of information, including character entities; elements are strings of entities that form a structural unit.

expression

The realization of a work, by expressing the abstract concept as actual words, sounds, images, etc.

fair use

A concept in copyright law that allows limited use of copyright material without requiring permission from the rights holders, e.g., for scholarship or review.

federated digital library

A group of digital libraries that support common standards and services, thus providing interoperability and a coherent service to users.

field, subfield

An individual item of information in a structured record, such as a catalog or database record.

fielded searching

Methods for searching textual materials, including catalogs, where search terms are matched against the content of specified fields.

finding aid

A textual document that describes holdings of an archive, library, or museum.

firewall

A computer system that screens data passing between network segments, used to provide security for a private network at the point of connection to the Internet.

first sale

A concept in copyright law that permits the purchaser of a book or other object to transfer it to somebody else, without requiring permission from the rights holders.

FTP (File Transfer Protocol)

A protocol used to transmit files between computers on the Internet.

full text searching

Methods for searching textual materials where the entire text is matched against a query.

gatherer

A program that automatically assembles indexing information from digital library collections.

gazetteer

A database used to translate between different representations of geospatial references, such as place names and geographic coordinates.

genre

The class or category of an object when considered as an intellectual work.

geospatial information

Information that is reference by a geographic location.

gif

A format for storing compressed images.

Google

A web search program that ranks web pages in a list of hits by giving weight to the links that reference a specific page.

gopher

A pre-web protocol used for building digital libraries, now largely obsolete.

handle

A system of globally-unique names for Internet resources and a computer system for managing them, developed by CNRI; a form of URN.

Harvest

A research project that developed an architecture for distributed searching, including protocols and formats.

hash

A short value calculated from digital data that serves to distinguish it from other data.

HighWire Press

A publishing venture, from Stanford University Libraries, that provides electronic versions of journals, on behalf of learned and professional societies.

hit

1. An incoming request to a web server or other computer system.
2. In information retrieval, a document that is discovered in response to a query.

home page

The introductory page to a collection of information on the web.

HTML (Hyper-Text Mark-up Language)

A simple mark-up and formatting language for text, with links to other objects, used with the web.

HTTP (Hyper-Text Transport Protocol)

The basic protocol of the web, used for communication between browsers and web sites.

hyperlink

A network link from one item in a digital library or web site to another.

ICPSR (International Consortium for Political and Social Science Research)

An archive of social science datasets, based at the University of Michigan.

identifier

A string of characters that identifies a specific resource in a digital library or on a network.

IETF (Internet Engineering Task Force)

The body that coordinates the technological development of the Internet, including standards.

InfoBus

An approach to interoperability that uses proxies as interfaces between existing systems, developed at Stanford University.

information discovery

General term covering all strategies and methods of finding information in a digital library.

information retrieval

Searching a body of information for objects that match a search query.

Informedia

A research program and digital library of segments of video, based at Carnegie Mellon University.

Inspec

An indexing service for physics, engineering, computer science, and related fields.

Internet

An international network, consisting of independently managed networks using the TCP/IP protocols and a shared naming system. A successor to the ARPAnet.

Internet RFC series

The technical documentation of the Internet, provided by the Internet Engineering Task Force. Internet Drafts are preliminary versions of RFCs.

interoperability

The task of building coherent services for users from components that are technically different and independently managed.

inverted file

A list of the words in a set of documents and their locations within those documents; an inverted list is the list of locations for a given word.

item

A specific piece of material in a digital library; a single instance or copy of a manifestation.

Java

A programming language used for writing mobile code, especially for user interfaces, developed by Sun Microsystems.

JavaScript

A scripting language used to embed executable instructions in a web page.

JPEG

A format for storing compressed images.

JSTOR

A subscription service, initiated by the Andrew W. Mellon Foundation, to convert back runs of important journals and make them available to academic libraries.

key

A digital code used to encrypt or decrypt messages. Private key encryption uses a single, secret key. Dual key (public key) encryption uses two keys of which one is secret and one is public.

legacy system

An existing system, usually a computer system, that must be accommodated in building new systems.

lexicon

A linguistic tool with information about the morphological variations and grammatical usage of words.

Lexis

A legal information service, a pioneer of full-text information online.

Los Alamos E-Print Archives

An open-access site for rapid distribution of research papers in physics and related disciplines.

manifestation

Form given to an expression of a work, e.g., by representing it in digital form.

MARC (Machine-Readable Cataloging)

A format used by libraries to store and exchange catalog records.

mark-up language

Codes embedded in a document that describe its structure and/or its format.

Medline

An indexing service for research in medicine and related fields, provided by the National Library of Medicine.

MELVYL

A shared digital library system for academic institutions in California; part of the California Digital Library.

Memex

A concept of an online library suggested by Vannevar Bush in 1945.

Mercury

An experimental digital library project to mount scientific journals online at Carnegie Mellon University from 1987 to 1993.

MeSH (Medical Subject Headings)

A set of subject term and associated thesaurus used to describe medical research, maintained by the National Library of Medicine.

metadata

Data about other data, commonly divided into descriptive metadata such as bibliographic information, structural metadata about formats and structures, and administrative metadata, which is used to manage information.

migration

Preservation of digital content, where the underlying information is retained but older formats and internal structures are replaced by newer.

MIME (Internet Media Type)

A scheme for specifying the data type of digital material.

mirror

A computer system that contains a duplicate copy of information stored in another system.

mobile code

Computer programs or parts of programs that are transmitted across a network and executed by a remote computer.

morphology

Grammatical and other variants of words that are derived from the same root or stem.

Mosaic

The first widely-used web browser, developed at the University of Illinois.

MPEG

A family of formats for compressing and storing digitized video and sound.

multimedia

A combination of several media types in a single digital object or collection, e.g., images, audio, video.

natural language processing

Use of computers to interpret and manipulate words as part of a language.

NCSTRL (Networked Computer Science Technical Reports Library)

An international distributed library of computer science materials and services, based at Cornell University.

Netlib

A digital library of mathematical software and related collections.

NSF (National Science Foundation)

U.S. government agency that supports science and engineering, including digital libraries research.

object

A technical computing term for an independent piece of computer code with its data. Hence, object-oriented programming, and distributed objects, where objects are connected over a network.

object model

A description of the structural relationships among components of a library object including its metadata.

OCLC (Online Computer Library System)

An organization that provides, among other services, a bibliographic utility for libraries to share catalog records.

OPAC (online public access catalog)

An online library catalog used by library patrons.

open access

Resources that are openly available to users with no requirements for authentication or payment.

optical character recognition

Automatic conversion of text from a digitized image to computer text.

Pad++

A experimental user interface for access to large collections of information, based on semantic zooming.

page description language

A system for encoding documents that precisely describes their appearance when rendered for printing or display.

PDF (Portable Document Format)

A page description language developed by Adobe Corporation to store and render images of pages.

peer review

The procedure by which academic journal articles are reviewed by other researchers before being accepted for publication.

Perseus

A digital library of hyperlinked sources in classics and related disciplines, based at Tufts University.

policy

A rule established by the manager of a digital library that specifies which users should be authorized to have what access to which materials.

port

A method used by TCP to specify which program running on a computer should process a message arriving over the Internet.

PostScript

A programming language to create graphical output for printing, used as a page description language.

precision

In information retrieval, the percentage of hits found by a search that satisfy the request that generated the query.

presentation profile

Guidelines associated with a digital object that suggest how it might be presented to a user.

protocol

A set of rules that describe the sequence of messages sent across a network, specifying both syntax and semantics.

proxy

A computer that acts as a bridge between two computer systems that use different standards, formats, or protocols.

publish

To make information available and distribute it to the public.

PURL (Persistent URL)

A method of providing persistent identifiers using standard web protocols, developed by OCLC.

query

A textual string, possibly structured, that is used in information retrieval, the task being to find objects that match the words in the query.

ranked searching

Methods of information retrieval that return a list of documents, ranked in order of how well each matches the query,

RDF (Resource Description Framework)

A method for specifying the syntax of metadata, used to exchange metadata.

RealAudio

A format and protocol for compressing and storing digitized sound, and transmitting it over a network to be played in real time.

recall

In informational retrieval, the percentage of the items in a body of material which would satisfy a request that are actually found by a search.

refresh

To make an exact copy of data from older media to newer for long-term preservation.

render

To transform digital information in the form received from a repository into a display on a computer screen, or for other presentation to the user.

replication

Make copies of digital material for backup, performance, reliability, or preservation.

repository

A computer system used to store digital library collections and disseminate them to users.

RSA encryption

A method of dual key (public key) encryption.

scanning

Method of conversion in which a physical object, e.g., a printed page, is represented by a digital grid of pixels

search term

A single term within a query, usually a single word or short phrase.

secondary information

Information sources that describe other (primary) information, e.g., catalogs, indexes, and abstracts; used to find information and manage collections.

security

Techniques and practices that preserve the integrity of computer systems, and digital library services and collections.

server

Any computer on a network, other than a client, that stores collections or provides services.

SGML (Standard Generalized Markup Language)

A system for creating mark-up languages that represent the structure of a document.

SICI (Serial Item and Contribution Identifier)

An identifier for an issue of a serial or an article contained within a serial.

speech recognition

Automatic conversion of spoken words to computer text.

STARTS

An experimental protocol for use in distributed searching, which enables a client to combine results from several search engines.

stemming

In informational retrieval, reduction of morphological variants of a word to a common stem.

stop word

A word that is so common that it is ignored in information retrieval. A set of such words is called a stop list.

structural type

Metadata that indicates the structural category of a digital object.

style sheet

A set of rules that specify how mark-up in a document translates into the appearance of the document when rendered.

subscription

In a digital library, a payment made by a person or an organization for access to specific collections and services, usually for a fixed period, e.g., one year.

subsequent use

Use made of digital materials after they leave the control of a digital library.

tag

A special string of characters embedded in marked-up text to indicate the structure or format.

TCP/IP

The base protocols of the Internet. IP uses numeric IP addresses to join network segments; TCP provides reliable delivery of messages between networked computers.

TEI (Text Encoding Initiative)

A project to represent texts in digital form, emphasizing the needs of humanities scholars. Also the DTD used by the program.

TeX

A method of encoding text that precisely describes its appearance when printed, especially good for mathematical notation. LaTeX is a version of TeX.

thesaurus

A linguistic tool that relates words by meaning.

Ticer Summer School

A program at Tilburg University to educate experienced librarians about digital libraries.

Tipster

A DARPA program of research to improve the quality of text processing methods, including information retrieval.

transliteration

A systematic way to convert characters in one alphabet or phonetic sounds into another alphabet.

TREC (Text Retrieval Conferences)

Annual conferences in which methods of text processing are evaluated against standard collections and tasks.

truncation

Use of the first few letters of a word as a search term in information retrieval.

Tulip

An experiment in which Elsevier Science scanned material science journals and a group of universities mounted them on local computers.

UDP

An Internet protocol which transmits data packets without error checking.

Unicode

A 16-bit code to represent the characters used in most of the world's scripts. UTF-8 is an alternative encoding in which one or more 8-bit bytes represents each Unicode character.

union catalog

A single catalog that contains records about materials in several collections or libraries.

URL (Uniform Resource Locator)

A reference to a resource on the Internet, specifying a protocol, a computer, a file on that computer, and parameters. An absolute URL specifies a location as a domain name or IP address; a relative URL specifies a location relative to the current file.

URN (Uniform Resource Name)

Location-independent names for Internet resources.

WAIS

An early version of Z39.50, used in digital libraries before the web, now largely obsolete.

Warwick Framework

A general model that describes the various parts of a complex object, including the various categories of metadata.

watermark

A code embedded into digital material that can be used to establish ownership, may be visible or invisible to the user.

web crawler

A web indexing program that builds an index by following hyperlinks continuously from web page to web page.

webmaster

A person who manages web sites.

web search services

Commercial services that provide searching of the web, including: Yahoo, Altavista, Excite, Lycos, Infoseek, etc.

web site

A collection of information on the web; usually stored on a web server.

Westlaw

A legal information service provided by West Publishing.

World Wide Web (web)

An interlinked set of information sources on the Internet, and the technology they use, including HTML, HTTP, URLs, and MIME.

World Wide Web Consortium (W3C)

A international consortium based at M.I.T. that coordinates technical developments of the web.

work

The underlying intellectual abstraction behind some material in a digital library.

Xerox Digital Property Rights Language

Syntax and rules for expressing rights, conditions, and fees for digital works.

XLS (eXtensible Style Language)

System of style sheets for use with XML, derived from CSS.

XML (eXtensible Mark-up Language)

A simplified version of SGML intended for use with online information.

Z39.50

A protocol that allows a computer to search collections of information on a remote system, create sets of results for further manipulation, and retrieve information; mainly used for bibliographic information.

"William Arms offers a comprehensive look at digital libraries from many perspectives. He's right: we're just at the beginning of this story. The best is yet to come."

—Vint Cerf, *Senior Vice President, Internet Architecture and Technology, MCI WorldCom*

The MIT Press Massachusetts Institute of Technology Cambridge, Massachusetts 02142
http://mitpress.mit.edu

ARMDH 0-262-01180-8