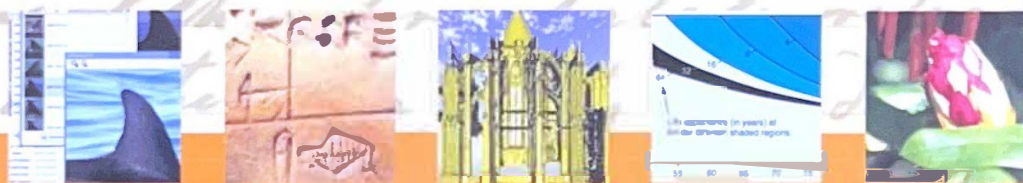




Michael **LESK**



Understanding Digital Libraries

SECOND EDITION

MK
JOHANN KAUFMANN

META-1008
IPR2025-00430

The Morgan Kaufmann Series in Multimedia Information and Systems

Series Editor, Edward A. Fox, Virginia Polytechnic University

Understanding Digital Libraries, Second Edition

Michael Lesk

Bioinformatics: Managing Scientific Data

Zoe Lacroix and Terence Critchlow

How to Build a Digital Library

Ian H. Witten and David Bainbridge

Digital Watermarking

Ingemar J. Cox, Matthew L. Miller, and Jeffrey A. Bloom

Readings in Multimedia Computing and Networking

Edited by Kevin Jeffay and HongJiang Zhang

Introduction to Data Compression, Second Edition

Khalid Sayood

Multimedia Servers: Applications, Environments, and Design

Dinkar Sitaram and Asit Dan

*Managing Gigabytes: Compressing and Indexing Documents and Images,
Second Edition*

Ian H. Witten, Alistair Moffat, and Timothy C. Bell

Digital Compression for Multimedia: Principles and Standards

Jerry D. Gibson, Toby Berger, Tom Lookabaugh, Dave Lindbergh, and Richard L. Baker

Readings in Information Retrieval

Edited by Karen Sparck Jones and Peter Willett

Understanding Digital Libraries

Second Edition

Michael Lesk



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

MORGAN KAUFMANN PUBLISHERS IS AN IMPRINT OF ELSEVIER



MORGAN KAUFMANN PUBLISHERS

Publishing Director
Senior Editor
Acquisitions Editor
Publishing Services Manager
Project Manager
Developmental Editor
Editorial Coordinator
Editorial Assistant
Cover Design
Composition
Technical Illustration
Copyeditor
Proofreader
Indexer
Interior printer
Cover printer

Diane D. Cerra
Rick Adams
Lothlórén Homet
Simon Crump
Dan Stone
Corina Derman
Mona Buehler
Asma Stephan
Frances Baca
CEPHA Imaging Pvt. Ltd.
Dartmouth Publishing Inc.
Yonie Overton
Daniel Easler
Kevin Broccoli
The Maple-Vail Book Manufacturing Group
Phoenix Color

Morgan Kaufmann Publishers is an imprint of Elsevier.
500 Sansome Street, Suite 400, San Francisco, CA 94111

© 2005 by Elsevier Inc. All rights reserved.

Designations used by companies to distinguish their products are often claimed as trademarks or registered trademarks. In all instances in which Morgan Kaufmann Publishers is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopying, scanning, or otherwise—without prior written permission of the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone: (+44) 1865 843830, fax: (+44) 1865 853333, e-mail: permissions@elsevier.com.uk. You may also complete your request on-line via the Elsevier homepage (<http://elsevier.com>) by selecting "Customer Support" and then "Obtaining Permissions."

Library of Congress Cataloging-in-Publication Data
APPLICATION SUBMITTED

ISBN: 1-55860-924-5

For information on all Morgan Kaufmann publications,
visit our Web site at www.mkp.com or www.books.elsevier.com

Printed in the United States of America
05 06 07 08 09 5 4 3 2 1

Preface

Changes Since the First Edition

Digital library research and practice has exploded since the first edition of this book, led by the growth of the Web as an interface and access route for information retrieval on a global basis. Online content has flooded onto the Web from publishers, libraries, museums, undergraduates, and practically anyone else with a keyboard or scanner. Library catalogs, museum exhibitions, and historic manuscripts are now available in this format, along with the ubiquitous advertisements and product manuals. A uniform interface has encouraged the proliferation of online content. For instance, in mid-2003 Google claimed to be indexing more than 3 billion Web pages. The Web has more than 150 terabytes of text today, more than all but the very largest research libraries.

Unfortunately, the economic and legal problems discussed in the first edition are still with us. Since the original publication, we have witnessed the dot-com explosion and crash. Despite the many startups and new ventures devoted to online information, we are still looking for a sustainable way to support digital libraries. The copyright tangle has become yet more difficult as a result of the worldwide extension of the period of copyright by 20 years and aggressive steps by the recorded music industry to enforce online restrictions.

The updates in this second edition focus largely on the impact of the Web, as well as on new digital library research projects. For the first time, the average person recognizes something that comes from a digital library research project: Google. This spin-off from the Stanford University Digital Library project, run by Hector Garcia-Molina, is now responsible for more than half the searches

4

Multimedia Storage and Retrieval

Libraries do not just consist of text pages, whether ASCII or image. The current rage for “multimedia” encourages us to plan for collections that include sound, video, software, and everything else imaginable. Unlike the optimistic conclusions in the last two chapters, for which there were suitable methods and successful examples of digital libraries with millions of pages of ASCII and hundreds of thousands of images, building systems for multimedia is harder. In particular, video remains a problem for libraries. The ubiquitous VHS cartridge is fragile and poor in quality. Color photographic film is expensive. And digitizing video today is still expensive and little video is free of copyright restraints; nor do we have good ways of searching it once stored. This chapter will discuss where we stand on handling nontextual material; it may not be straightforward today, but there is active progress. This is probably the area of greatest research interest and progress today.

4.1 Sound Formats: Vinyl, Tape, and CD

The simplest form of nontext material to deal with is audio. Audio has been collected by libraries for years; the Library of Congress even has thousands of

player piano rolls. The once-standard vinyl record gave way to cassette tape and then to CD. In one of the faster complete replacements of one technology by another, CD made vinyl obsolete in about 10 years. Cassette tape still survives for the moment since its players are cheaper and, until recently, more standard in automobiles.

There are two common levels of formatting for audio: a high-quality standard for music and a low-quality one for voice. The format for storing music is driven by the entertainment industry, which now uses digital encoding for both studio and consumer formats. Ideally, one would like recorded sound to approximate a live performance in its impression on the listener's ear; a typical person can hear up to 15–20 kHz and a dynamic range of 30–60 dB in loudness. kHz (kilohertz, or thousands of cycles per second) is a measure of frequency. Middle C on a piano keyboard is 0.256 kHz, the A above it is 0.440 kHz, and most people can not hear above 20 kHz, although dogs can usually hear 40 kHz, and cats up to 65 kHz. When digitizing sound, the frequency range in kHz determines the sampling rate, and the dynamic range (the ratio of the loudest sound to the smallest useful difference in loudness) determines the number of bits per sample. Since it requires a minimum of two samples per cycle to give the ear a clue to the existence of a frequency, the sampling rate must be double the highest frequency that is to be captured. Thus, to represent the range up to 15 kHz requires 30,000 samples per second or higher. In other words, a value of the sound wave intensity must be recorded about every 30 microseconds (30 millionths of a second). The decibel scale raises the sound intensity by a factor of 10 every 10 decibels, so that a 65 dB sound is ten times as large as a 55 dB sound. Roughly, 1 dB is the quietest perceptible sound (for someone with good hearing), 30 dB is a quiet room, 40 dB is conversation, 60 dB is loud music, 80 dB is a noisy car, and 100 dB is painful (e.g., an express subway train rushing through a station). To represent the difference between the 30 dB quiet parts and the 60 dB loud parts of a musical performance requires a range of a thousand in amplitude, or at least 10 bits of dynamic range. Music is typically handled by computers in CD format: 44,000 samples per second at 16 bits/sample on each of two audio tracks, comfortably above what is needed to represent what the average ear can perceive. This means that audio CD players must handle 176-KB/second (the standard 1X CD-ROM reader). This format is sufficiently high-quality in terms of both acoustic range and accuracy that most people accept it as a permanent representation, except for a few audiophiles who insist that analog sound has a different and preferable texture. Regrettably, the Audio Engineering Society (2003) is circulating a draft proposal for preservation of old sound recordings which asks for sampling rates of at least 88 kHz and amplitude digitization at 24 bits. This is well beyond the capacity of either the human ear or the equipment on which the

5

Knowledge Representation Schemes

The quantities of material stored in digital libraries, as in all libraries, pose the problem of finding what you want among the stacks, so to speak. In some collections of information, the user is simply overwhelmed. Queries return either nothing or thousands of answers, and users are often baffled about what to do. Attempts have been made over the years to provide a way to find information conveniently, often based on the idea of organizing everything by subject. Often, this was a byproduct of the need to put books on shelves. On disk drives the physical location of any particular byte is irrelevant to the librarian, but the intellectual organization of the collection still matters. This chapter will describe the ways we have tried to organize and arrange knowledge, with the idea of making searching simple.

If we had a single knowledge representation scheme that let us put each idea in one place, and if the users knew this scheme and could place each of their queries in it, subject retrieval would be straightforward. There would still be a need for items at different levels of sophistication, and a need for quality checking, but

fish oil causes certain blood changes and 34 articles showing how the same blood changes would alleviate the problems of Reynaud's disease, but no articles discussing fish oil and Reynaud's disease. And, indeed, later experiments showed that fish oil appears to be of some use in treating Reynaud's disease. Similarly, he found that there is considerable literature about migraine and epilepsy and about epilepsy and magnesium deficiency, but no papers about dietary magnesium and migraine. After Swanson's publication, several papers appeared with evidence that magnesium alleviates migraine. Despite these examples, it has proved difficult to follow up this work. Neither Swanson nor his followers have been able to mechanically find other examples of this sort.

Will usage-based searching be important in digital libraries as well? Digitally, it is much easier to track usage of different items. As yet, there are no systems that say, "Show me things lots of people have liked," or systems that say, "Show me things other people have overlooked." Such a system is called *community-rated* information. The idea here is that you can model the search you want to do by looking at what other people have done. This is the logic of "Anything Joe recommends, I will like," done mathematically and with a large set of people.

An early example was the "video recommender" of Will Hill and Mark Rosenstein at Bellcore (Hill et al., 1995). They set up an email address which asked people to rate movies from 1 to 10. They sent out a list of 500 movies; amazingly, the average person who returned this list rated about 200 movies. They wound up with a database of about 25,000 ratings. When a new person entered ratings, the database was searched to find the 10 people with the closest agreement. This new person could then be modelled as a linear combination of the best 10 matching people. Given this way of modelling the person, the system could then suggest movies that the model predicted the user would like, but that the user had not rated (and presumably had not seen).

Tests showed that this was a very effective way of recommending movies. Table 5.6 shows different methods of suggesting what videotapes you might want to rent. The correlation coefficient is between the actual rating and the rating predicted by the method given.

Similar work has been done for audio (popular music CDs) by Patti Maes and her group at the MIT Media Lab (Shardanand and Maes, 1995). The technique

Table 5.6 Effectiveness of recommending movies.

Method	Correlation
Choosing movies at random	0.16
Using recommendations of published movie critics	0.22
Community-rating, as described	0.62

7

Usability and Retrieval Evaluation

In Chapter 5 we talked about finding the “Holy Grail” of knowledge representation. Making computers easy to use is the “Holy Grail” of the software industry. Frustration with computers is very common; we all know jokes like “To err is human, to really foul things up takes a computer.” The SBT Accounting Group suggested that the average office worker wastes 5 hours a week because of computer problems, costing the US economy \$100 billion a year (Gibbs, 1997; Shneiderman, 2000). In 2002 the National Institute of Standards and Technology (NIST) estimated that just software bugs were costing the economy \$60 billion per year (see Tasse, 2002). Frustration and even “computer rage” are now common, often as a result of difficult interfaces. The BBC reported that three-quarters of users admitted to swearing at their computers (Welsh, 1999). Libraries accept a wide range of the public as their patrons and cannot force them to accept training. Users of Google and the other search engines are accustomed to just using them, with no training at all, and they are generally satisfied with the result. Library schools give a semester-long course in online searching. Yes, the trained people get better results and waste less time, but a lot of people enjoy searching and are happy with what happens. Nevertheless, some people are frustrated, and perhaps more important is the fact that many people could get better

results than they do. Early experiments by Chris Borgman showed that a quarter of Stanford undergraduates had difficulty learning one of the commercial search systems. Searching is still often hit-or-miss; it is easy to find information on a person with an unusual name, say John Galsworthy, while if you are looking for information on the island of Java you will need to phrase the search carefully, perhaps looking for Java and Indonesia. A practical digital library has to build systems everyone can use. This chapter will emphasize the user and the considerations necessary to make systems effective, not just available.

7.1 General Human Factors Considerations

The general science of human factors is called *ergonomics* and it began with the study of physical tools and equipment, such as the shaping of handles and levers. Ergonomics is still referred to derisively as the “comfy chair” science. The basic principles of ergonomics come from cognitive psychology, and, as it happens, the research methods from that area are used in evaluating software as well. Two alternative technical solutions are implemented and tested on a group of subjects, traditionally, on undergraduates in elementary psychology courses.

Tests of this sort follow a traditional paradigm of scientific research. The subjects are divided into groups, each group using a different interface or trying a different task. For example, one group might be using an index language, while another is using a free-text search system. Statistical measures are used to decide whether there is a significant difference between the conditions. The measures may be based on the accuracy, speed, or preferences of the users. Preferences may be based either on some choice (e.g., the amount of time the users choose to spend on some condition), or on questionnaires given to the users. Time taken should always be recorded, as much good cognitive psychology data is based on human reaction times as a measure of how much processing is required in the brain.

Studies on human subjects need to be done with care. For example, it is always desirable to run some preliminary tests before the main experiments. Subjects will think of things the experimenter did not anticipate, and it may be necessary to revise the experiment in consequence. Subjects have to be balanced in terms of their computer expertise, experience with the subject matter of the experiment, and so on. Normally this is done by randomization: the subjects are assigned at random to two different conditions. If two computer systems are being compared, they need to be balanced as well. If one is on color screens and the other on black and white, unless that is the point of the experiment, the data are *confounded*, and it is likely to be difficult to learn anything from such an experiment. A critical point is response time of the systems, since people

are relatively intolerant of slow response time and, again, unless this is the point of the comparison, response times should be balanced across experimental conditions. An even more serious issue is bugs in the software. All too often, an experiment produces confused results because the users hit new bugs in the process of testing which distort the results.

Often, an analysis of variance is done on the data to decide how much of the difference between the results can be assigned to different characteristics of the experiment. For example, an experiment might involve 16 students, 8 of whom had previous programming experience and 8 of whom did not. Half might be given one text editor, and half another. Thus, 4 students of each degree of experience would see each experimental condition. At the end, analysis of variance would indicate the relative importance of the programming experience versus the particular text editor. Such experiments are often discouraging, since the variance among individuals in many computer tasks is very high. Thus, the result of many programming comparisons has been to discover that some people are better programmers than others. Compared with individual skills and with the difficulty level of the particular task to be done, the programming language used is almost unimportant.

In fact, the history of experiments of this sort is that they have been valuable in developing the details of interfaces, but have not been as successful in generating new top-level paradigms. In the early days of computer, time-sharing attempts were made to do systematic comparisons of batch processing and time-sharing, and the results were inconclusive until time-sharing completely overran batch processing in the marketplace (to be pushed out in its turn by individual workstations). Similarly, the number of truly appalling human interfaces which have been successful in the marketplace is testimony more to the ability of determined users to overcome obstacles in their path than to the success of human factors research in program design. Getting to the marketplace first and establishing some kind of standard is more important than having good research support for one's design.

Human factors research is going to become more important, however. The development of the computer interface is a major part of new software systems. More than half the code in many new software projects is running the interface, and the interface is often critical in someone's decision about which program to use (when there is a choice). However, the design of good interfaces is still as much art as science, since there is still insufficient experimental data and good models of how to build an interface.

One principle from studies is the speed-accuracy trade-off that sometimes exists. If people work more quickly, they may make more errors. This isn't true across the board; more experienced people often work both more quickly and more accurately. We all know that someone competent in a language can speak

8

User Needs

Despite the widespread acceptance of digital content and digital searching in libraries today, the match to what users actually want and what they can do is very weak. Users are neither clear about what they want, able to operate the systems well, or doing much to get help. However, they're satisfied with the results. To the extent that we can tell, the acceptance of new systems is partly based on the inability of users to tell how badly things are actually working, and partly on the probability that older systems were also not being used very effectively. At least, we have opportunities for improvement, if we can pay enough attention to the users when designing the systems.

8.1 Overview

Despite jokes about librarians who think their job is to keep the books safe from the readers, libraries exist to provide information to people, today and in the future. Digital libraries also provide information and services to their users. How will those services change and what new things will be happening?

Traditional libraries provided services that extended from entertainment to advanced research, but did not cover as wide a range as that on the Web. For example, libraries often rented videotapes, but rarely encouraged patrons to use library television sets and VCRs to watch them on the premises. Libraries could

feasibly provide daily newspapers for financial and weather information with 24-hour delay, but do not have current stock tickers or weather reports. On the other hand, as compared with the Web, libraries have librarians who can help users with their requests, and they provide a range of services to children (such as reading aloud to them) which are not realistic or practical online.

We have a model of the “information seeking” user in which the user has an “information need” of some kind, which is then expressed in some form of query, and then receives an answer in the form of a citation to some document. The user then reads the document and decides whether and how to rephrase the query. Once upon a time the user might have looked in some paper index, such as *Reader's Guide to Periodical Literature*, and found a list of traditional bibliographic citations; nowadays, the query goes to Google and the user gets a set of URLs, which are easier to turn into real documents. The user might also approach a reference librarian with the query, and the librarian might use a variety of sources, both print and digital, as well as personal memory and experience, to suggest a list of possible answers.

However, this model is oversimplified. Often, the result of seeing a possible list of answers is not just a change in the way the information need is expressed, but a change in the original need itself. That is, it's one thing to be unable to spell Alan Ayckbourn, but it's another kind of change to decide you really needed to look up Michael Frayn. In addition, many queries are really very vague and fuzzy. Although half the people in a research library enter with a particular item in mind that they want to find, the majority of the people who walk into a public branch library or browse the racks at an airport bookstall are just looking for something interesting to read; they have no particular book in mind.

Even to the extent that people can describe something that they want, often the description is not in a form that lends itself to description in keywords for either Google or a printed index. Qualities such as the language of the text or the place or date of publication are often in catalogs and may be supported by a search engine or OPAC (online public access catalog). Some additional qualities are imaginable as catalog contents: the user might want a short book, or a children's book, or an illustrated book, or a popularization as opposed to a book for experts. Often people say that they remember the color or shelf position of a book they want to find; regrettably, their memory is often wrong or at least insufficiently precise to be a useful search term (and of course shelf position can change over time, whether in a library, bookstore, or home).

But user wishes can be far less specific. The airport book buyer probably wants a book that is “exciting” or “absorbing.” Christmas card publishers want landscape scenes that are “placid” and full of snow, while art catalogers write down the name of the place depicted and the painter's name. Allison Druin, who

studies digital libraries for children, remembers one of her young patrons asking for a book "that will make me feel good" (Druin et al., 2003).

The process of retrieval is not just "turn a crank," but rather a negotiation between the user and the search agent, whether a person or a piece of software. The user is both elaborating and changing his or her needs; the system is responding to what it can deal with. How to deal with this dialog is more fundamental than screen color or font size.

8.2 User Services

The primary user service in the traditional library is access to books and journals, but most libraries also provide reference services, circulation, catalogs, photocopying, and simple study space. More elaborate services can include interlibrary loan, reading aloud to children, training classes, sound and video collecting, maintaining book reserves, running book sales, advice to tourists, running exhibitions, and more. The user population is usually defined in some way, whether residents of a particular town or students, faculty, and staff at some educational institution. However, enforcement of entry requirements is often minimal, the main reason for library cards being the need to keep track of those who borrow books. Typically, even in some large university libraries, anyone is allowed to walk in and browse (many state universities must, by law, admit every citizen for on-site services). Geography is enough to see that almost every user of a typical library will be somebody from the community which is paying for it.

Digital libraries have a corresponding but changed list of activities. Access can be larger and broader, with no need to limit collecting by shelf space. Circulation is unnecessary for Web access, since, in general, no physical copies of anything are being circulated. There are some services, such as Netlibrary, which do manage their materials on a simultaneous use basis. This kind of analogy to traditional borrowing lets each copy be read by only one user at a time; each user in turn gains access to the digital file for a time set by the library and then "returns" it so that some other reader may read it. The library can buy as many "copies" of each book as are needed to provide for simultaneous readers. The loaning of physical "e-books" may also involve a more traditional circulation.

More challenging problems are the provision of reference and help services. For those users in the library, it is much as before; but many of the users of a digital library service are not physically present and may be continents and hours away. Libraries offer telephone, email, and online chat alternatives; they even make arrangements with libraries around the world so that someone with a problem in the middle of the night in one place can try to get assistance

from someplace in an appropriate time zone. For example, Global Librarian is a collaboration of libraries in the United Kingdom, Canada, and Australia; somewhere in this list, it is likely to be normal operating hours. However, it's impossible to provide the same degree of encouragement or service to those not on the premises.

Exhibitions in the digital world are, if anything, more straightforward than with paper. There are no issues of fragile materials that can't be exposed to light, of transporting loans from one place to another, or of removing materials from use while on exhibit. For many major libraries, their first activity in digitization of their own holdings (as opposed to purchase or access to the digital materials of others) involves the creation of Web-based exhibits.

8.3 Acceptance in the Past: Why Not Microfilm?

The Web, of course, is far from the first new technology to reach libraries. Some fifty years ago there was a great deal of hype about microfilm. H. G. Wells, Fred Kilgour, and Vannevar Bush all said that microfilm would be the basis of a revolution in information availability. To quote Wells' more colorful phrasing, a microfilmed encyclopedia would have "at once the concentration of a craniate animal and the diffused vitality of an amoeba . . . foreshadows a real intellectual unification of our race." Librarians were attracted to microfilm since a relatively cheap photographic process produced enormous space reductions and also a "user copy" that could be employed in place of fragile originals. Newspapers were particularly appropriate for film, since they were bulky, printed on volatile paper, and the combination of large size and weak paper meant that they were particularly likely to tear or be damaged by heavy use.

What went wrong with microfilm? Technically, it works just fine, but in every library microfilm readers sit idle while computer terminals have to have time limits. Let's take a look at the pluses and minuses of microfilm in contrast to the Web. We'll start with the downside. Disadvantages of microfilm include

1. Microfilm has no full-text search capability. A few things can be bar coded, but at most this usually gives you access to a frame number. Readers have the tedious job of paging through, frame by frame.
2. Microfilm requires visiting the library. You can't do it from your office or home.
3. Microfilm readers are often difficult to use, and in actual practice are often broken (partly because so many of the readers are unfamiliar with them and do not know how to use them), and sometimes produce either headache, eyestrain, or backache as a result of the chairs and lighting

conditions available. Film (or fiche), especially older photography, can be difficult to read.

4. Material on microfilm is usually not as up to date as even printed material, since normally it takes at least a few months from the time something is printed until the microfilm edition is shipped (if only because a number of issues of the journal or paper have to be collected). For those interested in timely information, microfilm won't do.
5. One microfilm can only be read by one person at a time; digital copies can be read simultaneously by many. This is rarely a problem, given the low use of most filmed publications; and of course what is listed here as an advantage of digital is considered a disadvantage by publishers.

Advantages of microfilm include

1. Microfilm contents are, in a way, twice-selected, at least for those microforms which are alternative versions of print publications (as opposed to films of original manuscripts or business records). This material was originally chosen for publication, and then it was selected for filming rather than just discarding. Thus, it's usually reliable and high-quality information.
2. Microfilm is known to be durable and accepted as a long-term preservation medium. This is more interesting to the librarians rather than the users.

Similarly, we could look at the CD-ROM versions of publications which flourished through the late 1980s and early 1990s. Like microfilm, CD-ROMs saved shelf space and avoided damage to paper originals. However, they share many of the disadvantages of microfilm. The user must go to the library to use them; a special machine is required; and often no search capability can be offered (since many early CD-ROMs were just page images of the printed version). Again, CD-ROMs are losing out to online digital.

8.4 Finding Library Materials

Perhaps most important among the problems we have discussed is that of search and access. Comparing the combination of card catalog and shelf browsing with the modern electronic full-text search, certainly we can find content more accurately, but not necessarily style or genre. The question of what kinds of queries users have, of course, varies not only with the users but also with the library system and the user experience. Users will learn to ask queries that they have found can be answered. Further, users choose which library to go to depending on what they want; there is little point in going to a research archive if you want a children's book or to a science library if you want literary commentary.