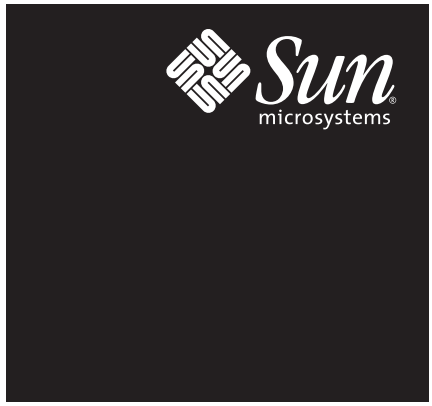
The Digital Library Toolkit



January 2003
Third Edition

The Digital Library Tool Kit

by Dr. Peter Noerr

Preface

The *Digital Library Tool Kit* was sponsored by Sun Microsystems Computer Company in the hopes of addressing some of the leading questions that academic institutions, public libraries, government agencies, and museums face in trying to develop digital content and distribute it on the Worldwide Web. Librarians and Campus CIOs are dealing with a plethora of new technologies and issues in the realm of Digital Libraries. The evolution and coalescing of Java applications, digital object standards, Internet access, electronic commerce, digital media management models, search engines, and library automation systems, is causing educators, CIOs, and librarians to rethink many of their traditional goals and modes of operation. As one of the leading technology providers to the Education and Library Communities, Sun discerned the need for a comprehensive, state-of-the-art document that could give some guidance in the nascent field of Digital Libraries. We hope *The Digital Library Tool Kit* is of use to you as you explore new directions and technologies.

Art Pasquinelli
Group Marketing Manager, Knowledge Enterprise
Global Education and Research
Sun Microsystems, Inc.

Purpose and Scope

This third edition is an update and expansion of the previous editions. It contains more of everything. In particular, the resources section has been expanded and updated, and a chapter on the impact of the Internet is added.

This document is designed to help those who are contemplating setting up a digital library. Whether this is a first time computerization effort or an extension of an existing library's services, there are questions to be answered, decisions to be made, and work to be done. This document covers all those stages and more.

©2003 Sun Microsystems, Inc.

The first part (Chapters 1–2) is a series of questions to ask yourself and your organization. The questions are designed generally to raise issues rather than to provide definitive answers. It also includes in Chapter 2 a look at the issues raised by the pervasiveness of the Internet and World Wide Web.

The second part (Chapters 3–6) discusses the planning and implementation of a digital library. It raises some issues which are specific-specific, but does contain answers and information to help answer these and a host of other aspects of a digital library project.

The third part (Chapters 7–8) includes resources and a look at current research, existing digital library systems, and the future. These chapters enable you to find additional resources and help as well as show you where to look for interesting examples of the current state of the art.

This document was produced on commission to Sun Microsystems Computer Company. It is available as an Adobe PDF document on the Sun Web site at www.sun.com/edu, or on the www.EduLib.com and www.MuseGlobal.com Web sites.

Dr. Peter Noerr
peter.noerr@museglobal.com

Table of Contents

Part 1

52 Questions to Ask Before Creating a Digital Library

...Or "Do I Want To Do This?"

This section raises questions and discusses the issues raised so that you have an insight into the full ramifications of the project you are about to undertake.

For most of the questions there are no "correct" answers. The intention is to make sure you consider a wide enough range of potential options and think seriously about whether the "obvious" answers actually are that obvious.

It is also important to consider how the topics will affect, and be affected by, your specific project and its wider context. Doing this topic by topic will focus your attention and eventually give rise to a picture of the whole of the project and its interaction with the current and future library world.

Chapter 1

The Questions (And Possibly Some Answers)

The questions are divided into broad areas. Each question has a brief discussion of its topic and/or a number of subsidiary questions. This is intended to raise issues that may have been overlooked, and to give the reader a reason to pause and consider what he/she is contemplating doing.

The discussions center on practical issues and considerations rather than on theoretical ones. There is no intent that this document should be used as a defining thesis. It is intended as the first stage of a "how to" guide.

What Is...

...a Digital Library?

Conventionally there are two possibilities:

- A library that contains material in digitized form
- A library that contains digital material.

The difference is sometimes very subtle and is discussed in "...Digital Material?" on page 4.

The really important point is that a digital library has material stored in a computer system in a form that allows it to be manipulated (for instance, for improved retrieval) and delivered (for instance, as a sound file for playing on a computer) in ways that the conventional version of the material cannot be.

©2003 Sun Microsystems, Inc.

An automated library is not, per se, a digital library as a library consisting entirely of conventional physical material (such as only printed books) may be very highly automated. This automation does not make it "digital" in the sense we are considering here. However, it is true that a digital library must be automated in some of its essential functions.

Because the material is in digital (or computer readable) form, some new possibilities are opened to the digital library that are not there for a conventional library, even one with the same material.

As an example: the material delivery process can be very different from the removal of a book from a shelf and checking it out. Because the "book" in digitized form can be copied to a user's computer for reading, but still remain in the computer "stacks," it can immediately be "loaned" to another user. This implies that holds (reservations) could become a thing of the past for a fully digital library, at the expense of a very much more complex usage tracking system.

This example has added complexity when the digital full text may be stored as a repository outside the library. The material may be downloaded to a portable reading device (an "e-book") or it may be sent to the reader's private permanent store (a personal digital library?). Ownership, rights management, and commercial considerations become much more complex in this environment.

...Digital Material?

In this computerized day and age information and the medium on which it is recorded can be considered as either digitized or not. There are many other ways of categorizing the material, but computer readability is the important criterion here.

"Digital" can be taken as a synonym for "computer readable." This is a serious generality, but it is this aspect of information that is most relevant to a digital library. The creation of digital information from conventional is generally a two-stage process.

The first stage is digitization. This is essentially the conversion of the physical medium into a digital representation of that physical medium. It takes no account of any information content of the original material, in the sense people would generally recognize the term.

Consider the process of scanning a piece of paper. This produces a computer readable (digitized) image of the paper within the computer. It is stored as a computer file that can be manipulated as any other file (i.e., it can be sent from computer to computer, or be copied or deleted). The original paper could be a page of text or a picture or even blank. We recognize different information content in each of these cases. The computer only recognizes a digital "picture" of them.

The second stage of the computerization process is to have the computer extract information from the digitized image. For text this is done by Optical Character Recognition (OCR) software that recognizes the shapes of the letters of the alphabet and produces a file exactly the same as one produced by a word processor used to type in the same text. Feature recognition software can perform an analogous process on a digitized image of an original picture, or on a sound file from a tape or record original. This stage allows some of the information from the original page to be made available to the computer. Thus, it is now able to index the text for retrieval and is also able to reformat the text for different forms of output. Note: At the digitized image stage it is only possible to perform so called "graphical manipulation" such as stretching, compressing, turning color to black and white, etc. on the image.

All this processing applies to material that comes originally in non-digital form. Most of the existing stocks of libraries are in this form. For this material to become part of a digital library the material must be at least digitized or, more usefully, be converted to computer-manipulable form. This is the process that takes time and money.

Once this process has been completed for an original object, such as a book, the library now has three alternative representations of the same object. They each have different properties and allow for different activities.

It is important to realize both the power and dangers of this information in its different forms.

This process of digitization and conversion is not perfect. There are losses of information. However, with some intellectual input it is possible to apply a reciprocal process to reproduce a facsimile of the original. In its simplest form this is merely printing out the text file form of the original document. If enough information has been captured or recorded by human input, then the reproduction may approach the level of a duplicate, matching (in the case of a printed document) size, style and typography.

	Original	**Digital Image**	**Digital Information**
Physical Form	Physical object (book, video)	Computer file	Computer file
Format	Varied (English text, VHS)	Graphical file (.BMP, .MPG, etc.)	Structured file (.DOC, .MPG) Database and index records
Readability	Human or special equipment	Computer graphics program	Computer text, video or database program
Reproduction	Physically duplicate original (photocopy, duplicate)	Copy file and print any number of exact duplicates	Produce original information in different form (re-print book in large Italic type, play video with different sound track)
Manipulation	Physically modify (write in margins, cut and splice tape)	Mark electronically and manipulate graphically (add user specific notes, reduce/ enlarge, re-sequence change colors, paste alternate images)	Edit the original information, produce derivative work, copy and distribute endlessly

Today increasing amounts of material are originally produced in digital form. These have, in one sense, no physical presence other than the computer file that is their original form. Thus there are an infinite number of ways they can be realized for human consumption all of which are valid, yet all of which vary in the sorts of detail that abound in physical objects.

As a particularly nasty example consider a computer program. This has two very clearly distinct methods of reproduction for human consumption. One is a listing of the program commands in a programming language. The other is the program itself when it runs and interacts with the user. Which is the more “correct” representation? And how should it (they?) be represented in the digital library?

©2003 Sun Microsystems, Inc.

A more common example would be the text displayed or printed from a digital book. Most people would see no problem in utilizing a copy of a textbook in Times Roman font. However rendering the text of a Medieval manuscript into the same format would be seen as a great loss because of the departure of the illuminated script and possible marginalia and so forth. However the "plain text" version of that same manuscript is now amenable to textual analysis to help determine authorship and other interesting information that would have been impractical except as a lifetime's work.

...the Bleeding Edge of Technology?

This is where many digital library projects have foundered. As an example consider the digitization and conversion process which is at the heart of many libraries' problems.

Digitizing and conversion of the images to information are very difficult exercises. The computer hardware and, particularly, software which perform these functions are good and practical, but less than perfect. Many companies, Sun included, provide hardware and software that are excellent choices for digitizing and conversion projects.

- Optical scanners have suitably high resolution, but are mechanical devices. They break down, and they need single sheets of paper, which often means photocopying the material first.
- Computer disks fill up and files get lost or overwritten.
- Optical Character Recognition has errors that have to be manually corrected or ignored. Characters are not recognized or are incorrectly recognized. Non-Roman character sets cause havoc.

All of this means that the process which seems so swift and painless in the salesman's hands with standard texts and simple requirements, may become a painfully expensive reality. This is particularly true if material is old, in mixed languages, faded, or just voluminous. The question to ask here is definitely "Is it worth it?"

Just as much material will never be catalogued onto an automated library system because it is not used, so much should not be shoveled, in its entirety, into the computer as digitized images or information.

Even once digitized, the problems may not be over. Searching, and finding, material can still be a rather hit and miss affair for digital material. Delivery mechanisms such as streaming download of a video to the user's home computer sound fine and perform well in the laboratory. But the real life Internet gets in the way and low bandwidth connections, clogged trunk routes and incompatible Browsers all mean a less than happy experience for many users.

...Automatic Indexing?

This is the extraction of the information for a bibliographic record (the "metadata") directly from the original text by a computer program. It is particularly concerned with the extraction of keywords as an indication of the content of the document. Often it is called free text or full text indexing. They do differ, but an important part of their appeal is the automatic extraction of the indexes from the text.

Its advantage is that there is no human intervention. Thus it can be run continuously and cheaply. The extraction (indexing) process is the same for all documents and thus avoids the idiosyncrasies of individual cataloguers. Authority files (or lists) can be used extensively to further ensure consistency.

Its disadvantage is that there is no human intervention. Thus a document is characterized by the frequency of certain words and these may give a very wrong picture of the actual content. Standardized subject headings are difficult to apply, as they must be matched through statistical means. Here again the chances for deviation are quite large.

However Web searching is performed against such indexes of Web sites and is thus a common form of user interaction. Until more powerful tools come along this may be the way your users will be forced to interact with the digital library, and automatic indexing will be what you use to serve them.

Policy

Is There a Need for a Digital Library?

Does the library have a collection of purely digital material? Is it required that this material be delivered directly to users' computers? Do users need to search in "non traditional" ways for the material they need? Which material types in the existing library would benefit from being digitized? How? Is there a need for multiple copy distribution at the same time? Is there a need for the material to be modified and returned to the library?

Are there no other sources for searching and retrieving the material that you wish to digitize? Is the material unique or confidential? Are the users geographically and temporally widespread?

Generally the more "yes" answers, the more a digital library is a sensible proposition.

Is the Current Library Expanding?

If it is, then is the new material obtainable in digital form? Particularly if the library is essentially an organizational report repository then material can be acquired in digital form and the costs of setting up the operation as a digital library are reduced. The same question applies to externally acquired material. If it can be acquired in digital form (or can be referenced in digital form) then the ongoing cost of the library is reduced.

If the library is static then the cost equation is different as it is the availability of the existing material in digital form that makes a difference to costs.

Is the Library Central to the Specific Project?

Does it contain material that the workers in the organization need to access as part or all of their work? A library maintaining a current report repository would be central whereas a report archive would not.

The question really is whether the cost of digitizing the material is justified in terms of the use that will be made of it. Will digitizing actually make the material more used? Will it make it easier (and hence quicker and cheaper) to find the material? Will the material be in a more useable form?

Here the cost/benefit matrix becomes more complex. A report-holding library where the reports have to be printed in hard copy to be marked up (filled in) and then have to be re-digitized does not seem to have any benefit over a paper library. A library where the videos are always retrieved by producer name (and they are filed that way) does not need advanced searching features.

©2003 Sun Microsystems, Inc.

How Valuable is the Library's Information?

This is a sub-question of the previous one. If the library is central then, presumably, the information is important. If the library is used only occasionally, but intensively (as in the reference section of a library attached to a research laboratory), then it is still valuable. The best benefit may be obtained through improved search tools for internal and external information rather than digitizing everything in sight

If the library is used infrequently or is an archive then the cost of digitizing may still be justified, but on different grounds.

The library may have been bypassed as an information dissemination center and digitizing its collection(s) may be a way to re-position the library into a more central role. This may well have cost implications for the organization as a whole rather than just the library or its parent division.

Is the Information Changing?

If information is changing, then the ongoing costs of re-entering the information into the system will need to be considered. Also important is the policy for handling different versions of documents within the digital library.

An alternative scenario is that the focus of the library (and presumably the organization) is changing. If this is the case then it is prudent to consider if now is the time to start a digitization program. Much of the material may be irrelevant in some months' time. However, this may be just the time to start the process as the new material can be acquired and processed directly into a digital library system. The material may be acquired in digital form or the costs may be justified by one-time processing as the new material is acquired.

Do the Library or Organization Want an In-House Digital Library?

This is a methodology question. It may be possible to outsource all, or most, of the functions of a digital library. Alternatively, it may be possible to buy access to external search and delivery services that cover most of the library's requirements in the digital area.

This can save on initial digitization and on-going administration costs. Service organizations exist which will operate the whole digital library remotely in complete security and confidentiality, from the initial digitization through the ongoing provision of the search and delivery services.

However, they do not, in general, have the subject expertise of the in-house library staff and there may still be security and service availability issues to consider. A factor, which modifies this discussion, is the growing prevalence of Application Service Providers (ASPs). These organizations do nothing but provide the library with access to its particular application requirements in an environment that is staffed with experts in the particular application area. Although this is not a common business model in the Library Automation area at present, it is a growing trend and this is a factor that can change many of the economic numbers if a suitable partner ASP can be found.

Should a Digital Library Coexist with a Conventional One?

This is really the "all or nothing" question. Is it intended to replace the existing library with an all-digital material version, or is it intended to supplement what the existing library does with new services?

The question here impinges on the perceived role of the library in the organization, its efficiency, how the organization is going to handle internal (and external) communications in the future, and if there are different functions and facilities provided by the physical library and the digital one. Bearing on this are the issues of branch (or local campus) libraries, specialist knowledge, specialist collections, centralization vs. decentralization, and the other spin-off functions of the library, both physical and digital. Serendipity is the word here, especially in the supply of tangential material, the answering of "strange" questions, the place to work quietly, the superior analysis tools available on-line, and the benefit to the organization of informal communications.

Is the Object to Run a Library or Manage Material?

Is the organization trying just to provide just facts or to provide more extensive services? Is the real purpose of the library to be an archive? Is it to track the routing of reports and files? Is it to perform research and analysis services for the organization?

These sometimes hard questions need to be considered, as the capabilities of a digital library are very different to those of a conventional (physical) one. The true function of the library in the organization should be written large on the justification for a digital library.

Audience

Is There a Demand for New Services and/or Material?

Is the digital library proposal coming from the library users or is it generated by the library staff or the computer (MIS) department? This is really a question addressing the issue of how the organization decides on the introduction of new services. Is it "market driven" or does it follow a planned introduction of services or technology?

If management is proposing the digital library, it is important to determine that it will prove beneficial for the organization and its potential users. If the potential user demand extends outside the organization it becomes a marketing and business case to determine which, if any, digital library services and information sources are justified.

Has the Market Been Sized?

How much use will be made of the library and its services? Is the (potential) user population large enough to achieve the organization's goals, whether they are cost recovery, better information flow, or even corporate publicity?

There are no absolute numbers for the user population as it depends on the services, their cost and the desired return. However, an estimate of user numbers must be made so the global benefit can be discussed.

How is it Composed?

Is the user population internal or external to the organization? Does each of these groups consist of different types of users (e.g., students, research workers, teachers, etc.)? Do the different types of user want (or form a potential market for) different digital library services or information resources? Are they willing to "pay" different amounts for the different information? How do they use existing conventional services and how will a digital library affect this use?

©2003 Sun Microsystems, Inc.

How Will a Digital Library Be Used?

Will users be offered different (new) services that are not currently available? Will some of the services replace those of the conventional library? What changes in the way users work will be introduced by the advent of the digital library? What changes in the rest of the organization will be needed to accommodate the new method of operation? What will be needed to make best use of the digital library? Is there likely to be an improvement in conventional library services as a result of converting some services and information to digital format?

As an example, the delivery of material directly to the user's desktop can have a profound impact on work patterns. For a magazine publisher where the library holds the picture library, careful design and integration of the remote digital library holding the magazine's pictures means that the journalist can drop straight from his/her word processor into the library search routine, find an appropriate picture, and copy it directly into the page. This may be appropriate where the journalist is also responsible for page make-up, however, it may be a better use of resources if the journalist gets only a low resolution "place holder" picture and the high resolution image is queued at the final printing equipment for inclusion at that stage. In this example both the method of working and the nature of the information held by the digital library are candidates for modification. This is generally an iterative process to achieve the best result.

How Will a Digital Library Be Accessed?

Since digital libraries are held within computers it is important to realize the possibilities for access that are offered and are denied. Access can be permitted from the user's desk wherever that may be. Traveling workers can be given access directly or via the Internet if desired.

If users do not all have computers then public access must be provided. Is this best done within the physical confines of the library or should/could it be distributed across different buildings? If these "library access stations" are used then they will need equipment and maintenance, but will be physically close to the users. It may be a good idea to add some extra facilities to the stations, such as reference material, chained copy of the staff handbook, color local printer, etc., so that they become specific work points. This is leading the deployment of the digital library within the organization in a particular direction with certain benefits and costs.

Another access model is to assume all access will be from desktop computers. These do not need to be fully functional PC's. Network computers or clustered workstations (such as the Sun™ Sun Ray™ machines) can be tied to a local server and provide inexpensive access to users who do not have full PC's. This model has the advantage of cheap deployment, easy desktop access and access to a wide variety of applications, possibly held within the digital library itself.

There are many models, even within a defined IT strategy, and the choice of one (and its modification to local requirements) needs as thorough a study of the users as of the library itself.

Since access via the World Wide Web is becoming so important, discussion of this has now grown to its own chapter (Chapter 2), where many of this issues specific to Internet and Web access will be raised, and many of the general access issues will be seen in a particular context.

Is There Competition?

If a great deal of the library's resources comes from outside the organization, then there may be alternative sources that either the library or the end users can access. This could enhance the library's case if it acted as the conduit for this material and add value. Or it could diminish the library's case if the material was easily and, possibly, freely available through external sources directly to the users desk.

The advent of specialized ASPs and their services mean that even the information market is being served in different ways. A specialized search service that was individually subscribed to and could deliver results on a par with those of the library could be a very powerful competitor for the hearts and minds of the potential audience. If the library can set up a portal of its own (see Chapter 2) and provide access to its own and external information sources, then it is providing a benefit that will win it many friends. The question here is really how much of a step the library is willing to take?

Reasons

To Expand Services?

Are digital services being added to expand the repertoire of the existing library? Or are they seen as a replacement for existing services? Are the services (and information sources) being added complementary to the existing ones or do they break into new service territory? Are there resources to expand the services? Are the new services supposed to be self-funding? If so, how is this calculated and what are the expected figures?

To Make the Library More Central to the Organization?

Is the library taking on a more central role in the organization and is thus expanding its services or, are the services being added to make the library more central? Are the new information and the services fundamental to the operation of the organization? Again, if the library is expanding and taking a more critical role, does it have the resources and equipment to fulfill this position?

To Generate Income?

Is the intention to sell the new digital services and/or information? If so, are they to be sold internally or externally? Are they to be sold for real money or some form of internal credit transfer? How are the prices to be set? How will the users come by the money to pay for the services? Will priced services just drive them away? How will the costs be recorded and accounted for? Bear in mind that most library software does not have the capability to charge for services on a per transaction basis. Even recording "logged on" time is often not possible and this has a very detrimental effect on user satisfaction.

If income is generated, then the matter of copyright payments and even taxes becomes important as well as possibly changing the status of the library. In addition to these regulatory and legal matters, it may well be that the bookkeeping associated with charging and collecting the fees costs more than the income generated.

©2003 Sun Microsystems, Inc.

To Promote Collections?

Some libraries have unique collections and the promoting of more widespread use of them is one common aim. This is particularly true where the collection is one of rare and expensive material. Fragile material with special handling needs is another good reason to digitize the collection. In all these cases once the digitization process has been undertaken the original material can be returned to its preservation environment. Because the digitization is done by experts and only done once, it can be painstaking and therefore mindful of the preservation and security needs of the collection. The originals will then be required for study much less often as the only reason to use them is now to study the actual construction of the objects, not their "information" content. Obvious examples are rare books, manuscripts, pictures and the like.

It may be that a single collection (or exceptionally a single work—think of the British Library's "Beowulf" project) warrants the expense of digitization because of its rarity or value or utility. Since the decision may be made on publicity or "public benefit" grounds, the actual cost may not be important.

To Raise the Library's Profile?

The library may feel it needs to undertake a project to raise its profile either internally within its parent organization, or externally. This is a perfectly good motive, but it should be understood as the motive and not hidden.

Because of Staff Pressure?

Similarly to the above, it may be staff pressure that suggests the creation of digital collections or the conversion of the library to a digital format. This may be from a desire to better serve the library users, to better exploit the collections, to experiment with new technology and techniques, to continue to be part of some external organization (such as an information-sharing consortium), or to undertake interesting and challenging projects. Staff may wish to enhance their professional training and remain abreast of current technology and thinking in their field.

Alternatives

Do Nothing?

Instead of creating a digital library, it may be just as effective to continue with a conventional library or to upgrade the library or information service in some other way. This may be a service consideration or it may, eventually, be a cost-based one.

If the library is providing a good service and the users are well adjusted to, and happy with, what is provided, then there may be no case for digitizing even part of the collection. One of the problems here is assessing if best use is being made of the collections and if the users are being most effectively served. To do this it is necessary to take a larger perspective than just that of the library. Often users do not know what alternatives could be made available to them and thus they "don't know what they are missing." They may be satisfied, but not realize more is possible. The same is true for library staff who may well be doing a sterling job with the resources they have, not realizing that alternatives are available.

Introducing a digital library just because it is a technology that has caught someone's eye is wrong. What is needed is to consider both the needs of the users, the resources in the library, the requirements of the organization, and the whole spectrum of available improvements.

This document asks many of these questions without a bias towards digital libraries. It is important that there is a person on the decision-making team who is able to suggest suitable alternatives.

Out-Source?

Even if it is decided that a digital library is needed, it may be that the best way to achieve it and run it is to give that task to another organization.

The process of retrospectively digitizing the material from the chosen collections may be highly specialized and need expensive equipment for best results. Since this will be done once, contracting for the service may be the most cost-effective way. Even ongoing conversions may be best handled on a bureau basis. Equally some of the conversion work may be mechanically very simple and repetitive and hence not a good use of a highly skilled librarian's time. There are companies which undertake nothing but this type of work.

Once the material has been digitized it is worth considering the costs of making it available, particularly if the service is to be provided externally and will be paid for. Availability around the clock may be required. Large computers may be needed to handle peak loads. Staff and facilities will be needed to back-up and secure the data. These, and other considerations may make it a sensible proposition to contract out the actual running of the service. There are all sorts of levels at which this can be done, with different degrees of security and service, and, of course, different costs.

It may be that running a service for external users through the World Wide Web is considered an option. This may be best run through a commercial provider (an IPP or Internet Presence Provider), particularly in the early days. This lessens the requirement for what might be extensive capital investment and can provide much needed expertise in areas where the librarians and the organization itself may have little or none.

Several library automation vendors use the ASP (Application Service Provider) model, but this hasn't proven to be nearly as effective or popular as it was initially thought. This still may provide yet another avenue for moving much of the cost or the risk into a purely financial monthly fee. The ASP has the specialist skills for migration both of the material and of any existing library processing. They have invested in all the expensive capital costs of equipment and infrastructure. The widespread use of Intranets means that such an organization may be able to service both the external world and internal users.

Provide a Gateway?

If the library is considering primarily providing access to already digitized material and/or material acquired from other parties, it may be sensible to consider a gateway operation. In this the library is running a service that is only a re-direction of the users' questions to the holders of the original digital material. The gateway may provide its own indexing and search services and it may combine original resources from a number of different providers. Gateways are becoming quite common in the library world.

The major difference between outsourcing and running a gateway is that the outsourcing is entirely of your information and is an operation being run for you by a third party. A gateway is where your operation is linking to independent third party sources.

©2003 Sun Microsystems, Inc.

As an example, in an outsourcing operation the library would have to acquire copyright clearance to disseminate all the material it acquires from a third party. It would do this at the time of purchase of the material in the form of a contract that, almost certainly, required the library to be responsible for counting usage and assessing and paying the fees to the owner. In the case of the gateway operation the third party source has negotiated such a license and it merely bills the library for each access to a particular piece of information. It is the third party that is responsible and undertakes all the accounting.

Both outsourcing and a gateway service can be provided through the organization's computer addresses, with the organization's logo and house style for the screens. If designed properly, the eventual users would be unable to tell if the digital library were in-house, outsourced, or a gateway operation.

Costs

What are Start-Up Costs?

The variability in start up costs depends mostly on the material to be included in the digital library. Volumes are an obvious factor. The type of material and the degree of digitization and the completeness ("resolution") of the digitization also affect the cost. Care and attention to fragile material adds to the cost.

All the above apply to local material to be digitized. If the material is to be acquired in digital form then it has an obvious cost.

Once the material has been digitized, it has to be loaded into a suitable library application. This will store the digital material (often in a database or in a file store), index it and add it to the library housekeeping database. There will need to be normal library housekeeping operations such as authority file maintenance, assigning material to access classes, determining library policies, etc., just as in any library system. The library system will have to run on computers. Either a large server and workstations or a server and network computers or a network of computers will be needed. The configuration and the need for specialized computers such as video servers, must be determined in consultation with the supplier of the library system software.

Networks and network and application servers may need to be set up or upgraded. Bear in mind that digital files are generally very big and thus take a lot of storage space and are slow to transfer across a network. Thus a network that is perfectly adequate for office automation may well be totally inadequate for allowing users to view video, even one at a time.

Added to all the above capital expenses are the staff training costs for both library staff and end users.

Disruption inside and outside the library adds a cost to the whole exercise. If the service is to be made widely available (particularly outside the organization) then it must be advertised and promoted in some way.

If the service is to be made publicly available then there may well be registration and licensing costs involved as well as trademark and name protection.

If some services will come from external organizations, then there is an obvious (though not necessarily easily quantified) cost attached to them.

What are Ongoing Costs?

Addition of new material to the library incurs the same processing (or purchasing) costs as when the retrospective conversion was done. If there is a regular flow of material then it will be possible to negotiate a reduced per unit rate with a conversion specialist.

For externally networked services there may be telecommunications charges, particularly if the service is transferring large amounts of data across a third party network (such as the Internet or a public carrier network).

Every so often the physical capacity of the computers either to store material or to handle the number of users will be exceeded and they will need to be upgraded. The organization's IT policy and infrastructure will change and the application and data must be migrated. Regular backups of the data must be made, checked, and archived.

Staff will change and new staff need to be trained. Users will need regular training.

How to Reduce Costs?

First, determine exactly where the services have to be delivered. Accurately size the amount of material for both conversion costs and computer costs. Try to acquire gateway access to material which is not your own. Then determine which are the core collections and which are most needed. If necessary, aim for an 80/20 solution by dropping services or collections which are not extensively required or are infrequently used. Consider access to other suppliers for those services on an "as needed" basis. Determine exactly what use will be made of the material and, particularly, how it will be retrieved. Do not buy a retrieval (catalog) system which charges for features that your users do not want or will not use. Be careful of this last suggestion as it could lead you into a "low functionality" trap. It is always better to buy the more functional system rather than to sacrifice the possible function or service. The cost savings here are not generally large in the overall scheme of things.

Income

If the library is going to attempt to at least cover costs then all possible sources of revenue must be considered. Organizational limitations and licensing requirements may restrict what can be done. See also Chapter 2 with further information on revenue attraction, and the cost/benefits of various methods.

Direct payment for information is the most obvious income source. Usually this is for access to the full text document or the full video, etc. Access to index and catalogue records is not usually charged for, though this may occur in some cases. This payment may be on a "pay-per-view" basis that is cheaper for the occasional user, but a serious administrative problem for the library to collect the fees. Better is a subscription basis for either some counted use (number of visits, number of hours on-line, searches, texts read or printed, etc.) or for unlimited use (within certain functional areas) for a fixed period. It is quite common to allow on-screen reading, but not printing, as this has an added cost level.

Also a possibility is the provision of advertising to obtain revenue. This option has to be considered very carefully as inappropriate advertising (and even any advertising at all) may appear to compromise the integrity of the site. Sponsorship may be more appropriate than on screen advertisements, but advertising must still be carefully considered as an option.

Hosting of collections and services for other libraries is a possibility, but requires the investment in the professional staff and equipment to make this possible. It is not a "garage" operation any more.

©2003 Sun Microsystems, Inc.

Providing a gateway service or Portal for a group of libraries may be an option. Here you acquire the portal services (possibly outsourced to a third party), and on-sell part of those services to other libraries. This is becoming a fairly common e-business model. It makes most sense when there is some commonality with the other libraries, whether in material coverage or audience. Then there can be seen an obvious advantage to the libraries and the users to such an aggregation under one portal. This may be an actual service you run or it may be an idea you initiate at some level of regional co-operation so you may share in the portal as a partner rather than as the principle. Your organization or higher authority may not allow other than shared participation.

Sources of Material

Internal Sources?

Does the organization generate the material itself? Does it generate it in the original, physical form or in digital form or both? If a digital form is not currently produced, can the creator easily produce it?

A typical example here would be where one of the collections is to be the full text of internal reports. If the report writers were to be asked (or required) to submit an electronic form of the report as well as the current paper copy then a lot of time and expense could be saved in creating the digital collection. It would also save the necessity of correcting the mistakes introduced by the imperfect digitization process.

Does the organization have the means already at its disposal to digitize its material? Is most of the material going to come from existing collections? How much of these collections is unique and will have to be converted instead of possibly buying-in the pre-converted information?

Archives, Etc.?

If archives are involved then the biggest question is “Is it really worth it?” If the archives are rarely used then there is little reason to go to the expense of digitizing them, unless the reason for their non-use is that they are inaccessible in their present form. Obviously security and collection promotion considerations can radically alter the worth of the collection in digital form.

An alternative to wholesale retrospective conversion is to create a computerized index to facilitate access and then to digitize as required by use. In this scenario it would probably be sufficient to just create images of the archive material and not try to extract the information from them (if they are text). This would offer quicker and cheaper methods and would supply the advantage of electronic delivery.

External Original Sources?

If the source of the material is external to the organization then it is important to ask whether a gateway operation would be more suited to what is required. The copyright and commercial issues of dealing with this material become much more complicated than for owned material.

One possible reason for locally held, externally originated, material is to enhance the value of the whole library. Here it may be desired to manipulate the contents of the material and this has to be very seriously considered before it is undertaken.

The external material may be bought outright, its use may be licensed, its use may be leased, or its use may be on a "pay-per-view" basis. The material may be acquired through exchange or gifts just as conventional material is. Sometimes the attraction of the (proposed) library site to external users is such that the library may charge for external material to be made available through the library.

Delivery

Local Material and Delivery?

Is the original material in the library that will deliver it to the users? The digital component of the library may be a portion of the original material held in digital form for searching (such as full text) or for manipulation (such as a 10 second video clip to determine if the tape is the "right" one).

The general assumption about the use of a digital library is that the material will be delivered to the user at their desktop totally digitally. Thus the collections held in the digital library must be of sufficient quality for the user's needs. It goes without saying that for digital delivery the whole of the object must be available since the librarian has no means of knowing what aspect of the object the user is interested in.

If all the material is not stored digitally then some method of delivery must be devised. The user may be required to borrow the original physical item. The item may be digitized on demand. The item may be available from a branch library or remote delivery point. Note that the "digitize on demand" option does not necessarily require that the whole digitization process is undertaken. Just scanning and delivering the images will probably suffice in most cases. Some system designs envision delivery by fax in an updated version of the Inter-Library Loan (ILL) process.

Proxy Delivery?

If the user cannot be expected to have a suitably functional computer to receive the material or play it then an alternative is required. Most computers are perfectly capable of handling pages of documents as either text or images. But video presentations may be beyond their capabilities. Thus the concept of the proxy client.

This is a computer local to the user or at some special site (such as a computer lab) that is guaranteed to be capable of handling the material. Unless the library supplies special highly specified computers, the machines will have to be found from those available. This means that different machines may be needed for different material types. An unsatisfactory state of affairs. Thus it may be that the library has to either limit what it is prepared to deliver depending on what the user can play or read, or it has to undertake the supply of the necessary computers and limit access to certain digital collections to those machines.

Another aspect of delivery is the bandwidth consumed by high-resolution images or video. These may be a reason to look at a special delivery mechanism such as satellite delivery. This again requires specialized equipment at the user end, but it may be practical with a closed, or limited, audience for whom the added cost is acceptable or can be justified.

An example of this would be in an academic institution where the library plays a part in the storage and delivery of distance education material. This often involves video, simulations, programs to be run locally, etc. This material is bandwidth hungry, but the audience is finite and fairly small. The receiving equipment could be loaned to the individual students or the schools taking part in the lessons, or the cost could be reduced by grant funding or other means.

©2003 Sun Microsystems, Inc.

Where is it Delivered?

Is it to be delivered to the user's desktop? Is it sufficient to deliver to the user's building? Is it sufficient to deliver to the user's local library? Must the material remain in the library?

How Permanent is the Delivery?

Even if the material can be successfully delivered at a technical level, there is the question of ownership to consider. If a copy of the material is delivered for the user to do as they will, then they effectively have as much of the material as the digital library and could re-distribute it. This is different to the physical situation where there are only a certain number of copies and the act of copying is either physically impossible or prohibitively difficult.

One possibility is to consider delivering non-permanent material. These are files that are encoded and usually compressed so that they are played through the included decompression program. After a certain time the program refuses to play and the material cannot be used. This effectively introduces a "borrowing time" to the material just as with physical material.

If the material is delivered on a physical medium (say CD-ROM) then it is necessary to ask if the CD has to be returned. If so, then the library has to run a circulation system as for physical material. The ill-fated DiVX DVD format for commercial videos attempted to address this issue by allowing ownership of the medium (the DVD disk), but requiring the user to pay for access to the information (the video) stored on it.

What Capabilities are Required?

Are high bandwidth communications needed to deliver the material before the user has lost all interest in it? This would be the case for digitized video if it were sent over a network. A possible solution here is to use streaming technology and play the video for the user as it is transmitted. However, the user may wish to have an editable copy rather than just viewing it. In this case streaming will not help. It may be that hard copy delivery of a CD-ROM or DVD-ROM is the best answer even though it reverts to physical object delivery. For this a CD (or DVD) writer would be needed at the library or a stock of the CDs would have to be held as with any conventional material. This introduces the physical object handling (circulation) problems or the decision that the material is not returned. This has both cost and copyright implications. Much software today is available over the Web for free download or it is available on CD for a modest charge ($5–$10) to cover the medium and postage and handling.

If large files are to be delivered to the user then she/he (or the remote computer at which she/he is working) must have the capacity to store them. In the case of a researcher this may mean very large capacity indeed and, if the work is being done at a library carrel, the library may be involved in some of this cost.

If material is to be printed, then fast capable printers are needed and the delivery system (the library, the file transfer, and the printing systems) must be capable of handling network interruption, paper jams, etc., without imposing an undue load on the rest of the operation.

Special multimedia presentations (often not included under the digital library's umbrella, but they increasingly will be) may need sound cards, graphics accelerators/3D graphics cards, big screen monitors, etc., to be played properly and these will generally not be available at the user's desktop. The solution here may be a proxy delivery (see "Proxy Delivery?" on page 17).

Deliver Security

Various methods exist to ensure the security of material delivered over the Internet. This covers both the prevention of unauthorized access if the material goes astray, and the insurance that the material does not go astray in the first place.

The standard file delivery mechanisms such as via HTTP or FTP protocols have built-in mechanisms to ensure that all the packets of your material are delivered and re-constructed correctly. These, mechanisms combined with the robustness and security of the underlying TCP/IP delivery infrastructure should ensure that the files reach their intended recipient in pristine order.

However files can be "snooped" on in transit and copies made. This is where the encryption and secure delivery mechanism come in. Even if a copy of your material falls into the wrong hands, they should not be able to read it. Signed delivery and digital certificates which authenticate the parties at both ends of a transaction are mechanisms you can put in place. More recent mechanisms such as Microsoft's rather discredited Hailstorm/wallet and the Liberty Alliance where a third party vouches for both parties can be used, specially for paid for transactions.

Recent encryption techniques developed for the music industry claim that they are encrypted in such a way that files cannot be unencrypted without the key, and cannot even be copied. For the latest in protecting files in transit look at the RIAA (Recording Industry Association of America) Web site as they are pushing this development.

Copyright/IPR

Who Owns the Material?

The material may belong to the organization, or it may belong to an affiliate or subsidiary of the organization. It may belong to a third party. It may be in the public domain. It may belong to a foreign organization subject to different copyright laws. It may belong to an individual. Or, of course, it may belong to a mixture of the above.

Having a copy of the material does not necessarily constitute ownership in terms of copyright laws. There is only one copyright owner however many copies are made. This is true for computer copies (digitized or otherwise) as well as physical copies. Also be aware that the right to re-distribute material usually is not acquired when a copy of the material is bought.

Re-Use and Dissemination

Many countries allow for material to be copied for research purposes by individuals ("fair use"). However, making copies for re-sale or re-distribution is usually a matter for a commercial contract between the copyright owner and the organization wishing to re-distribute.

It is usual that the owners require that payment is made if all or part of the whole of the material is disseminated. It may even be that the owners require that they are the source for the eventual distribution of the whole material object. This is the case for many publishers where they allow their journal articles to be catalogued and indexed in retrieval systems, but the publisher must do the delivery of the full text of an article.

Remember that the bibliographic records describing the objects in the library are themselves "intellectual works" and have copyright. They belong to the person/organization creating them so be clear of the limits on re-use if your catalog records come from some form of shared cataloguing resource.

©2003 Sun Microsystems, Inc.

Charging?

The digital library may not wish to charge for material, but may have no choice so far as copyright fees are concerned. If these fees are payable to the copyright holder then the organization will have to pay them. It may decide to absorb these fees itself as a service to its users. This is often the case where the library is exclusively used internally to the organization.

If the library intends to charge for the information it disseminates then the terms of its license with the copyright owner may change and the price the owner charges may be much higher than for "free" material.

Again, it must be pointed out that all the accounting for these costs and charges must be done for each individual transaction and thus it will be necessary to utilize distribution software that is capable of this level of detail.

Partial Delivery?

Accounting becomes more difficult when an "order" cannot be completely fulfilled. Part of the material has to be charged for and other bits not. This is particularly a problem where network distribution is being used, as the delivery has to be secure from end to end across the networks. This occurs across public networks such as the Internet (see Chapter 2).

If a network problem causes a failed delivery, or a corrupt file at the user end, but the sending system believes the material has been correctly sent and should be charged for, there will be a problem when the user is asked to pay for material which did not arrive. Equally if the material is sent entirely before the transaction is posted then it may be possible for a user to break the connection and obtain the material while the sending system believes it has not been correctly received.

These problems have been solved for the online delivery of software, but they are not simple and complex solutions would be needed by the library.

An increasing trend in software delivery across the Internet is for the user to acquire the file and then "activate" it by means of a key obtained independently from the distributing organization. A library could utilize this method for delivery so long as an "installation program" was sent to the user and that program could communicate with the library sending program. This is not a trivial exercise to set up and will certainly involve extra cost, which third party fulfillment would not.

Act as an Agent?

If the library does not own all the material it wishes to offer then it may act as an agent for the eventual owner. This removes many of the problems discussed above. However, if any of the material is owned by the library's organization it will have to address the issues.

If the library decides to distribute material which it charges for and invests in the necessary software and network safeguards, then it may wish to consider making those facilities available for other libraries to use. It could thus become a "database spinner" or host, for other libraries wishing to make digital collections available.

Fair Use

The concept of "fair use" is recognized by many legal systems and it allows users to make a certain number of copies of copyright material for purposes of private study. This was instituted upon the advent of photocopiers, but has become more important with the arrival of digital forms of material.

To make a photocopy it is necessary to have a photocopier and most individuals do not have them, certainly not of sufficient performance to contemplate copying whole books. However, every user of digitized material has the means to copy that material built right into the basic equipment. It is thus very easy for users to copy material and pass the copies on. The library would have no knowledge of this. This does not mean that the library is necessarily blameless for such copying, unless it took sufficient steps to make it difficult.

Security

Just like physical material, the digital material of the library is valuable. Access to it must be guarded and its well-being must be ensured.

Security for the digitized material must be provided in the form of restricted access to the computers that hold the material. This includes both physical access and electronic access across a network. The precautions are elementary and are well known in the commercial world. However, they may be unknown to the library. Unlike the physical stock of the library, the digital stock must be copied and secured. This protects it from natural disaster, malicious damage, and software errors.

Access security at least, must be allowed for in the digital library. The library may be freely accessible to all, but certain sections must be protected. Just as certain sections of the physical library are protected no matter how open the catalogues and stacks may be, certain areas of the digital library and certain functionality must be protected. If the material is not freely available then a method of restricting access to those who are allowed with a minimum of inconvenience must be arrived at. If the material is to be paid for then it must be secure until the payment is made (or promised) and then it must be correctly accounted for. The users must also be given an option to back out of any situation where they are about to commit a sum of money (see also Chapter 2).

The increasing use of smart cards and the concept of "user customization" means that security is now both more possible and more difficult than a few years ago. Systems which allow customization do not always extend that from the "look and feel" to internal functionality. Thus it may not be possible to block the dangerous functions when required, except by the introduction of passwords for all.

Specialized software now exists (such as EduLib's STOPit system) for use on library workstations which allows the functionality of those machines to be linked to individual users via their library card or some form of key. This solves some of the security issues and can add extra features such as gathering usage statistics and allow for interface design.

Watermarks and Other Protections

Earlier it was mentioned that the user's computer is inherently capable of copying any digital material on it. One protection against this practice is to introduce watermarks into the library's digital material. A watermark will not prevent copying, but it will mean that the owner of the copied material can be recognized. Modern systems can do this even if only part of the material (say a part of a picture) is copied. It is also not possible to overwrite one watermark with another without a special key.

©2003 Sun Microsystems, Inc.

The same protection can be provided to all digital material. There are various methods of watermarking, however, none of them are yet foolproof and they are often incompatible. Some require user-side software for full protection and this means they are limited to where this software can be mandated for access to the library. This is possible within an organization and may be possible where the collection is unique and valuable enough.

Material may be disseminated in degraded form or it may be only partially disseminated. In these cases either further verification of the user is required before the full copy of the material is delivered, or the material is delivered by some other means—usually as the physical delivery of a CD of the digital material.

Technology

Standards?

As with all matters to do with computers there are standards that impinge on the area of digital libraries. Unfortunately, since these libraries are at a cross-road, there are a number of standards which might be appropriate. Of course, some of these standards are mutually contradictory or even exclusive.

The standards fall into three areas: material description, user access and systems architecture.

Material Description

In terms of material description by far the strongest standards come from the library profession. Two forms of description have to be considered; the abstracted information (or metadata) which constitutes the bibliographic description in conventional library systems; and the material itself. In a truly multimedia digital library it is also necessary to consider the relationship between the various items and pieces of material and their different forms and formats.

Descriptive standards such as AACR2 and MARC from the library side here compete with SGML and HTML from the "Web" part of the computer industry and document description standards such as .PDF (Page Description Format) from the document handling community. These standards are not mutually exclusive, but there is a lot of overlap and converting from one to another is a function to be considered at the design and acquisition stage.

Recent entrants on to the stage are the re-vamping and rationalizing of MARC to MARC21, the increasing use of the Dublin Core set of descriptors (attributes), and the conversion of both of these and the full material to document types within XML(eXtensible Markup Language). These changes are promising to bring digital library systems closer to commercial systems in terms of interchanging actual material, but the difference in approach at the cataloging and processing (circulation vs. sale) stages is still large.

Non-bibliographic material (pictures, sound, etc.) is handled by the MARC format, but there are competing standards. For instance, for geographic information there are descriptive standards coming from the cartographic professions. Many of these are actually more interested in describing the original (the terrain) than in describing the physical material (the map) or its digitized equivalent.

This strongly suggests that until universal description frameworks (standards) are in place, it is very important to decide what the material is, what needs to be described, who it is intended for, how it will be retrieved, and how it will be processed and used before deciding on a scheme for its description.

The logical format of the digitized material and how convertible it is from one form to another is an important consideration as the wrong choice could limit the number of users who can "read" your material. This applies to how the material is held in the database or files as much as how it is described for retrieval.

A good example here of where an early decision may later prove costly is in the area of multilingual texts. Where they are not encoded in a unified encoding scheme (such as Unicode), they may not be readable except by specialized client software.

User Access

Basically, there are two methods by which users may access your digital library. One is via a dedicated network and the other is over public networks. Within both of these it is possible to have users access via dedicated clients or general-purpose browsers (see "Future Possibilities" on page 26).

Where the network is private and the client software is dedicated then the standards used are unimportant, with the proviso that any protocol that is not extensible and only supports the current functions of the library is unlikely to be a sensible choice, as it will become obsolete very quickly.

Where public networks are concerned two standards for system access exist in the catalog search area. One is the HTTP standard from the "Web'; the other is the Z39.50 standard from the information retrieval and library world. They are actually standards for different purposes from different backgrounds. But they can be made to perform the same search and display functions. They are seen as competitors and a system that supports only one may limit the types of users who can access your library. For general access from the Web an HTTP interface is needed. For access from other library systems a Z39.50 system is needed. To confuse the issue (but actually to help) there are gateway computers on the Internet that convert from one to the other, and services that can handle both (see also Chapter 2).

General-purpose browsers (Microsoft's Internet Explorer or Netscape's Navigator are the most widespread examples) are widely available (often for free). Thus, access by them is a requirement if the desire is to have the library accessible by the widest possible audience. However, they are page-oriented devices and are not ideally suited to the material structure and list-oriented nature of much library searching. By their nature they are not specialized and their capabilities come from the "plug-ins" or Java™ applets (programs) which can be added temporarily to them through the downloading of the program. Thus they are not as suited to specific tasks as specialized clients and software. However, the capabilities of Java enabled browsers are increasing at an extremely rapid pace.

It may be that the library offers a number of alternative methods of access depending on the requirements of the user. This provides a better service for the user, but is at the cost of the development and maintenance of the alternative access methods.

Systems Architecture

In most important respects a digital library is no different to a conventional library automation system. As such all the remarks, which can be made about the system architecture of library automation systems, in general apply.

The major differences are in the volumes of material to be stored within the computer and to be disseminated to the users in real time. These requirements suggest specialized subsystems to handle the work. This usually translates into independent computers to act as the servers and the

©2003 Sun Microsystems, Inc.

appropriate networking to ensure the information is delivered. These specialized servers have to work in conjunction with the regular library catalog and other modules. The software must have been designed from the outset to link and control these servers with minimum input from the user.

Architectural practices rather than formal standards are the norm here. The majority of systems are "client/server." This distributes the workload across the library's (server) computer and the user's (client) computer. An extension of the architecture splits the server "tier" into two so that there are three tiers. The two server ones are the "database" (or repository or resource) tier and the "application" (or business rules or intermediate) tier. The resource tier handles the storage and retrieval of the raw information, indexes, image files, etc. The application tier handles the processing of this information into a form suitable for the user.

There are a number of important questions. Can the software support these architectures? Can the server tiers all be held on a single physical computer (useful for initial smaller installations)? Can they be distributed across multiple computers (to handle growth)? Can they be distributed across a wide area network or the Internet?

If the software does not have a multi-tier architecture does it have sufficient capacity and resilience to handle the whole of your expected traffic on one computer? And what do you do when that computer goes down?

The current trend is in distributed, even widespread, computing and this tends to keep the capital costs down by allowing an organization to buy equipment incrementally and to utilize existing equipment and capacity.

The de facto networking protocol (standard) is TCP/IP. It is the standard for Internet traffic and as such has permeated most other networks. It is really an essential for any serious digital library unless the audience is very small and tightly networked on a different protocol.

Similarly the de facto operating system for "mission critical" projects today is UNIX®.

Proprietary Solutions

Many (if not all) of the components of a digital library can be bought or produced in-house. This applies to the creation of the digital material from the originals as much as to the system software. In both cases it is important to consider the economics of doing it in-house vs. buying in a solution. However, there is another aspect to consider in this debate and that is the issue of "proprietary solutions." Essentially a proprietary solution is one where the organization does the work itself, either literally using its own staff or it commissions a solution from an external supplier.

A proprietary solution has a number of plus points: the solution is exactly fitted to the organization's requirements, the organization has absolute control of future development, there are no license fees and conditions. However, there are, of course, a number of negative points: the organization is on its own, there are no other users, the external world may adopt standards or conventions which bypass or conflict with the organization's solution, development has to be done by the organization.

While many of the activities of creating a digital library naturally fall to the organization (such as the creation of the digital material and the cataloguing of it), much of the infrastructure does not (such as the design and programming of a DBMS and retrieval system). Many of the questions to ask here are very similar to those about standards.

Is the digital library to be publicly available? Will unskilled users access it? Will they use standard client software (such as Web browsers)? Will they be unfamiliar with how the material is organized? Will they need (or be allowed) to download material?

If most of the answers to the above are yes, then the system should be standards-based and that generally means commercially acquired. Or, at least, produced by a systems house knowledgeable about the standards in use and in prospect. The time and effort in re-inventing many of the wheels needed for a digital library are not worth it. Compliance and continued development are a large drain on resources.

Use of so-called "commercial" solutions means they will be externally developed and conform to standards. The organization can concentrate on building the digital library where it has the expertise. The components which are serious candidates for buying in are: Database Management System, Information Retrieval System, Library Automation System, Web Server, Delivery and Accounting System, Rights Management System. Not all of these are necessary (and the list is not complete).

Purchasing components trade off an initial capital purchase against a longer term saving on running costs. If purchasing components seems to be the way for some (if not all, and it does not have to be all) of the bits of your digital library then it will be important to utilize a systems integrator which is skilled in this area to ensure that the components are designed to fit together and to ensure that they do so fit.

ASPs are essentially systems integrators and this model provides possibly a middle path where the ASP provides the software and services from standards compliant suppliers, and puts them together in a bespoke fashion for the digital library. A perfect "mix and match" world is still some way off, but things are moving in that direction and standards are the essential glue that will keep it all together.

Scalability

Whatever the initial size and predicted growth for your digital library, everyone hopes it will be an instant success and the world will beat a path to its door. While sober judgment acknowledges this to be unlikely, it is something that must be considered.

How would the organization handle a phenomenal success? Does it have the infrastructure to handle it? What departmental restructuring would be needed if the digital library were 5 or 10 times more successful (in terms of visits or revenue or workload) than predicted? Could the organization capitalize on this success?

How would the staff handle a phenomenal success? What sort of increase in traffic and workload could the existing staff handle? How much extra traffic could an extra member of staff handle? Are there positions that are now "doubled up" onto one person that would require more? What about vacations and other absences?

How would the library systems handle a phenomenal success? Does the hardware have sufficient capacity? Can it be upgraded or must it be replaced? Can the software handle the traffic? Store the data? Retrieve the data? Handle the requests? Will it have to be replaced? How difficult will that be?

Most capacity planning exercises in the library field size a system (hardware and software) for a five-year life at the projected growth. Having done that, re-size at a growth rate of 50% greater. Then consider the effect of initial capacity at three times (or five times if you're feeling optimistic) the assumed value. Take the largest of these and try to accommodate those figures.

Future Possibilities

Within the technology area almost anything is possible. It is, after all, the advent of technology that has allowed the concept of a digital library to start becoming fulfilled.

Faster and more powerful computers are certain from suppliers such as Sun. This is good both for the servers and for the user's desktop computers. It is good because the software producers will find more and more things to do, all of which will demand more processing power.

The advent of the network computer (e.g., Sun's Sun Ray™ Information Appliance) allows a new architecture to emerge where the user's computers are not themselves heavy powerful computers, but the display devices for a "user server" that provides shared resources at the client end. This allows better matching of requirements to hardware and also allows for the automated distribution of software. This latter point should not be underestimated, as it is potentially a great saver of library time and resources. It is a feature you should ask your potential library automation vendor about.

New hardware for data capture (image scanners, video capture, etc.) will allow the real world objects to be digitized in greater detail and much faster. The extra detail will make them better representations, but will have downstream implications where storage and bandwidth may become the bottlenecks.

Improved networking technologies have brought 1 Mbit bandwidth to user's homes via dial-up lines. Internal LANs have become 100 Mbit or even 1 Gbit. This means that static computers will be able to handle multimedia requirements. 1 Mbit will play videos in real time. These numbers will continue to increase for the foreseeable future.

Wireless networks will become faster and cheaper. This means users will be able to access the information stores from more places. Background retrieval of information will probably become more common.

More complex data structuring will become possible within mainstream applications. This means that the library's material may be more completely described and linked to its immediate and broader context. This allows relevant answers and better information to be supplied to the user. However, it requires an order of magnitude increase in the storage and processing performance.

Preservation/Handling

Is Material Irreplaceable?

If the material is unique to the organization then can it be replaced? Is it subject to decay? Is it a collection of "one only" objects? Does it need to be handled to "read" it? Would an image provide a suitable substitute for most purposes? Do "copies" exist for security (such as photographs of manuscripts)?

Physical handling is one of the most destructive things that can happen to a fragile object. One of the best ways to preserve it is to limit physical access to it. This is a very strong case for creating a digital library of such objects. Most use does not require the actual object.

If it is irreplaceable, then undertake a preservation and security plan whatever the decision about making it available in digital form.

Is the Material Multi-Use?

A number of problems to do with multiple use can be solved through a library of digitized forms of material.

The first is that unlimited numbers of copies can be made available. This can be subject to commercial or legal restrictions. While there are some technical caveats, it is generally true that as many people as want can have a copy of a digitized object.

The second is that the users can view these objects wherever they want. (Again, some technical, and any imposed legal restrictions limit the universality of this statement somewhat.) This is more convenient for the user, more convenient for the library and generally more cost effective.

Physical limitations of the material or library can be overcome. Since users do not have to be in the physical library to use the object there is no restriction on numbers placed by the buildings or equipment. Of course, if the material is only available through equipment in the library then the limit is re-imposed, but probably with higher numbers. Preservation considerations of bright lights and acid fingers are no longer a concern.

Material can be simultaneously accessed. Thus a whole class can see an image in whatever detail is required during a lecture. Of course the digitized form has no physical substance and so some physical characteristics cannot be studied.

Material can be simultaneously interacted with. This use extends the current boundaries of multimedia and what is included in a digital library, but an interactive computer game can accommodate hundreds playing against each other, which could not be achieved in the real world. Educational programs, which involve the whole class in role-playing, are being developed and these can only be experienced in their digital form.

Material can be modified from the user's copy. It is very easy to "cut and paste" images from a digitized library book into a school report. With suitable technology the link can remain and live digital material (such as stock exchange prices) found through the library can be incorporated into reports that reference back to the original source in real time.

Chapter 2

No Digital Library is an Island

Increasingly access to digital libraries is via the Internet and the World Wide Web (the "Web" from now on). Not only is access increasing from the Web, but also many of the functions and content offered by a digital library may well be located remotely across the Web. The digital library may access content from other sites as part of its own site to provide a more complete and satisfying experience for the user. The digital library may utilize functions from across the Web, or that utilize the Web or Internet to make the library more functional for the user. For example, a library may access an online classification scheme to organize its collection, or it may add a meta search engine to its site to allow the user to collect associated information from other sources when reading one of its items.

We need a brief look at the technology involved and then to consider what the Internet and the Web can offer the digital library in the way of additional resources and content, how it can promote co-operation with other institutions, and how it can act as an enormous front door to bring people to your collections.

Internet and Web Technology

The Internet is becoming pervasive in modern computing and information processing. It is now taken for granted in many parts of the world, and its reach is extending ever further. It will be, if it is not now, the main access method for most digital libraries. At its heart the Internet is a simple concept—the devil is in the details.

©2003 Sun Microsystems, Inc.

More familiar to most people now is the World Wide Web. This is just one of the services residing on the Internet (others are things like email, chat rooms, user groups, video conferencing, telephone communications, the vast text-only Gopher databases, etc.). The Web is increasingly the access method of choice as it embodies graphics, sound and video as well as text, and is becoming much more interactive and universally available. It is now possible to access Web sites through cell phones and Personal Digital Assistants (PDAs) as well as through the traditional Personal Computers.

Basics

The Internet is basically a vast collection of computers all talking to each other through a network that links them. These may be large computers permanently connected to this vast network—these are called servers or hosts—or they may be personal computers connected only by a telephone link while the user is active. The network connections are copper cable or optic fiber or microwave relays or even satellite links. The nature of the link is not important to understanding the Internet and what it can do for you. All that is important is to know many paths link all the computers so that the whole system is very secure, and that these links are of different speeds at different places.

Every server on the Internet has an address. In fact it has two. One is an Internet Protocol (IP) address that is a number and is represented like 123.321.213.312. This is the number the computers use to find each other. It is the direct equivalent of a telephone number. The number of numbers in this is running out and a new numbering scheme, called IPv6—much longer and able to accommodate growth for many years to come—is being implemented. Newer technologies mean these numbers will be increasing hidden from the user, and the URL will become the paramount method of locating servers and services.

The second address is associated with the services the computer offers. It is a Uniform Resource Locator (or URL) that is a textual name for that computer. This is represented as a name and a suffix that indicates the type of service and possibly the country where the computer is located. This looks like www.sun.co.uk or www.museglobal.com. The exact meaning of the parts is not important here; you can find information about them in countless books on the Internet. What is important is the existence of an automated “telephone directory” which translates these URLs into IP addresses so that we can type the much more friendly URL and the computers can make the connection.

One part of the URL is important here and that is the initial “www.” This signifies that this URL refers to a World Wide Web service on that computer. In practice this means a Web Site.

A Web site is a collection of pages that contain whatever contents the designer wishes to place there. It may be static like names and addresses and product descriptions, or it may be dynamic as in the results of a search. This is where your digital library hits the Web. To make it available you have to create a Web site or work with someone who has one you can be part of.

To make the Web work there need to be two pieces of software involved. One is at the server where all your content resides. This is a Web Server. Its job is to respond to user requests and send your content to those who request it. It wraps all your content up in those Web page designs in a language called HTML (Hyper Text Markup Language) and sends them out to the user in a protocol called http (hypertext transport protocol). This is the “language” of the Web.

The other major piece of software lives on the user's computer and it receives this HTML and converts it into the text and images seen on the screen, which represent your content.

You will find details of Web servers, other types of servers, other types of services, and browsers in the Resources chapter. Suffice it to say here that the vast majority of the servers running the Internet right now are Sun machines and the original browser (called Mozilla), from which almost all are descended, was developed in part by Sun researchers. Sun has been in the Internet since the beginning.

Connections

While the technology of the Internet and Web is of great interest, it is not important to your digital library. What is important is how the Web in particular, facilitates connections at the level of the content and services of your library.

The basis of the Web is the idea of Hyperlinking (and Hypertext and even just plain linking). This notion is based on the viewing a page of content and being able to link from any object on that page to another page. The purpose of this is to allow the simple object on the page to be expanded and to provide detailed information or context or related objects to the user when needed, while keeping them out of view for simplicity and clarity on the original page. Thus an author's name could be hyperlinked to a page giving his or her biography. An image could be linked to a page giving the physical description of the image and its provenance. A name within that provenance page could be linked to details of that owner, and so on. The idea of linking in this way is not new, data modeling for databases had been using it for years and even some library systems utilized linking (or navigation) before the advent of the Web. What happened with the Web is that it became generalized and formalized and the method of doing it was spread across many places very quickly. It became a standard. And it arrived at just the time when user interfaces were moving from text based to graphical, so it had a natural mechanism in the hyperlink and clicking on it to see more.

The potential of this widespread linking method is enormous. It allows collections to self refer. It allows a collection to refer to background context. It allows a collection to refer to other collections. It allows a collection to refer to other works. It allows a collection to refer to external background and support context. It allows a collection to link to external services. It allows a digital library to be as big as the designer wants to make it.

This is tremendously alluring, especially to someone who has slaved for a long time to produce a very detailed specialist collection. The drive is there to utilize this technology to enhance the collection as much as possible and to expand it to make it a world-leading port of call for those interested in this type of material. This is a siren song with serious ramifications.

For each of the possibilities mentioned above there are advantages and disadvantages. It is your decision as to how much you want to use this technique.

©2003 Sun Microsystems, Inc.

Self-Referral

Self-referral means links from one object in the library to one or more others. There may be links for the name of a painting's artist to show all of his or her work. There may be links from description keywords to all objects of that type. There may be historical timeline links showing temporal or thematic development.

Advantages	**Disadvantages**
Multiple access points	May be confusing
Related objects	Reasons for link obscure
Object context	Linking in new objects
Easier to find objects	
A more friendly collection	
Basically no maintenance	

Background Context

This means links from an object to supporting material. These links could be biographies, information about the historical period, the appropriate school of thought for the work, definitions of terms, provenance or history of the object.

Advantages	**Disadvantages**
Really deep context	Creating the information
Starts alternative trains of thought	Maintaining the information
Access to peripheral data	Clear method of display

Other Collections

This means links from a collection or, more usually, a single object to other collections. Thus a collection of objects from the estate of an important person may include a painting. The link could be from that painting to the gallery where it is now housed, to the artist, the history of the company the owner worked for at the time the painting was acquired.

Advantages	**Disadvantages**
Adds information for little cost	Broken links
Provides access to other collections	Permission to use
Starts alternative lines of inquiry	Differing layout
Broad range	User confusion
	Quality of information

The user confusion may be because of the differing layout and style of content. Users may not realize they are in a different Web site. They may not know how to return.

You have no control over the linked to site, so it may change at any time. You also have no control over the quality, and any bias, of the information presented.

Other Works

This is a more detailed version of the above where a particular object in your collection links to a particular object in one or more other collections. Thus a copy of a Gutenberg bible may link to all the other copies.

All the advantages and disadvantages above apply with an additional advantage: Comparison possibilities.

External Context

Links to peripheral data held on external Web sites. The only difference between this and the links to other collections above is the type of material and quality of data provided.

External Services

These are links from your site to services that may either enhance the user's use of your site generally, or may allow services that extend the "boundaries" of your site.

A search engine, which allows the user to search the content of your site, is an example of the former. A meta search engine that allows the user to search external reference sources using terms or phrases from your site is an example of the latter.

Advantages	**Disadvantages**
Functionality at little cost	Services may disappear
Vastly more useful site	Services may have bias, or be unsuitable
	Services may start charging

Irrespective of the individual pluses and minuses listed above, one over-riding question must be answered.

How Big is Big Enough?

Every link placed on a Web site or in a collection, whether it is put there manually or by a program, must be there for a reason. Does it add value? Does it add confusion? Is it something most users will want to use? Is it there just because it was easy to do?

Every link once placed (and that is no small task), must be maintained. One of the biggest problems of the Web is broken links. This is worse than problems of traffic congestion (which just slows things down) or unavailable servers (which will be back again shortly), this is a permanent dead end. A broken link is a link which points to a page that has moved and left no forwarding address. It is not only annoying because the information is not there, it is also annoying because the user's hopes have been raised by the existence of the link. It is indicative of the Webmaster not caring about the site anymore. It is a lot of work to keep links current. It is so much work that there are programs which will check for broken links for you. What they cannot do is fix them. That is an ongoing maintenance task. And it could take all your time. (As an example: every one of the links in the Resources section of this book was checked for each edition. It was done through a database of this information, which generates both a Web site, and also the section of this book. It was the most time consuming part of creating the new edition!)

©2003 Sun Microsystems, Inc.

Services

Mention was made of external services in the section above which could be added to your Web site to make it more functional and attractive to visitors. This section lists some of those types of services. Some are of benefit to the user; some are for your benefit.

The most useful service to add to a Web site is a search service. This whole topic is discussed in “Search Engines” on page 35.

Various services exist which display news feeds and dynamic information to your site. It may be the case that weather or stock market information is useful and appropriate to your users. Many services offer these information delivery services. You can arrange the details with the provider and add a small piece of HTML code to your Web site at the appropriate place. Information can be supplied on:

- News—Make sure it is appropriate to your content
- Weather
- Stock market
- Background music—Could be inappropriate or annoying
- Travel—Could be linked to predefined searches
- Date and time

Advertising may not be an option (see “Income” on page 15 and “Measuring and Charging” on page 82), but if it is, then there are many services that will deliver adverts to your site and pay you a fee. This involves signing up with the service and then adding a small piece of HTML code to your pages.

Usage statistics are always useful and various services will provide statistics of visitors. Some of these are free. It is important to decide how much of this information you want, and what you are prepared to pay to get it. Simple visitor counters can be found on the Web and downloaded for free. More detailed analysis of the log files from your Web server could be expensive.

Forms for users to respond to you are useful and can be embedded in your Web site at an appropriate place. If you collect personal information from the user, make sure you have a privacy policy about how you will use this information, and that you put this information on the Web site for the user to see before they fill the form in.

Email capabilities for the user to send you comments are another possibility and the remarks about forms above apply here.

Your Web Site

With the capabilities of the Web in addition to all the content you have generated for your digital library it is easy to get carried away and build a site which has everything in it. This can be confusing and expensive to maintain, and may well detract from the value of your content. Resist the urge of “bigger is better.”

Resources and Content

Your digital library as a Web site consists mainly of your content, and it should be featured prominently “front and center.” You have the option of keeping things simple which will mean the content you have will be prominent, but the site may seem spare and unattractive. Or you can design a complex site which attracts many people. However they may be only peripherally interested in your content. This may be your intention.

Search Engines

Search engines provide two possible services to your digital library. They can allow the user to search the content of your site to find material, or they can allow the user to search on more external sites to add value to objects found on your site.

Internal Searching

Many of the commercial search engines (such as Alta Vista or Google) have versions which you can download to the server running your site, which will then allow users to search the content of your site. The search can be made to cover as much of the site as you designate (so that secure areas are not included), but the search will be limited to text searching. This is not a serious limitation as virtually all searching, even of digital libraries of images, is done against the text description accompanying the object.

The search engine will index every word on your site and present the user with a list of titles and page URLs for those which match the search criteria. It is very familiar to most users and an easy way to add powerful access to objects which might otherwise be difficult to find.

External Searching

To achieve this your site will connect to a search service running outside your site. This may be a major Web search engine (again, such as Alta Vista or Google), it may be a reference search engine (such as those from the Library of Congress or the British Library), it may be a service which allows you to broadcast a search to Web search engines (such as Metasearch or MuseSearch), or it may be one that allows you to direct your search to particular resources (such as MuseSearch or Webfeat).

Connecting directly to a major Web search engine is easy. Make the agreement with them (often this can be done on line and has no cost), add the appropriate HTML to your site and you are done. This allows a user to access a text box on the browser toolbar or somewhere coded into the Web page. The results of their search will be typed in (or dragged and dropped for Google) and will be shown just as if they were on the search engine Web site. It is easy and convenient for the user, but the user leaves your Web site as soon as they use the search, and you have no control over the look and feel of the results, or any adverts which might be on the search engine results page.

Connecting to an external reference resource is no more difficult. Make the agreement (also often available on the organization's Web site) and download the HTML code and place it on one or more pages. This provides essentially the same services as the Web search engines, except the results are limited to those from that resource. This may well be an advantage, as you will have chosen the resource because it is useful and has an assurance of information quality. Again your user will be on the remote service Web site and you will have no control over look and feel, though it is very unlikely that you will access a resource which has advertising. You can connect individually to as many of these resources as you wish from different pages of your Web site.

Connecting to a Web metasearch service generally is the same situation as connecting to a single search engine directly. All the conditions above will apply. The exception is where the service is a paid-for one (such as MuseSearch) where you then have control over look and feel and the user is provided with a direct return to your Web site. In fact if the results page design is consistent with that of your site, the user will probably never know s/he has left your Web site.

The services, which allow you to choose the resources to search and allow them to be searched simultaneously, are all paid-for services. They allow for customized look and feel and for the choice of resources, so that the user can be provided with exactly the research capabilities you

©2003 Sun Microsystems, Inc.

feel are appropriate and in a manner which integrates with your Web site. If you have a local search engine then these services can even search your Web site at the same time as the designated external resources.

Searching is pivotal to use of a digital library and it is worth spending some time in consideration of what you need for your library.

Portals

Portals are popular in 2002. A portal is an overgrown Web site. It provides either a lot of functionality of its own, or it provides a single point of access to an aggregation of information (other Web sites usually), which are likely to be of use to a particular user. Thus you will find portals dedicated to everything you never wanted to know about cats or dogs or the fear of flying. One of the interesting things about the structure of the Web is that you can build endless interlinked hierarchies of pages or sites or portals. Thus for every Web site that brings together a number of pages, there is a portal which brings together a number of Web sites. And there will be other portals that bring together other portals, and so on endlessly. There is no data model for this and no attempt to restrict or structure this behavior, so it is quite possible to have a looped hierarchy, and the concept (in linking terms) of a Web site being its own grandfather is easily achievable.

Leaving aside all the semantic niceties of the previous paragraph, portals are an important fact of life for digital libraries.

A digital library may turn itself into a portal. It can do this by scouring the Web and other resources to obtain as much information about its topic or genre as possible, then bring them together within its Web site, which then becomes a portal. Add processing and search functions, some fun things and some relevant and easy to access information which is kept up to date, and you have the ingredients for a portal. The idea of a portal is that it is the first place people will think to go to look for a particular type of information. Therefore, an important distinction for portals is whether they are broad or deep.

The Web search engines are the broadest possible portals. You can go to one of them to find out about anything on the Web. That is the theory. In practice none of them can index the whole Web in any reasonable time. Fortunately they do crawl or spider (the terms for the action of reading and indexing Web pages) different parts of the Web, so they provide nearly perfect coverage if you can search them all. Enter the meta-search engines that search the search engines and thus cover a wider area of the Web than any single search engine. Then there are the meta-meta-search engines that search...; you get the idea. There is no broadest and you should not aim for complete coverage. It is an impossible task.

There are many deep sites dedicated to one, or a small number, of the aspects of a particular specialty. This is where most digital libraries reside. They are centers of excellence or deep pools of knowledge reflecting the expertise and care and attention their creator has lavished on them.

Content

However you structure your Web site, it is your content which is king. You must not allow the bells and whistles of building a Web site to hide the content and make access difficult or obscure.

While building (or designing) a Web site it may occur to you that great benefit can be gained from adding material to your collection. As has been mentioned before, this could lead to viewer confusion or a very increased maintenance load, but it could also multiply the attraction and

usefulness of your site many times. Seriously consider the possible benefits before you add more content. Try to think as a user (both novice and expert) and determine if you would like the new content.

If the content is unique then the argument for including it is more compelling and all the practical constraints are the ones that will limit what you add. If the content is being added merely "for completeness" then consider linking to other sites that already have this content. The joy of the Web is that it is a co-operative effort and you don't have to do everything yourself.

Working with Others

Three ways of working with other Web sites present themselves. You can operate your site within a suitable portal. You can link to other interesting sites. You can add the sites of others to yours, and then you become the portal. Of course you can ignore these and go it alone.

The reasons for working with other sites are:

- Sharing of workload
- More comprehensive content
- Better exposure to your market
- Sharing experience and expertise

The first and last reasons are of benefit to you, if they can be realized. In the best situation they allow you to be part of a much bigger, better Web presence, without taking every waking moment to achieve it.

The second reason is for the benefit of your viewers

The third is of benefit to everyone, the viewers who will be more likely to find your site and benefit from it; to your collaborators as they will get traffic from viewers you brought in; and to you as you will have viewers "wander over" from the other sites and make your more well known.

The commercial and legal means of working with other sites range through:

- They pay you to link to them
- An informal co-operative agreement
- Linking with no agreement
- Linking through a "boiler plate" agreement
- Paying to be linked to

How you go about it is up to you. Just be sure the sites you work with have the same values and audience you do and will not upset your audience.

Accessing your Digital Library

Getting people to your digital library is possibly the most frustrating and heartbreaking part of the whole process. If you have a captive audience within your organization then you are part of an extremely small and lucky minority. Most digital libraries are just out there on the Web and have to take their chances with the rest of the world.

There are things you can do to advertise your presence and to attract the right sort of people, and even keep the crowds away if that becomes a problem.

©2003 Sun Microsystems, Inc.

Direct Access

Speak to other Web site owners and get them to "mention" you in various ways.

Direct Links

The other site adds a textual hyperlink from your site to theirs. This is usually just the name of your site within the middle of a description of it, or a list of similar sites. Instead of a text link you may have an image (a logo or recognizable image—remember copyright on this small image), which is the link. It will need to be very obvious so people will know where they are going without an explanation. Of course both text and image can be combined.

The owner of the other site will have final say over where and how your link is displayed, so speak to him or her before agreeing to this.

If you have a well-regarded site then the owners of other sites may approach you about making the link.

Citations and References

These may be entries in an electronic catalog, such as the librarian's yellow pages, or a reference in one of the "selected lists" which various academics and consultants and publishing organizations put on their Web sites to make them attractive.

Many of these you will have to find and then approach the owner of the "other site" yourself about having your site details added. It will be easier if you can present your site information in a form that matches that of the list or catalog. In most cases this will not cost you anything. Many of the consultants run e-zines or newsletters, so you could get a write up in these if they think what you are doing is interesting.

Then there are good old-fashioned paper publications. Getting an author to mention your site in his or her book or article is a good way to bring specialist visitors. This book is a good example. In the latter sections there is nothing but references to Web sites. These may be products, or organizations or interesting Web site examples. In this case I have used the tools of my company (and my years of experience) to find the information I wanted on the Web and then bring it to this book. If you want to tell me about your site for the next edition (or a possible Web site with resource links), please do, but don't expect this alone to make you fame and fortune.

Access Through Aggregators

If going it alone or making one by one contact with other sites seems like too much hard work (and it is hard work), then you might like to join a crowd.

You can have your site hosted on the computer of an aggregator within your specialty. They will then promote you as one of the "tenants" of their site. This is a bit like opening a shop in a shopping mall. You can expect the mall operators to do some of the administration for you, and you can expect them to do some advertising as well. Choose the hosting organization carefully, not only for what they will charge you, but also for the technical terms (such as the allowed number of visits or amount of data that can be downloaded without incurring extra charges), and the general tenor of the other tenants.

You can become part of a portal. This will probably not involve hosting your site, but your site will lose some of its identity as one of the competing or complementary set of sites within the portal. A portal that has only links to sites with pictures of birds will get a lot of traffic from bird lovers, but your site will be only one of many they could go to. On the other hand you could be the

only bird pictures site on a “nature” portal or a “paintings” portal. You will get more varied traffic, many of whom will pass you by completely after a first quick look. But you will be “the” bird picture place within that portal.

You could provide your content to a publisher. You lose all identity, but your content is made available through their site search engines, and is promoted along with all their other material. And you have virtually nothing to do. If you want your material to have exposure above all else, then this is an attractive option.

How Do they Find You?

If you are working as a stand-alone site, then you need to let people know you are out there waiting for them.

Search Engine Listings

The best way is to get yourself listed on the major search engines and those that are specialized in your area. To do this you can just wait. They will get around to you eventually. It takes the big engines up to two months to get to you the first time. Thereafter it may become more often if you prove a popular site (in their terms) so they will see what you have more frequently.

You can get software that “guarantees” to get your site in the top 10 of the search engines. Don’t believe it. Specially don’t believe it if you have to pay for the software. At the best these pieces of software will send notifications on a regular basis to a list of search engines on your behalf. At worst they will do nothing or make sure you site is listed under “XXX” and “Sex” and other “popular” terms. This is firstly misleading, and secondly just adds you to the list of hundreds of thousands of other sites indexed that way. And for every trick to “beat the system” the search engine owners often develop a counter measure before they release the system. Play it straight.

Make sure your site is accurately and completely listed. The text on any one page of your site might not tell the whole truth about you. To overcome this use the <METADATA> tag in those pages you want to be specially noticed. In this tag you list all the keywords you want your site to be indexed by. Use dictionaries and thesauri to find terms. Use specialist classifications. Be creative. Be technical. The more you use, the better your chances of being found. The more specific each term is, the more chance you have of being one of a few sites found. There is no realistic limit to the number of terms you can use, but once you find yourself becoming less specific it is probably time to decide enough is enough.

Most of the big search engines have a page on their Web site where you can tell them that you are new in town and would like them to come and look you over. These pages may be hard to find (you may have to contact customer support), but they will get you listed sooner than just waiting.

Make sure you contact the search engines you use yourself and let them know about your site.

Reviews

Contact the journals and magazines and annual reviews in your field and let them know about you. This may mean a press release. It may mean a phone call. It may mean contact at a trade show. The trade show or annual convention is possibly the best place to try this. Go armed with a one-page description of your site, what it has on it, who it is for, why it is great. Put in a few quotes (from yourself if necessary), and make sure your Web site and contact details are on it.

©2003 Sun Microsystems, Inc.

Talk to everyone you can find and give them a copy; get their email and send them another one afterwards. Try and get "interviewed" by reporters and consultants. This gets editorial coverage. If all else fails...

Advertise

Especially if you have created the digital library and its site as part of a business (either actual or prospective) consider whether the cost of advertising is worth the exposure.

This may be Web site advertising on a professional association site, or adverts in journals, or it may take the form of attending the exhibition at an annual convention. This has the advantage that you can pursue the "reviews" line of attack at the same time as you bring people to your booth.

Closing the Gates—Security Issues

Sometimes there is too much of a good thing and you may need to limit access to your site and content.

Controlling Access

Just keeping people out is not the problem for most Web sites, but there may be parts of your site where you'd like to know who is wandering around.

ID and Password Access

Think of having part of your site (or all of it) hidden behind a logon page. Typically this would ask the viewer to enter an ID (or Name) and a Password. Users can get these by registering at another part of the site or with you in some off-line fashion.

Only people with the correct password can get to the protected parts of the site and thus you know when they have logged on. However, you must be careful about privacy—set a policy and publish it. Remember, this is personally identifiable information if the ID allows you to find out the user's name.

Turn Off Robots

You may wish to keep your site "hidden." The main ways people find sites are through somebody telling them the URL, or they read it in a magazine, etc., or they find it on a search engine.

To a large extent, you can prevent people telling others about your site by not telling them. But what can you do about the search engines?

Tell them not to index you. This is as simple as putting a <ROBOTS> metatag in each page (or at the head of a tree of pages) where you don't want indexes. You may want to route all visitors to the home page of your site. In which case put an off tag in all the other pages. If you want to keep viewers away altogether put the tag in the home page of your site and tell it to allow indexing of none of the site. This is simple HTML and your site builder can put it there for you. Search engine spiders do not have to obey the command, they could still index your site, but all the major ones will obey what you ask for—they have enough sites wishing to be indexed without forcing you to be.

This will not stop people falling over your site by accident, especially if you have links from other sites, but it will make it very difficult for people to find you by accident.

Subscriptions

One good way to keep people away is to make them pay for access. The mechanism still involves names and passwords, but now you charge them for access. To do this make sure you feel people will actually pay—you might just drive everybody away. Managing payments and subscriptions is beyond the scope of this book, but there are plenty in the book shops and libraries which will tell you how to do it. You could even look on the Web.

Protecting Your Assets

The major assets of a digital library are the actual content and the manner in which it is organized. Its organization is pretty safe, but the content can be at risk through a Web site. To look at the general things you should be aware of, see the section on Intellectual Property Rights beginning on page 19.

The particular problem of Web site distribution is that the electronic form of the object has to be downloaded onto the user's machine before it can be seen/heard. There are basically two ways to protect your content—encrypt it, or don't download it all.

Encryption

There are many schemes for this. Most of the work in this area has been done recently to deal with digital music. Various encrypted formats are still vying for supremacy in this area. So read about them and decide whether this is the way you want to go. For textual material (including embedded images) there are a number of schemes, but by far the easiest and most widely accepted is to use Adobe's Acrobat format. For audio and video material Real Networks offers versions of its formats which are encrypted.

The prime difficulty of encryption is that the user has to load a decryption program onto their computer to be able to read the object. Acrobat is a model in this regard as it is a plug-in to a browser. It will download very easily and self install. It is free. It is in widespread use and will often already be installed. For Real video another plug-in is required, and so on. The good news is that most users see these plug-ins as a normal part of surfing the Web.

While many of the plug-in readers are free, the programs to produce the encrypted files are not. Here you have to make a choice of the competing priorities. Have a look in the Resources section for information about these tools.

Degradation/Thumbnails

One way of not sending the whole object is to send either a degraded version or a sample.

Thus a thumbnail image (which cannot be expanded to give a high resolution image) will give the viewer an impression of the picture, but not allow them to copy it and use it. An alternative is to provide a high-resolution image of a portion of the whole. Make sure it is not the important bit! Thumbnails, or image portions can be pre-generated and stored, or can be generated at the time the request is made. A trade off of storage space for response speed.

A few seconds from an audio recording or video serves the same purpose, with the same protection. The quality can be degraded (by re-sampling the stored high quality version at a lower quality) as well as only using a sample. Make sure any degradation is sufficient that it cannot be used, but is good enough to show what it is about.

Other tricks can be used to make a copy of the object that shows it, but makes it unusable. Add a color wash to an image. Use a graphics program to change one of the components of an image. Add a hiss to a sound recording. Add images and or sound to a video or change the color

©2003 Sun Microsystems, Inc.

balance. You can be creative about what will work for your objects. Just remember that any of the above tricks, which are put in by computer, can be taken out by computer with enough time spent on the task.

Catalogs

One, ultimately secure, way to deal with content is to not show it at all. The digital library consists only of its catalog. The objects are available under restricted conditions only.

For this to work, the user must be able to decide from the catalog entry only which object they are interested in. If you are asking them to pay before seeing, then the object description must be very good indeed. It will also be necessary to have some form of refund policy. Making the catalog more complete takes more time and effort, so again there is a trade-off.

The Wrong Users

Having built your collection it is a very sensible exercise to think about those who should not be your audience. It may be that your digital library contains material that will be seen as offensive, insensitive or immoral by others. If your collection contains so-called "adult" material, or it contains material which is sensitive (in the military, political or moral sense) then you should think carefully about if and how to make it available.

Remember that a Web site is available throughout the world and cultural mores differ widely. Having said that, you should not indulge in self-censorship to the extent that it harms the integrity of your collection or reduces its value to appropriate users. Decide for yourself if freedom of speech is more important than possibly giving offense, or vice versa, and control access to all or part of your Web site accordingly.

All of the above assumes that the content of your digital library is legal and honest and does not infringe on copyright or other intellectual property rights within the context of where you and your site are located. If there is the slightest doubt, then seek a legal opinion before you make anything public.

It is worth noting that in the USA the Digital Millennium Copyright Act will almost certainly apply to your digital library and the various laws, both federal and state, relating to the protection of children should be considered. Other countries and regions, particularly the European Union, have similar restrictions and a legal opinion is always the safest course, even after personally researching any possible problem areas.

Other People's Things

If you do make use of content from others, then make sure you acknowledge that you have done so, and that you have obtained copyright clearance for any objects that you use. As stressed above, the advice of a lawyer about provisions of fair use and scholarly comment and criticism is vital if you feel there is the slightest possibility of a problem. Even if you are absolutely convinced all is OK, a quick chat with a legal expert will not be amiss.

As well as possibly being a legal requirement for using someone else's material, an acknowledgment and a link to the original is of use to your viewers and will enhance your collection.

Of course, you may have to pay to use some material. So you have to decide if it worth it, or if a different image (non-copyrighted) will do just as well. And there is always the possibility that people will respect and admire what you have done so much, that they will approach you to use parts of your collection. This may lead to you collecting a royalty or license fee from them. But, as always, make sure it is yours to keep in the first place.

Part 2

Planning and Implementing

...Or "Let's Get On With It!"

Part 1 of The Digital Library Tool Kit contains questions to ask yourself before you decide on creating a digital library. Review them if you haven't already and make sure you both know what you're getting into and have a real need or desire to do so.

If you've got this far, you have decided that a digital library is in your future. This "Tool Kit" will help you get through the traumas of undertaking the planning, design, installation and initial running of your digital library. Once you get started then you are very much at the forefront of this form of library technology and you will be rather on your own. But don't despair—this document will help you through most of the things you have to achieve.

This section attempts to show you the way and provide help and guidance. However, the whole task is much too big to encompass in this many pages. Part 3 provides a large number of resources and references where you can find further information, help, and advice, as well as other projects struggling with just the problems you are facing.

Chapter 3

Selling the Concept to...

This chapter addresses the issues involved in getting a project accepted. Once the questions have been raised and discussed and "answers" found then the project has to have all its numbers added up and the "business" case has to be made. This is not a simple task. It is even more difficult if some of your management or staff and not wholeheartedly behind you.

Yourself

You have to be the biggest believer of all in the correctness of the project for your organization and of its practical "do-ability." You need to analyze the potential benefits and disadvantages involved in proceeding with the project. The benefits accrue (usually) if the project is completed successfully. The disadvantages happen (usually) if the project is started and is then abandoned or fails.

The benefits can be:

- Promotion
- Management of the new digital library
- Financial reward
- Job satisfaction
- New skills
- Peer praise

©2003 Sun Microsystems, Inc.

The disadvantages can be:

- Promotion prospects
- Frustration
- Peer envy (only if the project is successful)

You need to consider each of these in respect to your particular position and see that the benefits outweigh the disadvantages. If they do then the only thing you need to do is to commit to the project to achieve the benefits.

You are accepting responsibility for the project's feasibility. If you have reservations now is the time to lay them out very clearly. Do this firstly for yourself and if you can't see a solution then include them in your proposal to management.

Management

For the complete project you will have been asked to do some (or all) of the following things:

- Devise and plan the project
  - Define the scope of the project
  - Evaluate the feasibility of the project
  - Determine the benefits
  - Determine the costs
  - Determine a timescale
- Implement the project
- Manage the resultant service

This list covers all the stages that have to be gone through. You may not be responsible for all of them. For instance the decision to undertake the project may already have been taken; you "just" have to implement it.

You will have to submit to management, with solid arguments and facts and figures to back it, a proposal that answers the following questions:

- Can it be done?
- If not how much can?
- What will it cost?
- What are the benefits to the organization?
- How likely is it to succeed?

For most libraries the feasibility relies on the existence of tools and some expertise in their use. The tools (software, hardware, etc.) will be bought in and existing or new staff will be trained in their use.

The total cost comprises the costs of the tools and training and staff time and any services required (analysis, design, digitization, project management, installation, etc.)

The benefits are less easy to find and quantify. The questions in Chapter 1 cover most of the issues and possibilities. However, it is the case that, unless the library is planning to sell the digitized information, the benefits are mainly intangible and cannot be priced.

The likelihood of success is a different question to the first listed above (Can it be done?). Success involves the organization's environment as well as the purely technical feasibility of whether it can be done. Items to be considered are the organization's commitment to:

- Make resources available as required
- Continue when problems arise
- Try a new technique
- Move to a new service or business area
- Undertake a new business model

The external environment raises such questions as:

- Is this a "race to market"—to be the first with a new product/service
- The effect of new technology being introduced during the project
- "Competitors" appearing
- The long term viability of suppliers
- External services being able to deliver (quality, price, time)

Whatever the environment if there are untried steps involved then the proposal must highlight them so that the organization's decision to proceed can be made in the light of the best information and most informed opinions available—yours.

Staff

Your intention must be to show that the project is not threatening, not a waste of time, and not an attempt to re-organize the organization by subtle means. The positive aspects of job advancement, new skills and new business areas should be emphasized. The benefits to individuals as well as to the organization must be publicized; such as easier access to information, more complete information, reduced number of steps to obtain information, access to other more comprehensive resources, etc.

Keep the project staff as a team, but not too isolated from the rest of the organization.

In all cases accurate and timely information (including bad news) is better in the long run than trying to hide things. Good news can be allowed to leak (as it surely will) as long as it is not perceived that you are keeping things to yourself at the expense of others.

As a source of information on project and staff management there are any number of books and training courses at all levels. Your organization may even have the materials you need in the library or may offer training courses. Build some time into the schedule for this for you and senior members.

Chapter 4

Planning the Project

Like any project much of the work has to be done before any obvious start date. Even the assignment of a start date needs planning! Background research and fundamental decisions are needed so that wheels can be put into motion.

Planning

Stages of Planning

Careful planning at an early stages can save much time and heartache later. The amount and type of planning necessary depends on the size and scope of the project and those you can't know until the planning is done. So, like many of these activities, the planning is an iterative process. A first rough approximation gives you a feel for the size and complexity and difficulty and hence for the areas where more detailed planning is needed.

Detailed planning is always needed in two circumstances. A piece of work may be complex in itself or intricately linked to other activities. It becomes a critical task in the logistics sense. A small change in it can have large effects outside. Secondly a piece of work may be of unknown scope. It may be that the size of the work (e.g., volume of material) is unknown, or it may be that the extent of the work (e.g., how can these two subsystems be made to work together) is unknown. Only by a detailed analysis of the problem and successively breaking it down to more quantifiable pieces can an overall "number" be obtained to relate back to the project as a whole.

©2003 Sun Microsystems, Inc.

The major function of the rough cut or first draft plan is to cover all parts of the project and to identify the amount of planning work needed for each part. Cast a broad net for the first draft within the limits given in the management project brief. Try to include anything which may just have a reason to be in the project. It is much easier to remove things from the plan later than it is to add them in once work has started.

For example, the system is to contain the full text of company information. The original management idea may well have been Annual Reports, Technical Reports, Brochures, Product Specifications and Press Releases. However, in the first stage you should go and talk to Human Resources about personnel records, Manufacturing about process manuals and production reports, Finance about credit information, and so on. Certainly consider all the material in the library and contact every department head for material they would like to make available to the public, customers, staff and management and also for material they use from external sources, both inside the organization and outside.

The original brief from management was essentially for a publicity system, however, looking widely enough can show that the same system can serve internal and external users with both publicity and process and confidential information. Put it in the report that you studied it even if it was rejected, and say why.

Trade-Offs

As in any planning process, or even during any project, it is necessary to compromise some desires and to modify others. The usual reason is lack of resources. Resources may manifest themselves as staff or equipment or software or marketing effort, but they all can be reduced to cash terms at the end of the day.

The major planning question is whether, once it becomes clear all the goals cannot be achieved, to go for a completed less grandiose project, or to produce a partial version of the same grand plan.

If this is a one shot project and new funds are unlikely then a reduced scale project is the more sensible option—it will allow you to produce a “complete” service or product which will serve the requirements of a complete audience. Some people will be left unsatisfied, but you will have a chance at performing useful services for part of the original audience. Below is a list of trade-off areas. If you have to trade some of your audience (either directly by not providing access or indirectly by not supplying the material or services they want) then you have to consider carefully the organizational influence of the various possible “left out” segments as one of the factors in deciding where to reduce the scope of the project.

If further funding can be expected, either from subsequent budgets or from generated revenue, then it may be sensible to extend the development/implementation phase of the project so that it achieves its original goals, but over a longer timescale. The initial services/products to be offered must be carefully chosen taking into account both their usability to the intended audience and the difficulty and resources and timescale for future expansion. After all, it is no use providing access to all the library’s collections if none of the material can be delivered. Users will be no better off generally and will feel frustrated and that will translate into negative reports and the project may never get to the second stage.

Trade-off areas usually come in pairs as listed below. The odd man out is “cash” as it can be paired with any other. However, it usually is seen as the salvation for a late project. While this sometimes is the case it is more often true that the money does not bring the project back up to

schedule. The (very) simple reason for this is that it is only relatively simple tasks that can be run in parallel, and then only in certain circumstances. The process of creating a Digital Library is generally not suited very well to parallel actions. The processes are generally too complex or need too high a level of expertise.

- General—Cash vs. time
- Functions—Drop whole modules vs. reduce level of service
- Content—Leave out collections vs. reduce level of description
- Markets—Drop sections vs. restricted membership
- Quality—Cheap and nasty vs. exquisite
- Products—Simple vs. frills

Resource Limits

Some limits are already set, e.g., the organization (or library) has certain collections which are to become the content of the Digital Library. Others are set by you as the planner. e.g., the timetable. Others are set by external forces. e.g., the project budget or the availability of certain software.

These limits have to be factored into the overall plan. The two biggest constraining resources are equipment and expert staff time.

Lack of equipment can have an effect from delaying initial schedules to calling into question the actual technical feasibility of the project. While it is unlikely that you will wish to undertake anything which is technically more demanding than currently available hardware (servers and workstations, disk drives, networking, etc.) is capable of, it is possible. This is a new field and the limits are constantly being pushed.

An example of where you could go beyond what is technically possible, or what you can afford to do, is in the area of playing video to users across the Internet. Suppose you have a collection of videotaped conversations with famous people. It would generally not be possible at the moment to play these videos to users without expensive server and client equipment. The "normal" servers for a digital library (such as Personal Computers) are not capable of playing multiple video streams to multiple users and still providing the search and housekeeping services to other users and staff. This functionality requires either a larger general purpose server or a special Video-On-Demand (VOD) server. These require careful sizing to match demand to capacity.

In the broader sense "equipment" can extend to other pieces of hardware (such as scanners, video capture cards, audio equipment, etc.) and to software (such as library automation, video editing, Web server software, etc.)

An example could be a collection of documents in multiple languages where the search capabilities of the online catalogue do not allow for mixed language, or worse—character set, input. This would mean that free text access would be as if all document were in English with all the problems of homonyms, stopped words, wild character and fuzzy and "concept" search algorithms applied incorrectly. Even simple restriction of a search to a given language's terms may not be possible.

Expert staff time may not be a resource which you lack. It may be lacking for one of your suppliers. The lack may be in the number of staff, or their level of training and expertise. It may also be an unforeseen lack if people get sick or leave the project for whatever reason. Most projects of this nature are small and cannot afford to have many, if any, "spare" staff to provide strength in depth. Some capacity can be built in by training more than one person for a task even though only one person will be working on it. The "backup" person must be kept up to date on progress and must get some experience so they are not novices if they have to take over.

©2003 Sun Microsystems, Inc.

The biggest problem is lack of experience. This requires that staff learn on the job and mistakes will be made. Thus work has to be re-done or corrected and that means running late (and usually over budget). Staff will not be experienced because they will probably never have done these tasks before and theoretical training does not substitute for experience. Planning has to take account of this by allowing time for training and for the learning process afterwards. Alternatively some experienced staff could be hired for the project. This would reduce the need for re-doing work. It is important that permanent staff gain experience by working with these experienced contractors. This will detract from the contractor's productivity and this must be allowed for.

Market

The market (or audience or community or users) for the library's services may have already been determined. An existing library within an organization will be expected to continue to serve those users.

Even in this apparently static situation there is the possibility of increasing the audience for the library. The advent of new material or new ways of accessing or manipulating the existing material is a great excuse to market the library to the whole organization. New service possibilities such as campus bulletin boards or corporate Web sites are a way to promote the library within the organization.

In a new library the possibilities are even more open. For a library which has decided to market its services and content to the outside world, advertising those services becomes a business necessity.

This is an area where new (or revitalizing) opportunities can open up for the library. However, it is an area where caution and prudence must prevail if the new digital services are to be justified on the grounds of increased use of the library. Realism must reign when predicting new users and (particularly) new revenue.

The hi-tech glamour of the end result of creating a digital library can easily hide the costly hard work necessary to create it. It can also blind proponents to the "real" need for the new digital information. Advertising can alert users to the existence of the library and its new facilities and information. It probably will not convert many possible users who have already found alternatives to the library. Unless they have been consulted in the design process and their reasons for shunning the existing library have been overcome, they will stay with their old sources. These reasons may be as simple as not having desktop access and having to walk to the library. They may be as complex (to provide) as wishing to have the material from the library directly incorporated into the user's production process (as in a picture morgue within a newspaper). The issues and their resolution must be enumerated and assessed before the predicted user population and usage is given.

Products

The various products the library could offer must be compared to the requirements of the users (both captive organizational ones and casual public ones) as discovered during the user querying mentioned above.

It is possible to query all or a representative sample of an organization's workforce to find their needs and desires. It is not possible to do the same for public users. For them certain models of user requirements and behavior must be assumed and the library products aimed at those models. The distribution and numbers of those model users must be estimated.

The number of products must be tailored to the needs and wants of the users within the constraints of the available resources.

For example a TV station news library may have video stock footage for use by the news teams. They could digitize the complete library and make the clips available across the network so the production staff can digitally splice them into the current production. This sounds like an ideal situation. However, the quality of information required for broadcast transmission means that the 15 MB for a 1 minute clip discussed later is too low by a factor of about 32. Thus that one minute clip will require nearly 500 MB. This increase in size may make storage impractical. The simple 15 MB clip will enable the production staff to verify the correctness of the clip, but direct delivery is not a product that can be offered. As an aside to this example, it is imperative to verify that the editing systems can accept the video in the format it is stored in the library and that the two systems can co-operate in the delivery; otherwise the whole digitization process may be an expensive waste of time.

The simplest products are just for desktop accessibility and that may be the 80% solution for most users and anything beyond that is just not worth the effort even if the resources are available. It is then a case of deciding if the extras should be done for other reasons (preserving material, in expectation of future upgrades elsewhere in the organization, etc.) or not at all.

Remembering that the whole exercise will take longer than expected, now is a good time to tailor expectations to reality and not get carried away with the glamour of it all.

Access

For a digital library to be useful (for any library for that matter) there must be a user community and a means for those users to reach the library. Since the essence of the digital library is that all the material is machine held and manipulated, the digital library does not have a physical presence in the same way that a conventional library does. Its users will be connected to it for research and delivery via computers.

The simplest, closest and most restrictive access is via a desktop station connected directly to the server on which the digital library is held. This limits use to essentially in-house and is a model which is simple to maintain, but extremely unlikely in this age of networked information.

The most closed access in practice is across a LAN (Local Area Network) serving the organization's campus. This provides remote access but within the geographical limits of the organization. From the library's point of view it matters little if the network is a LAN or a Metropolitan or Wide Area Network (WAN)—the technology and management are the same.

Remote public (i.e. from users not of the organization's closed user community) access can be achieved through a direct dial-up connection. Since this offers very few advantages over, and a much higher telephone cost than, an Internet connection, it appears to be a method not worth considering.

Public and private access can be easily provided across the Internet as long as the library makes its services available that way. Given the graphical nature of most digital library material and the popularity of the medium, the World Wide Web (the "Web") would seem the outstanding choice for any library at this time.

To make the services available the library must run a "Web server" (a piece of software) which interfaces with its Integrated Library System. Most ILS vendors have such an interface and many will provide a specially tailored Web server as one of their components.

The administrative mechanics of setting up a Web site can easily be handled by the library and its supplier(s). For Web site design and management it is advisable to either hire external professional services or to hire specialist staff if the size of the site warrants it.

©2003 Sun Microsystems, Inc.

A Web site can be as open or as closed as the operator requires. It maybe free to all or it may require sophisticated passwords to enter. Or parts of the site may be free and parts private. A completely private site is known as an Intranet when it is available only to members of the organization.

One of the advantages of an ILS is that they provide specialized clients for the performance of specialist library tasks (cataloguing, serials, check-in, etc.). Many of these functions still exist in the digital library and their efficient operation is still a requirement. By using Java applets and/or network computers the library can still have specialist applications running on staff terminals across the Internet to allow efficient processing.

Operating a digital library through a Web site across the Internet provides the complete range of user access options. It is the only method of access that will be considered in this paper.

Volumes

One of the most important planning exercises for any library is the estimation of required storage space. A digital library needs this planning just as much as a conventional one. Just because the data is stored on computer disks and does not take up much physical space doesn't mean it can be ignored. Disks do take up space (well, their cabinets do) and consume electricity and require back up procedures and equipment and cost money.

Because of the nature of the information the delivery mechanism (a network) becomes a planning consideration in a way that the main doors to a conventional library never did.

Storage

Since the essence of a library is its accessible store of material, it is reasonable to expect that the digital library will hold at least some of the material to which it provides access. This material will have to be stored on one or more server computers. The amount of storage required can be a large part of the cost of the computing infrastructure required.

As well as the storage for raw data (text, image, video, etc.), there is the overhead required to index the material so it may be retrieved in the desired fashion. This can add an overhead of anywhere from 50% to 600% for text material. For other material type it is likely to be much smaller as the indexing will generally be of a textual description of the object and this itself will be small compared to the size of the actual data.

Security of the data must be assured in both the immediate and longer term. Immediate security can be provided by a RAID (Redundant Array of Inexpensive Disks) array which spreads the data across a number of disks in a way that allows one (or more) to fail and the system to still function while the failed component is replaced. In the longer term security is ensured by the regular taking (and checking) of back ups of the data base.

Some approximate figures for the storage requirements for objects of the different types are discussed below. Bear in mind that any decisions about resolution, color depth, sampling rate, even character encoding (all discussed in "Capture" on page 67) can change the numbers below dramatically.

In all cases, remember that any databases (either for storage or transactions or indexes) will need extra space to grow and for temporary files for housekeeping, etc. Storage requirements will have to be calculated for a reasonable period of service taking collection growth and transaction and administration and user files into account.

Text

Text, in general, is stored at one byte per character. This will be doubled if raw Unicode encoding is used, but will be reduced to about 1.2x if the UTF-8 format for Unicode files is used. Note that with some languages and scripts (particularly Chinese) the UTF-8 scheme is actually worse than holding the raw Unicode. If your data is likely to be multi-lingual, be safe and assume 2 bytes per character.

Indexing overhead can be as low as 50% of the document for just structural components (author, title, etc.) and can be as high as 600% for full positional and stemming free text searching. It is reasonable to assume that the overhead is 100%.

Records for the actual text of a document are simple and incur little database structuring overhead. However, unless each page is held as a single record, the problem of the variable length of the documents may be a significant factor. Assume a minimum record size of the average length of the documents (in pages) plus one page. This allows for the record to be stored in page size increments most efficiently.

Bibliographic and other metadata records are structurally complex and the database and indexing overhead for them may be as large as 800%. However, they are small compared to the full text of the document averaging about 500 bytes. Assume a 200% structural and indexing overhead.

Compression can be used on the raw text files and will give about 50% on average, with the original being perfectly recreated from the compressed record.

Thus an example small collection of 100,000 articles averaging 5 pages, all in English, to be stored in full text and indexed for proximity and structural searching will take:

Characters/Page	2,000	
Characters/Article	10,000	5 pages/article
Characters/Collection	1,000,000,000	100,000 articles/collection
Raw Data Bytes	1,000 MB	1 byte/character
Database Structure Overhead	200 MB	1 page/article = 2 KB/article
Index Overhead	1,000 MB	100% of raw data
Bib Records Overhead	150 MB	500 bytes + 200%/article
Subtotal	2,350 MB	
Processing, RAID, etc.	780 MB	33%
Total	**3,000 MB**	**= 3 GB**

Although 100,000 documents is a reasonable sized digital collection, the amount of storage calculated above is well within that offered by most entry level personal computers. Thus actual disk storage space is unlikely to be a serious bottleneck unless the collection is very large or has some special characteristics (such as multiple languages).

Some of the DLI projects discussed in “Bibliographic” on page 94 and projects like JSTOR (Chapter 6 for reference) address these issues of large volumes of text and how to store, manipulate and deliver them.

©2003 Sun Microsystems, Inc.

Images

Images will generally be stored as simple database records or in independent external files. In either case these BLObs (Binary Large Objects) are stored quite efficiently. Images tend to come in standard sizes because of the editorial decisions made during capture and processing and the desire to display the images on a fixed resolution display device. Thus a number of pre-defined record sizes can be used within a database to store the image data and the overhead can be kept very low (<1 KB).

Indexing will be of the bibliographic or metadata records and will thus be the same as for the equivalent text data.

If any extensive narrative description accompanies the images then they should be treated as text objects and sized accordingly (minus the bibliographic/metadata component).

Indexing of visual features is usually done by recognizing the features and storing them as keywords so the overhead is much the same as a simple keyword index on a relatively small piece of text, since only a small number of features (about 5/6) are usually recognized per image. This is partially an editorial decision and after 5/6 features in an image the remaining ones tend to be of little significance. This would add an overhead of only about 200 bytes per image.

Resolution and color depth of the images are the biggest variables in the image record size. 640x480 pixels is a common (small) size for images. Even at this size each image needs 300 KB (at 256 colors—it becomes 3x bigger for true color—16.7 million colors). A thumbnail (64x64 and 16 colors) will take only 2 KB. A full screen size (1280x1024 is the recommended high definition resolution for most 17-inch monitors) with true color (16.7 million colors—as many as the human eye can discriminate) will take (1280x1024x3) 4 MB.

Compression can make a very large difference as a compressed image can be almost 10x smaller than the raw version. Two cautions need to be borne in mind about image compression. The achieved compression is dependent on the actual image and the compression is "lossy." Information is lost during compression and cannot be recovered. A safe figure for compression, if it can be used, is about 50%.

The same small collection size of 100,000 images captured at 640x480 in 256 colors and with a bibliographic record (no commentary or description) for each which will be indexed for structural searches will need:

Bytes/Image	300,000	
Raw Data Bytes	30,000 MB	100,000 images
Database Structure Overhead	100 MB	1 KB/image
Feature Index Overhead	20 MB	200B/image
Bib Records Overhead	150 MB	500 bytes + 200%/image
Subtotal	30,370 MB	
Processing, RAID, etc.	10,115 MB	33%
Total	**40,460 MB**	**= 40 GB**

Compression of the raw image data can reduce this to about half (20 GB).

Audio

Audio data will be stored as files or as BLObs and the database overhead for them can generally be kept down to about 10%.

Indexing and bibliographic/metadata considerations are as for images.

Direct feature indexing of audio data is almost non-existent. If the audio is converted to text and then indexed the resultant text files should be sized as such.

Data size is affected most by sampling rate, sampling size and number of channels. Thus the spoken word in mono (using 8-bit size and 11 KHz rate) will take about 1 KB/second, whereas stereo music (16-bit and 44 KHz) will take about 17 KB/sec when stored as .WAV files.

Compression can be applied to audio using much the same techniques as for images. The compression is also "lossy" and the high frequencies are usually lost. The signals may also be processed to suppress hiss or boost the bass as with any audio recording. This has no effect on the record size.

Audio data will tend to be kept in individual files and so there will be no database structural overhead.

The same small collection size of 100,000 audio recordings, half sound (8-bit, 11 KHz and mono) and half music (16-bit, 44 KHz and stereo) of 10 minutes and with a bibliographic record (no commentary or description) for each which will be indexed for structural searches will need:

Bytes/Sound Clip	600 KB	10 minutes @ 1 KB/sec
Bytes/Music Clip	10,200 KB	10 minutes @ 17 KB/sec
Raw Data Bytes	540,000 MB	50,000 sound and 50,000 music
Bib Records Overhead	150 MB	500 bytes + 200%/article
Subtotal	540,150 MB	
Processing, RAID, etc.	180,000 MB	33%
Total	**720,150 MB**	**= 720 GB**

Compression can reduce the raw data size by about half so the compressed size is about 360 GB.

Video

Video is really a sequence of images and an audio track. However, its peculiarity that the images are only slightly different from those before and after it mean that a special form of compression can be used. Without this the raw data sizes are enormous. A second of video at 30 frames per second (fps) of the 640x480 image size (256 colors) considered above requires 9 MB. Even with good compression this amounts to about 1 MB/sec. Thus storage for a 90 minute feature video at this size would require (90x60x1MB) 5.5 GB.

Lower sampling rates (15 fps) and smaller sizes (320x200) can be used in conjunction with modern graphics cards to give large images from reasonable amounts of storage.

Recent advances in image extraction have allowed videos to be indexed by selected images and these to be analyzed for features. This indexing would add somewhat less overhead than the same number of images as there will be much common content.

Videos would typically be stored in individual files rather than a database so there is no structural overhead.

©2003 Sun Microsystems, Inc.

As an example consider a small collection of 100,000 video clips of 1 minute each at 320x200 and 256 colors at 15 fps. These are to be feature indexed at 10 image frames per clip. They have spoken dialogue and a bibliographic record (with no commentary or description) is required for each. They will be indexed for structural searches. The whole collection will need:

Bytes/Video	15 MB	0.25 MB/sec assuming 4:1 compression
Raw Data Bytes	1,500,000 MB	100,000 articles/collection
Feature Index Overhead	200 MB	10 images/clip and 200B/image
Bib Records Overhead	150 MB	500 bytes + 200%/article
Subtotal	1,500,350 MB	
Processing, RAID, etc.	500,000 MB	33%
Total	**2,001,350 MB**	**= 2,000 GB = 2 TB (Terabytes)**

This is an already compressed figure. Better compressions techniques, or a decision to lower quality, could halve this figure to about 1 TB.

Bandwidth

For amounts of data of the sizes given above the bandwidth needed to deliver them becomes a serious consideration.

For the sample objects considered above in the examples the delivery requirements for one object are:

One Text Article	10 KB	
One Image	300 KB	
One Audio Clip	600 KB	1 KB/s
One Video Clip	1,500 KB	250 KB/s

These sizes have to be compared against the available capacities of the delivery channels:

Internal PC Disk Channel (DMA)	33,000 KB/s
48x CD-ROM	7,200 KB/s
Ethernet LAN (100 Mbps)	10,000 KB/s
DSL/Cable	256 KB/s up; 5,000 KB/s down
ISDN Connection	128 KB/s
56.6 Modem	5.7 KB/s
28.8 Modem	2.9 KB/s

It becomes clear that the actual delivery time for each of the objects is quite considerable when networks are used as part of the delivery channel. In addition public networks have other traffic and the system is unlikely to achieve the above theoretical capacity. Public network (Internet) connections should be assumed to provide only 50% of their rated capacity with any reliability.

The problem is worst when the "continuous" materials (Audio and video) are delivered. They are the biggest so the download time is the longest (say 7/8 minutes for a 28.8 modem for a video clip). They can be played without a complete download using streaming technology where sufficient is downloaded so that the download will finish at about the end of the playing. For the video above this would only mean that the initial wait was reduced to 5/6 minutes. However, streaming technology uses even more efficient compression and more of the facilities of the accelerator video cards to shorten this time and improve delivery speed and quality. For small videos the technology works now with broadband connections (DSL and cable modems) and will improve rapidly in the next couple of years. Be aware that some of these technologies (like cable distribution) share bandwidth among a group of users and so, as the number of users within the group increases, so you proportion of the total bandwidth goes down. In extreme case to not much better than a dial up modem.

Processing

The amount of processing for delivery is not really large for the text and images, however, the requirement to move video (or audio) data from disk storage to a network connection requires significant processing power. An amount of compression is being done on-the-fly, but mostly it is just the movement of large amounts of data.

For this special hardware (Video servers) has been developed. If delivering video is a significant part of the functioning of your digital library then one (or more) of these are a must as the load video delivery places on a normal database or application server (specially if a Web server is added) is such that all the search and housekeeping routines slow to an unacceptable rate. Optimized server hardware (such as Sun's Enterprise servers) helps ease the load, but the calculation of retrieval requirements and delivery requirements must be done on a case by case basis.

Where very high bandwidth channels (such as 1 Gbit/s = 1000 MB/s) are available, then the limiting factor in performance can easily be either the processor or even the hard disk. At 100 MB/s a modern high performance hard disk cannot keep pace with a modern high performance network connection. The fact that there are likely to be multiple users on the network evens things out. Unless they are all waiting for downloads from that same server.

Systems

Two forms of systems need to be considered; hardware and software. The software may provide the functionality to make the digital library work but the hardware provides the underlying resources and processing.

Hardware

There are two components to any modern distributed client/server system; the server and the clients. Since the clients are the machines that reside on the user's desks there is generally little the library can do to enforce minimum levels of resources or performance. The library's system can contain a recommended (or indeed required) minimum level of equipment (and software) for the user to correctly and efficiently interact with the digital library, but users will still operate with whatever they have and expect to get some sort of response. Today wireless devices are becoming more widely used as clients. These have a restricted set capabilities (slower connections, small screens—often black and white, minimal or no keyboards, etc.). However they are viable clients and must be considered in overall system design.

©2003 Sun Microsystems, Inc.

If the library does have some control of clients such as staff machines or those for use in-house or on campus then there are two scenarios. The first involves ordinary PCs (or Macs) as clients, the second uses network computers such as SunRays.

The advantage of the PCs is that they already exist in many situations, they can have a wide range of peripherals (hard and floppy disks, modems, etc.) and they are familiar. The disadvantages are that they are more expensive, the peripherals allow data to be removed or other programs to be run, they are of fixed performance, and they have to be individually upgraded.

The advantages of Network Computers is that they are simple, inexpensive, do not have the peripherals, can utilize the processing power of a server, and are automatically upgraded every time they are run. The disadvantages are that they need a local server, applications have to download, and they must make use of network peripherals for file copying, etc.

Sun Rays are an attempt to address the maintenance and performance and upgrading problems of individual PCs. What their model does provide is an easy way to centrally maintain and distribute applications. These may be full blown retrieval workbenches or processing tools, or they may be Java applets running within an Internet Browser session. They can be loaded when needed. This impacts the way software for these systems is written as the software becomes more componentized and open. Individual applications have to work with each other. This feeds back to the whole development process and provides the environment for more controlled software development.

The server(s) for the digital library are pieces of hardware where the library has control. The number and power of the servers needed must be addressed for each installation. What is now possible is to think of using the specialized servers for different tasks so that there is some spreading of the workload and some redundancy.

Servers are basically specialized into three classes: database servers with large high speed disks and very fast local communications, applications servers with fast processors; communications servers with fast communications peripherals. They are usually adaptations of the same basic range of machines with specialist equipment and larger capacities added. This means a good basic platform can be utilized for all three classes. If the basic platform server is scalable (such as the Sun Enterprise Server series) then each of the specialist ones will be and the library will be able to grow in the areas where it needs to.

The real point of the specialized servers is that they allow the library to buy capacity (whether it is storage, processing power, or networking) where they need it without having to over purchase. The flexibility is there to grow as well as the security of redundancy in case of failures.

Overall there a large number of pieces of hardware which are needed for a computer server site, but these are not special for a digital library, with the possible exception of video delivery servers if they are needed. These may involve not only normal computer network connections, but may deliver their video by TV cable or even satellite.

Since digital libraries do require large amounts of storage whatever their content, it is a good idea to pay particular attention to the storage solution. Particularly important is the future flexibility of the subsystem. In this respect something like Sun's Intelligent Storage Network™ (an example of a SAN—Storage Area Network) shows the future direction where the physical storage devices are intelligently controlled and made available to a number of application and database computers. This creates the data as an independent resource which can be accessed (with permission) from any system.

Software

Like the hardware there is a division between the clients and the server(s). At each of these there is the operating system and the application software to consider.

For the clients the operating system is likely to be either a flavor of Microsoft Windows (95, 98, NT, 2000 or XP) or a system based on a Java virtual machine, or, just possibly, Linux.

Windows has the virtue of being familiar and has a large base of software. The Java VM allows all the promise of the network computers to be fulfilled and constitutes probably the only viable alternative to Windows as a user platform, especially when running dedicated applets and general purpose Browsers.

For the clients the application software is likely to be supplied by the supplier of the whole automation package. Individual standalone applications may be required, but the majority of the end user interaction with the digital library will come either through specialized client software or through a generalized Web Browser interface.

For the server operating system there are really only three alternatives: Microsoft Windows NT server or Linux for small installations, or a flavor of UNIX for larger systems. UNIX in various forms is the operating system used by most Web servers and by nearly all large library automation systems. Of the UNIX alternatives, Sun's Solaris™ is by far the current favorite and continues to gain market share each year, although Linux has a very respectable showing in the Internet servers in general. This helped by being available very cheaply.

The major decision of all of this section is the basic application software and all its components. It may be a system supplied by a single vendor or it may be a system with components added onto an open architecture frame work. Of all the myriad pieces of software needed, the central one, and the one which needs to be chosen first and most carefully is the library automation system (or Integrated Library System).

A number of digital libraries are being constructed at present utilizing a mix of information retrieval, media management and Web server packages. If these are not tied together with a unifying underlying framework then the whole system will have problems with growth and expansion. These systems are also generally no more than catalogues and do not address any of the housekeeping issues of conventional automated libraries or their systems. These digital libraries are today where library automation was about 20 years ago. They have started to tackle the front end part without considering the implications of long-term maintenance and management.

Although an ILS should be the core of a digital library system, the problem is that the field is moving very fast and is being driven not by the library community, but by academic research. This means the traditional ILS vendors are struggling to keep up with what is required. Thus it is a difficult choice and depends on a vendor's future plans and flexibility as much as its present offerings.

The resources chapter (Chapter 6) lists all the types of software which are likely to be needed and gives a selection of suppliers for each. For the fundamental ILS software the list contains the dozen or so largest vendors in the world and is essentially complete. There are (and always will be) small companies which can offer possibly an even better solution, but their corporate factors (stability, growth, support, coverage, etc.) have to be especially carefully weighed.

Resources

The most important resource for the whole exercise is staff time and expertise. Although there is a lot of hi-tech involved in creating and running a digital library, most of it is hard work.

©2003 Sun Microsystems, Inc.

In particular, because access to the library is so much easier, people will "drop in" more often. If what they see is always behind the times (last month is the latest issue of a journal) and initial teething problems do not get corrected (images appear in "false colors" on some workstations) then they will quickly stop coming back.

This is a plea which belongs in the next section (time) as well; do not publicly set too aggressive a timetable. Allow time (another scarce resource) to produce things test them correct them and then do it all again.

Data will always be converted in less than perfect form. If it is keyboarded there will be typos. If it is converted by OCR then there will be mismatches. If images are re-sized during capture, then some will be cropped instead and heads (and feet) will be lost. All these things have to be checked for either on a 100% or sampling basis.

Even once the data has been converted there is still the biggest skilled job of all to do; the cataloguing and indexing. Here editorial decisions and cataloguing rules may have the be developed as well as applied. It is certain that the exceptions will not appear until near the end of a batch of material requiring at least another review of all the previous objects.

Even with sufficient staff and people (possible experts from outside) there are the mundane resource problems like the main catalogue database computer developing a fault and stopping every one. These resources for emergencies need to be considered and contingency plans (stand-by machine(s), access to a remote machine, a loaner, etc.) need to be made.

Time

Whatever other resources you have available there will never be enough time.

Problems will arise and will set the timetable back. All that can be emphasized here is to plan as thoroughly as possible and to be conservative. Some things to keep in reserve:

- Do not plan parallel activities
- Assume that even off-the-shelf computers will need installation
- Assume that packaged software will need to be installed
- Assume that complex software (ILS, DBMS, IRS, etc.) will need modification once installed
- Assume that the next software version will not fix all the current shortcomings
- Remember that people take vacations, get sick, don't work 24 hours/day
- Plan for NO weekend working (strictly 5 days/week)
- Remember public holidays (including overseas, if that's where suppliers are)
- Look at history and calculate an ACCURATE number for staff who will leave
- Assume ALL new (and replacement) staff will know NOTHING about what you have been doing
- Remember on-the-job training takes time from the trainer
- Assume nothing useful from the trainee until fully trained
- Plan at least one major delivery with a one month delay
- Remember you will be away for periods of time—allow for them

These will provide an horribly long time compared to what management are expecting, but defend it as best you can as early in the project as you can. Get the suppliers working to a faster plan. It is better for you to have things sit on-site unused for now than for them to be holding you up.

Reconsider the plan regularly even if all appears to be going well. Inform people of problems and successes.

Chapter 5

Getting Started

Having planned what to do, who to use to do it, how to do it, it becomes time to actually get started. This section takes you through some of the major aspects of what to do.

This is strongly biased towards a library which is digitizing a collection (or more than one) which it already owns and is planning to make it available to both its organization's staff and the public (in some form). It assumes that an ILS will be used as the basis for the library functions, though this is by no means necessary if all that is required is to provide an open catalogue with direct delivery through the network. It is further assumed that access to the library will be provided over the Internet (probably the Web).

Administration/Management

A digital library project has all the problems of automating a conventional library plus a few of its own. The extras center around the need to digitize the existing material as part of the project.

This additional activity means that another group of staff or a contractor (or both) have to be managed and the results of their work integrated with that of others. Co-ordination is the biggest problem in a project of this nature. There are a large number of tasks of differing duration and complexity to bring together.

On anything except the very smallest project it is almost essential to use at least a project management tool if not full blown formal methods. A project management tool which allows tasks to be timelined and to be assigned to resources and reports conflicts makes it possible to:

- Avoid the worst of disasters
- Re-plan when things do go wrong

©2003 Sun Microsystems, Inc.

- Give a coherent overview of the project
- Allow easy management reporting

The project should be broken down into successively smaller tasks until the resultant task requires either only one resource or involves only one activity. These can be estimated for work and duration and then the bigger groups of tasks can be created as work units for a team of staff. Each of these "group" activities should be given a team leader to oversee its progress.

Once the "atomic" and "group" tasks have been identified they should be entered into the project planner. The most important result of this planning is the resultant timescale and resource requirement, but the most important inputs and the tasks and their dependencies.

Along with any activity where something new is being done or being delivered a "test/repair test" triplet of tasks should be added. The same should be added at the next level up when the results of the new activity will have to be incorporated with something else.

Tasks should be entered in the logical groups used to define them in the first place. Dependencies, to start with, should always be set so that the first of a pair of tasks is completed before the second is started. This is in many cases unrealistic for the final result, but it makes checking the logic of the plan much easier in the early stages.

Once the plan has been outlined with all the tasks and their dependencies, check it with the team leaders who will be responsible for those group activities. Use the whole plan rather than just that bit for that team. This way the team leaders get an appreciation of the scope of the whole project and may notice things being done in other groups which overlap with theirs or which have been omitted in theirs. Finding things at this "paper" stage is a lot better than when things have started. Involve the team leaders in all the stages of planning either individually or as a group.

Given the diverse nature of the whole project it may well be that a major activity for all staff and groups is "raining." This must be arranged in advance of when it will be needed, of course, but there is a danger in arranging it too early. Staff may forget what they learned, or confuse it with other recently learnt skills, or even leave before the task comes round.

Since there will be new skills and they will be diverse it is important to ensure there is cover for these skills. Both for formal occasions (such as when annual leave is being taken), and also for when things get behind and more hands are needed. Try to complement training so that it builds on previous skills, bearing in mind that a highly skilled person becomes very vulnerable (specially if working in a support role in a public institution). Make sure that none of the teams is left out of the training.

When underway things will become very hectic as the library will probably have to keep functioning as well as implement the digital library project. Thus preparation and house cleaning before the "official" start can be very useful. Data should be checked and as many corrections as possible entered into authority files and the like. But remember that the current library operations have to continue, possibly forever, in parallel.

It is easier for staff to learn a new activity on a familiar piece of equipment than to learn both at once. If computers are going to be introduced for cataloguing, then introduce them well before the cataloguing starts so that people get used to how they operate. However, it is essential to give the staff something useful to do with the equipment otherwise it will not be used and the training benefit will be lost.

Utilize your planning tool to produce regular internal (project team) reports and discuss them at meetings. Involve vendors, contractors, consultants as well. It is good for the staff to get to know them and to see that they are not alone in this venture.

Regularly use the tools to predict for/with your team leaders how the project is tracking according to the schedule. Be honest and if the timeline has to change then do it earlier rather than later. Time lost at the beginning is *never* made up later, however, good the rationale for believing it will be.

The major activity groups are discussed below with some of their special requirements.

Purchasing

You will almost certainly need to purchase some things for the project. The only situation when you will not is if you have done it before and are "merely" enlarging an existing system. In which case you should know how to go about it.

Broadly the things you will have to have are:

- Hardware
- Networking
- Library System
- Multimedia
- Content

Some of these you will already own and others you may have no choice as to what is acquired (e.g., the organization has a policy and contract for networking services). In these cases you have to learn about the products you will be using and make sure they have the functions and capacity to serve your expected needs. If they appear not to then check again and then contact senior management about an alternative or enhancement.

The requirements of each of these components is discussed in other parts of this document, and lists of suppliers are in Chapter 6. Each situation is different and you will have to make up your own list of what is vital, merely important, and just nice to have (usually "mandatory, highly desirable, desirable").

Make a very full list of requirements before shopping around. Even if the process is not a formal bid or tender process requiring a Request For Proposal (RFP), it is sensible to list all you think you might want before finding what is on offer.

This section lists some of the major things to be aware of rather than attempting a complete shopping list.

New Technology

This applies to both hardware and software. New is good, newer is probably better, "tomorrow's technology" is possibly worse. This is not to say the technology is bad, just that very new things (especially software things) have a habit of not working properly for quite a while. A home truth in the software industry is to never buy a version x.00 of anything, wait until x.01 when the bugs have been corrected. It may be you have no choice, what you want may have only just become possible or available. in this case you have to go with the very new technology, just build in a lot of testing time, get to know the developers well and decide if you can trust them, find out how important your project is to them, try to get any development done jointly, ask for a development discount, get something in return if you agree to be a "beta" test site. Then make sure you do test

©2003 Sun Microsystems, Inc.

it thoroughly, because if you say it is OK without proper testing it is much more difficult to get it fixed later. Testing may well become your major activity; be prepared for it—and tell your management so they are part of the decision.

Too Good to Be True

If one vendor offers a "deal" which appears to be too good to be true, then suspect it. Any price which is less than those of the competitors' (for the same product) by more than about 25% of the average needs careful consideration.

Ask the vendor about the price and listen skeptically to the answers. Get them in writing.

If nothing appears to be wrong consider how the vendor can be in business at that much less than their competitors. How is it that they haven't captured 120% of the market—after all no one wants to pay more than they have to.

Consider two things. Companies often "buy" business by selling to prestigious sites at less than cost. Do you qualify as one of these? Are all the others padding their prices for discounts later.

The lesson here is a skeptical approach to prices. Remember always that the vendor is in business to make a profit. And they will only stay in business if they do make a profit. Their prices must reflect this.

Remember that you will be a partner with your chosen vendor for a long time to come. Consider the long term stability of the vendor and the product line as well as the initial "deal." Also consider the benefits of long term loyalty to a vendor where you will get preferential treatment and access which you would not if you are continually swapping from brand to brand.

Total Cost of Ownership (TCO)

One way prices get to differ is that things get forgotten (sometimes legitimately) or have not been asked for.

Thus make sure that you consider the costs for you purchases for a period of time, not just purchase. Usually three or five years are the time to use. Make sure all the following are included:

- Maintenance
- Support
- Training and re-training
- Help services
- New releases
- Bug fixes or upgrades
- Back up facilities or loan equipment

Not all of these will be included in the base cost, but get the per year or per incident costs and estimate the Total Cost of Ownership (TCO) and compare these.

It's Not Enough

Make sure the number of licenses are sufficient to cover the users you have to serve. Check if they are simultaneous or connected or "seats."

What underlying software is being used and what is the license situation for that.

Get a figure for increasing the size of the system (number of users—of different types if need be, size of database, anything else which affects the licenses and hence the price).

Are licenses in perpetuity or annual?

Apples and Pears

If you are in a competitive purchasing situation then you will get prices for the various vendors for a particular item. Although you gave each the same set of requirements, you will be offered a variety of products which will generally fit your requirements.

What you hope to get are a number of product which meet or exceed your requirements. This will not always be the case and you will have to be diligent in ensuring that things have not been "overlooked" in the response. All your questions must be answered. Beware of missing answers. Take a hard knowledgeable look at "it's the same as." But do not rule out an alternative until you have thoroughly investigated it. It might actually be better than what you asked for.

The Numbers Game

If you give numeric quantities to the vendors and they come up with sizes or capacities or numbers (e.g., disk size. number of processor, number of OPAC licenses, etc.) different from yours ask them for their assumptions and arithmetic. For example a smaller hard disk capacity for the database may not be wrong if the database software includes compression or is very space efficient.

Fair's Fair

If you find after talking to a number of vendors that you have a changed requirement then go back and give the changes to all of them. It is only fair to them and to you. They may otherwise bid on the wrong thing because you forgot to ask for something.

If you do go back to a vendor to ask for a better offer (in any way—it does not have to be price) then give the same chance to the others. After all, they are all in business to get your custom and you want the best product at the best price.

A Bigger Picture

Don't get blinded by the bottom line cost of the product. Particularly if it crucial to your project (like ILS software or database server hardware) look at the vendor as a partner you will have for a long while to come.

Do you like their approach, are their people knowledgeable, do they treat you properly, are they "legal, decent and honest," is the company well managed, do they have experience, are they committed to this industry, do they have a good (or at least not bad) reputation, do you think you can work with them?

Answer all of these questions and avoid all of these pitfalls and you should be well on the way to having the sort of equipment that you need at the right price and in a way that will let you sleep at nights.

Capture

Capturing the essential components of the original material in digital form is the heart of the process of setting up a digital library. It is creating the information for the library.

Based on each type of original material the important decisions are:

- What elements of the original to capture
- How to capture them
- At what level of fidelity
- How completely

©2003 Sun Microsystems, Inc.

Many of the original material types (e.g., books, videos, tapes, reports, etc.) are actually multimedia in terms of the simple categorization of the media types commonly in use. Thus a book has text and images, videos include audio and can have images extracted. Since must of the media types will have to be separately captured (only images and text have the same initial physical capture process—scanning or digitizing) it is necessary to consider the workflow for each of the material types in the collection.

Consider for example a normal printed book or report.

At a media level it has both text and images. Are both of these important? Is it worthwhile capturing both of them? Is it worth capturing them separately or would a simple image of the whole of a page suffice?

At the structural level the book has covers, a title page, a verso page, possibly frontispiece(s) and probably index and contents pages. Are all of these required as well as the body of the book?

Having decided what to capture it must be decided at what level and how completely. Perhaps the index contains meaningful words and phrases which would be useful for subject retrieval, whereas the contents page does not. Possibly the contents page is captured as an image only and the index is converted to text.

The problem arises that not all books are created equal and so either a blanket decision according to material type must be made or a complex series of procedures and rules must be set up. Of course a mixture of the two or broad categorization (such as “scan indexes only from text books”) is also possible.

The designs for the workflow depend on factors such as:

- How the material will be searched for (fulltext vs. indices and cataloging)
- How the material will be used (read and transcribed vs. cut and pasted)
- How precisely material must be targeted (book, chapter, page, paragraph)
- The functionality of the software (capturing text is no use without fulltext searching)
- The system capacities available (disk space, processing power, network capacity)
- The capture equipment available
- The staff available (or the services that can be bought)
- The time available

Having decided on the components of workflow for the various material type it is necessary to consider how each type of medium can be captured. This enables a physical workflow to be laid down so that components of material are not missed and can be related back together once completed.

Since all the efforts of your capture will end up as computer files it is worth spending a little time on thinking about files names. Although the contents of these files will be accessed via a Library system, most of the original images or audio will stay in their files. These files will be referenced from within the library system and will be retrieved and played at the appropriate time.

The file names must be unique. This can be achieved by two factors. Name the files by material type and give them some form of sequential number.

The file type will be automatically determined if you use name extensions as the extension denotes the type of file. This is a generally adopted convention, although both three and four character extensions are in common use. Since both Windows and UNIX will recognize a 3 character extension it is sensible to stick with that.

The body of you file name can be any length (which often means up to 32 characters), but it makes sense to adopt a simple classification and coding scheme to keep file names to a manageable size.

Many capture programs will take a set of prefix characters and pad them to 8 (the limit on file names under DOS) with numbers and automatically increment for each file created. This is advantageous as long as you remember that it will be necessary to manually relate the generated file name to the physical document (or film, etc.) at the time of capture and for this information to return with the captured files. Some allow this information to be entered into a computer record, for others it will have to be on a worksheet. This forms an essential link in the processing chain and must be part of the quality control process.

Remember also that 8 numbers represents only 100,000,000. This seems a lot unless you want to put your institution's initials at the front. With two initials the number of files is reduced to 1 million and with three to one hundred thousand. These are not large numbers so, if you have to use 8 character names for even part of the process (they can be expanded later if it seems worth it) choose prefixes with an eye on volumes.

Text

Text comes in two forms for the purpose of capture. As print on a page and as a machine readable computer file.

Text — Printed

This is first processed physically as if it were an image; the image is then processed to convert it to encoded machine readable text.

Scanning paper or other physical material for text extraction has some differences from scanning for images.

Most text is black on white and the Optical Character Recognition (OCR) programs need black and white. So the capture should be black and white. Most scanner software has a "text" setting which gives optimal black and white images for OCR conversion. In fact, the setting usually produces a "gray scale" image rather than a stark black and white one. This allows for smoother edges and more rounded corners which both makes OCR more accurate and also makes the text more readable.

The advantage here is faster scanning, better OCR performance, better display, and smaller files. A win-win situation all round.

Even color text should be reduced to black and white images for consistency of conversion. If it is desired to keep the page image for display to the user, then this will have to be re-scanned. For any volumes it will be faster to have scanners dedicated to black and white and to color and move the documents between them. Make sure the process assigns similar file names to the two files.

Since text tends to come in large volumes (many pages) it is very beneficial to set up the physical scanning so that it can be automated as much as possible. All flat bed scanners have Automatic Document Feeders (ADF) which will feed pages in exactly the same way as a photocopier (The whole process is very similar.) For the smaller scanners the feeders are limited to about 50 sheets so they must be continually fed.

Even this capability is of no use for bound material. The processing of bound material may be by direct scanning (there are special scanners for bound books which do not require the spine to be to be fully opened) or you may decide to photocopy all the material for subsequent scanning.

©2003 Sun Microsystems, Inc.

This ensures the copied pages are then all of the same size, orientation, paper weight and quality and are all black and white. This all makes the scanning process more reliable at the expense of the extra step and taking the care of the physical material at that stage. It is also possible to batch and control number the copied pages for later use. Scanning can then also be carried out in parallel and even be outsourced without concerns about the handling of the originals. Any repeat runs are against the copies and so the originals are further protected. This protection is probably only warranted for expensive, delicate, fragile material. However, the improvements in paper handling and general efficiency are worth considering under all circumstances.

When scanning multi-page documents it will be necessary to decide how to consider them. The individual pages of a book will be scanned one by one and each will be an individual image. As part of the scanning process these single page images can be combined into one file which holds all the images of the whole book or some logical part. This is where bibliographic theory comes face to face with practical reality.

If the pages of a book (or article or report…) are scanned as one object and stored in one file then they must be treated as one in the retrieval system or catalogue. Even if individual chapters of a book are catalogued and appear separately in the catalogue; when the link to the full text is made the "book" will be opened at page 1.

If the individual chapters of a book are scanned as separate objects then they must be catalogued as such and retrieval will start at the beginning of the appropriate chapter. However, there will generally be no obvious way to have the chapters linked together so the user may just "page forward" through the whole book. Even if the bibliographic record for the chapters are linked in the catalogue, most library software is not capable of making the leap from Chapter 2 to Chapter 3 of a title.

If each page is scanned and stored separately then display of the relevant page to the user will be immediate, but something special will have to be done to allow that same user to read the next page of the chapter, let alone the next chapter of the book.

At present it is necessary to make this user-oriented decision at data capture time. The next generation of library systems and document readers should start addressing this issue. The good news for the future is that however you decide to serve your users now, the technology exists to split or merge the image and text files to provide the desired level of granularity when the presentation systems allow that luxury.

Once the text has been scanned it needs to be run through the OCR program to convert it to a machine readable encoded form. This can be done as part of the scanning process if the computer configuration is big enough to allow the files to be sent around a network automatically. If not then the files must be processed after the scanning. In any event it is necessary to have some human intervention to review the conversion and perform editing where necessary.

The problem is that OCR conversion is not an exact science and the quoted conversion rates of "better than 95%" are just not useful. They sound quite good and they are for running text which will only be read, not indexed.

Consider an accuracy rate of 99.7%—this is very good for the software, but it means that three characters in 1000 are wrong. Since the average English word is five characters, it means every 67th word has a spelling mistake on average. These words will be indexed so the spelling mistakes will get into the index. This leads to confusion for the users. It can be a monumental task to correct.

Errors take two forms; a character just cannot be recognized or a character is incorrectly recognized. The former problem is addressed by the OCR software when working in interactive mode by displaying the image and the piece of text to an editor and asking for the correct character to be entered. The second case is not recognized as a problem by the conversion software. However, most of the software has an "editing" component which runs a "spell check against the text and this will catch many of the replacement characters. This still doesn't catch all the replacements (e.g., CAT becomes OAT because of a smudgy C). A further stage can be run with the more expensive software to attempt a context match (This should catch the OAT if the article is about Cats. Unless, of course, it is about cats in fields of cereal.) this is only of limited use in subject specific literature as the dictionary has to learn (or be taught) the terminology. For mixed languages its more of a waste of time than a help.

Knowledgeable editors are a must for this stage of the work. That is why specialist conversion services exist. (For a more extensive description of the problems visit the site of Access Innovation Inc. at http://www.accessinn.com.)

Once the OCR has been run then the text exists in machine readable form

Other than the problems of mis-conversion mentioned above you must be aware that obtaining OCR programs for character sets other than Latin (used in Western European languages) is a very difficult job and is an area where conversion is probably best left to a specialist bureau. Even if the conversion can be performed the problem of character encoding (see below) remains.

Text—Machine Readable

Machine readable text results either from the above scanning and conversion process or from originally created files from a word processor or some other computer program.

For a digital library two things have to happen to the text file. It has to be stored so that it may be displayed to users when they request it. It has to be processed and indexed so that its content is available for searching.

The problems to be overcome arise from basic character encoding and document formatting or structure.

Character encoding is the assignment of a computer code to each of the letters in the document. This is done during the word processing creation or during the conversion of a scanned image. Document formatting comes mostly from word processed documents where different pieces of a document (for instance the title) are given different typographic characteristics (the title is printed in a larger font and in bold). The codes to control this (and even the fonts to be used) have to be included within the document file, yet they have no impact on the content of the document and must be ignored for indexing.

Character Encoding

If all the text files come from a single source and are from originals in the same language then there is no reason to expect any problems with character encoding—all the characters will have the same encoding.

If, however, the documents come from different sources and particularly if they are in different languages, then they may use different encoding schemes and will appear as gibberish to all but the reader and the indexing program.

©2003 Sun Microsystems, Inc.

There is a unique universal encoding scheme. It is Unicode (or the international standard ISO 10646) and allows almost all of the characters from the world's languages to be encoded unambiguously. The problem is that at present Unicode is not widely implemented. It is rapidly gaining popularity and will become the future standard. Both Netscape and Microsoft Web browsers will recognize it and display it (if the correct fonts to give the character shapes are downloaded).

Multiple languages and character sets will have to be re-encoded into Unicode (or another unambiguous scheme—there are others, but they are not universal and are not recommended). This has to be done on a file by file basis and requires suitable software, mapping tables for the encodings and an operator who can look at the original text file and determine which character set it is in. Again, this is a time consuming and specialist task which is best undertaken by a specialist bureau.

Formatting

Having got all the text into a consistent encoding it is necessary to process the text so that only the content will be indexed.

Jumping ahead to consider the display of the document it is necessary to decide if the formatting of the displayed document will be as in the original or if the document will be displayed in a standard consistent for the whole collection, or if this is a user option.

Whatever the decision it will be necessary to identify the formatting conventions and structural elements of the text. These items can then be marked in a consistent way so that the indexing can ignore them and the display program can ignore them and apply its "house style" or else obey them. In some instances this is a non-issue because there is no formatting. In others it is quite easy, if time consuming, because the originating program (especially for word processors) is identified and the pre-processing can then use the control codes of that program to seek out the required content. In the third case there is no such algorithmic method available and it is a skilled manual process to "mark up" the text in a special editor which introduces the codes the indexing program (and the display program if used) needs. Again, a job often best left to specialists unless the volumes are low and the formatting simple.

Indexing

Once the text is all nicely cleaned up then it is a relatively straightforward process to feed the text files into a database. The database may store then for retrieval or it may just index them for searching. The indexing program needs some decisions to be made before starting such as:

- Which areas (or components) of the document are to be indexed
- How are they recognized
- Is a stop list to be used
- Is more than one stop list to be used
- Are phrases as well as words to be indexed (how will they be recognized)
- Are the extracted terms to be processed (against a go list or a spelling checker)
- Are term positions to be recorded, if so at what level (for proximity searching)
- Is morphological analysis to be performed (word stems, different vowel forms)
- Are browse lists to be created as well as direct access indexes
- If so how many and where to the elements come from
- How are diacritics to be handled
- Is language to be taken into account

This is not a complete list. The list of decisions depends on the capabilities of the library system or information retrieval system being used as well as the characteristics desired for the eventual digital library catalogue.

Fortunately, text catalogues have been around by far the longest and have developed the most sophisticated mechanisms. Thus the other material types will not have nearly as many problems to solve and decisions to make. Equally language is our most important communications medium and most users of a digital library will spend most of their time interacting with a system operating in a text mode. This is the most important underpinning of the whole operation.

Images

Before any capture starts (except experimenting) there are a number of decisions to be made. The assumption for images is that they will be viewed as images rather than processed. This is not always true, but will be assumed here. The viewed image may be viewed in a number of environments and it is necessary to relate the quality of the capture to the display environment.

First consider scale. At two extremes are a postage stamp and a sheet map. (There are bigger and smaller examples, but these will do.) The stamp is physically small, but may need to be viewed at high magnification. The map is physically large, but may only be needed at an overview resolution.

The problem is that the resolution of the image is fixed at capture time, not as with the physical object where a magnifying glass can give any required resolution (within the limits of the printing, etc. process). If the stamp is captured at a resolution of 300 dots per inch (dpi) then that is the maximum amount of detail you can see. Any attempt to magnify the image just produces bigger dots not more detail. Having said that, there are programs which can manipulate the captured optical image and enhance it to greater resolution. This is done by a process of mathematical interpolation and gives a series of higher resolution (smaller) dots between any two that were actually scanned. This process gives better color gradations and can give the appearance of sharper edges, however, it cannot “invent” features—it merely “enhances” the details already there.

For flatbed scanners resolutions go up to 1200 dpi except for very expensive ones where about double this can be achieved. Drum and other special scanners can achieve resolutions of over 10,000 dpi. Flatbed scanners tend also to have two resolutions one along the bed and the other across it of half the lengthways resolution. Even the 1200 dpi resolution is generally as much as is required.

The problem of resolution is that the higher it is for a given picture size the larger the file needed to contain it. If the file is only capturing black or white then one bit would be enough for each resolved point and an image of 1 inch square would contain only 1200x1200 bits =1,440,000 bits (= 180,000 Bytes, =180 Kilobytes, = 180 KB, = 0.18 MB)

Add to this the fact the stamp is actually in color. If the realistic minimum color depth of 256 colors was used then the size of the file jumps by a factor 8 (to hold the color information for each resolved point) to 1.44 MB. For a more reasonable high resolution image each pixel would use not 8, but 24 or even 30 bits, thus bringing the size for one stamp to about 5 MB.

Thus a collection of 1,000 stamps would take 5 GB of space just for the raw data. Even at today’s prices this is a lot of hardware and becomes costly for a reasonable sized collection if the more costly high performance disks are to be used

©2003 Sun Microsystems, Inc.

As with all capture, the matter of encoding and formats is important. A number of pieces of information have to be recorded about the image; either so that it may be viewed (technical) or so that it is adequately described (cataloguing), or so that the ownership rights may be properly protected and observed (IPR).

Some of the technical information required is:

- Its file format (not always clear from the file name)
- Its image type (bit-mapped, vector)
- Is it compressed and if so using what scheme
- Its dimensions
- Its color encoding (RGB, CMYK, etc.)
- Its color resolution (color depth, luminance, etc.)
- Its completeness (detail, complete, part)

This information is distinct from that which describes the image so that the viewer may form an opinion about e.g., the original colors—for that information about the spectrum of the scanner, any filters, any color processing, etc. would have to be recorded.

The amount of technical and associated information that has to be recorded depends on the expected use for the material. The more the objects will themselves be the object of study and scrutiny, the more information has to be retained.

Audio

Like all the other media, Audio can be presented to the library as an original analog object or in a digital form. The first stage is to digitize the analog and then process the digitized form.

Digitizing

Audio objects will almost certainly be recordings (tapes, CDs, wax cylinders). To capture them is as simple as playing them on the appropriate player and recording the result into a digital recorder. This may be a self contained physical process where a tape is played on a tape player connected directly to the sound card of the computer. Or it may involve actual re-creation of the sound waves and capturing them with a microphone attached to the computer sound card.

Whichever is necessary there are obvious requirements to preserve both the fidelity and purity of the recording (no coughing, only one input source running). The most obvious editorial decisions to make here are:

- Sampling rate (11K/sec for speech, 44K/sec for music, etc.)
- Mono or stereo or quad or any other special scheme
- Number of tracks (or channels)
- Noise processing (Dolby, etc.)
- Digital format (.WAV, .SND, .MID, MP3, MP4, etc.)

Most of these decisions depend on the desired use of the audio track. The sampling rate determines the highest frequencies captured. Mono obviously loses any spatial information, but not many computer systems can (or will be able in the near future will be able to) re-play in quad sound. If the sound is being recorded for use in a music studio then making individual tracks or channels available may be much more useful than just the combined final effect.

Noise processing (or any other form of signal processing or generation) may be beneficial or it may completely ruin the utility of the digital recording; it depends on the use that will be made of it. If in doubt, capture at the highest quality and do no processing.

The final format is purely a matter of the user capabilities. Formats can be converted, but it takes time and processing power and so it is sensible to store the digital sound in the form in which it will most often be used. It may be that the sound files should be stored in the most popular distribution format (such as MP3 if the material is mainly music) to prevent this conversion problem, as long as a high definition copy is stored as a master.

Processing

Currently there is very little that can be done to make a sound clip directly searchable. Text files have been searchable in sophisticated ways for many years and now images may be searched for form and visual content. However, it is not possible (in general) to whistle or hum a tune to a computer and be told what it is. Audio files are almost entirely a display only medium.

There is one just emerging exception to this. That is the spoken word. Voice recognition software has recently become capable of reasonably accurate recognition of direct continuous spoken input. It is intended for interaction with the computer, but is equally capable of recognizing, and converting to machine readable text, a digital recording. This allows the spoken text to be searched for meaningful content in the form of words. This is just text searching performed automatically, but it an important advance over transcribing spoken recordings.

It is not generally possible, within a library system to search directly for frequency signatures, intonation patterns, etc. However much of this technology does exist in specialist systems and even some retrieval systems. The capabilities of those systems could be mated to the general organization, storage, administration and delivery mechanism of a library system to meet a specialist need.

Video

Video is the other major medium considered as part of a digital library. Its processing combines the worst aspects of images and audio (since that is what it is composed of).

Capture

Like audio video capture consists of playing the object and feeding the output into a video capture card in the computer.

If the original is in the form of film then a special piece of equipment must be used which effectively “shows” the film to a video camera and the resultant signal is fed into the capture card.

For video tapes the process is physically similar except that a VCR (or VDR or laser disk player as appropriate) is used, the result is still fed into the capture card on the computer.

Here again editorial decisions have to be made such as:

- Black and white or color
- Color depth
- Frames per second
- Digital format (.AVI, .MOV, .MPG, .QT, etc.)

The arguments are much the same as those for audio and relate mostly to intended use of the material and the expected capabilities of the audience’s playing equipment.

One extra dimension here is the actual volume of data recorded when a video is digitized. To achieve the optical quality of a 35mm film takes such an enormous amount of storage space and bandwidth that this is, for now and the near future, an impossibility (although larger disks and satellite and cable delivery are rapidly changing this). Thus video compression is used and this

©2003 Sun Microsystems, Inc.

leads to another editorial decision. Which CODEC to use? A CODEC is a "compressor/decompressor" (or a "coder/decoder") and is a piece of software which compresses and then decompresses the video data.

The problem with video compression—like that for still images—is that it is "lossy." This means that the compression will lose some data so that the decompressed image is not as "good" (in some way—usually sharpness of detail) as the original. This may or may not be important depending on the final use of the material.

Except for special cases (such as a film studio video library where absolute fidelity to the original is required) most video is compressed and it is originally captured according to the dimensions of the place where it will be played—the computer monitor. This problem of storage and transmission bandwidth is also being alleviated by graphics cards in the workstations which can decompress and expand the images in real time from their small transmitted size to a full screen display.

To achieve real-time video displays across a network like the Internet is not really possible at the moment without high speed links and a lot of luck. Other traffic on the net can cause unexpected interruptions.

Processing

Processing of video for direct searching has made great strides recently (see the UCSB and CMU projects in Chapter 7) and techniques such as keyframing and subsequent image feature recognition mean that a video clip can be "summarized" and even searched directly.

The process is automatic once the relatively simple decisions about the type of summary and feature recognition have been made.

Index storage is not a space problem compared to the storage required by the original video.

Other

The above four material types constitute the totality of what most people consider to be the compass of the digital library. However, there are other material types which may have to be handled by a library and could probably benefit from the rigors of library discipline over their care, maintenance and administration.

These are merely listed with a couple of brief comments on each.

Computer Programs

These are in one form only; text files like any other and can be stored and indexed as such. The problem with this is that the text of a computer program is that it says very little about the function of the program. Simple keyword extraction would list the variables used, but not anything about what they are used for.

Interestingly, the programs have a second form when they are running in a computer. In this form they have visual characteristics (screen shots) or possibly audio (sound effects) which may give an indication about what they do and may be a more sensible way to categorize them. As far as the author is aware there are presently no digital library systems which deal with programs at this level even though there are definite advantages to be obtained.

Computer Presentations

Presentations which are produced on and shown through a computer are another material type having peculiar characteristics. They generate the images and sound (and possibly video) of a program, yet they have none of the logic which a program would need to create the same display. The intellectual content can be easily extracted from the text of the screens, but it is merely a shell and does not really have content without the accompanying verbal and/or printed presentation.

Multimedia Packages

These may contain elements of all the above. Almost by definition they can contain any other thing that can be imagined, including things of the type "multimedia." Thus storing and processing them can be done according to their components. The problem then becomes how to draw those components together to represent the whole.

As they say, "more work is needed on this."

CD-ROMs and DVD with Multilingual Sound Tracks

The interesting aspect here is not the medium, but that the capacity of the medium allows for not only different language sound tracks (and sub titles if desired), but also alternative endings (or middles or beginnings theoretically).

The questions are how to handle this from a descriptive point of view. Is there a "master" version which defines the content or is it the sum of all the variants. Is each language version treated as a separate work and catalogued so?

This is one which will be here before we know it; and we are not ready.

Cataloguing/Loading

Cataloguing has been touched on in the preceding section, as some of what has to be done is material specific.

The big issues are the completeness of the physical description part of the catalogue record and the granularity of the cataloguing.

Much digital material has been loaded into retrieval systems and the result called a digital library. Since the extraction of the data for the bibliographic record is usually done automatically, the resultant record is skimpy, to say the least. The development of the Dublin Core specification was an attempt to address this problem by defining a set of attributes which could, by and large, be extracted automatically or be assigned by unskilled people (such as the author). This means the bibliographic record is now very much the descriptive data not the content designation. Certainly controlled vocabulary and other quality control measures are much reduced.

This is not to say that automated assistance for cataloguing is a waste of time. Far from it. Automatic extraction of data and information can be a great time saver. However, the thrust of library automation systems has been towards sharing the results of human cataloguing rather than developing computer assisted cataloguing in the above sense. The vast amounts of data on the Web and in the new material types means that automation will be coming to library systems or they won't be able to keep up.

However, for now, cataloguing is a highly skilled manual process (as it will stay for a long time) where each item is considered on its own. Thus the degree of granularity of the new material types becomes an important question in terms of the workload of creating and maintaining a digital library.

©2003 Sun Microsystems, Inc.

The granularity required depends on the users and their requirements. If they are different from those of a traditional library then the granularity must change. Traditionally the cataloguing description has been at the level of a complete work. Its components have not bee described. Thus users who can now retrieve a single image from a document may want to be able to search directly to the level of the image, not via the level of the document. However, the majority will require what is done now in the well working model of the conventional material library.

As with any library, standards and rules must be strenuously applied to the cataloguing of the material. This does not change. The strength and utility of the catalogue depends directly on the quality and consistency of the cataloguing.

Loading and indexing the records has been touched on above in the capture sections. The indexing which is done from the bibliographic record is a part of the library automation system and is a series of decisions no different from those of a conventional library.

Indexing from the original material is material type dependent and end user and use dependent. What will result from a completely mixed material library is a series of indexes (and search tools) which are each organized round the particular material type. Thus images and videos may have an index of the material by the type and number of shapes in them. However, they can share a subject description index (in text form) with all the other material types.

The common denominator of these is the textual description applied (or extracted) during cataloguing.

In all cases the capabilities of the library and/or retrieval software will determine the scope of what can be offered to the users. In most cases it will fall short of what is theoretically possible.

Services

The library is not merely a searching tool for people to log onto over the Web or some other access medium. Conventional libraries provide a large number of other service. Many of these can and should be continued or extended in the digital library. Below are listed some of the major library services and how they could appear in a digital library.

All provide benefits for the library's users. There is nothing to say that these benefits have to come free. If the library has the remit and the software capabilities it may charge for all or some of the services it offers including basic access to its catalogue if it wishes.

Searching

Catalogues and their access are the most visible aspect of libraries, particularly when they are on-line In the on-line case they may have virtually no other presence. This is a shame as libraries generally have a lot more to offer than just an "it's over there" answering service.

The capabilities of the search system are generally going to be fixed by the software that you purchase. How you implement them and how the users use them is more under your control.

Vocabulary control is an essential part of the cataloguing process and is almost completely absent from full text searching. Searching using a controlled (and well known) vocabulary is a service which is particularly useful in the subject descriptions and makes for much better search precision than simple keyword searching. It also reduces user frustration at the extremes of "0 hits" or "3,465,789 hits" results. When controlled by a sensible thesaurus with descriptions and relationships it makes the user's job of finding the required meaning of a word much easier. It also allows for translation and thus the consistent use of the catalogue in a number of languages.

Natural language search statements are accepted by some systems, but are handled much better if processed by a human rather than a system. Thus a service to users could be a librarian mediated search service. This could derive the search from a written description emailed to the library and the results sent back the same way. It is not a feasible real-time service, but can have a rapid turn round.

An extension of this would allow searches to be spoken into a phone and then converted to searches as above.

Rather than mediate, the search library staff can filter the results to produce "better" output from the user's own input. Removal of duplicates is one not trivial, but very useful, service.

Delivery

Results of searches are generally delivered online, but may be offered via email or hard copy if the volumes are large.

The other delivery possibility which exists in a digital library is that of delivering the final object to the user. This would generally be online, but may also be by some bulk transfer (such as email attachment, ftp, CD-W, etc.) if large or if access is restricted.

Format/Quality

Particularly where results (whether results "hit lists" or the final required objects) come from a number of sources the quality and format of the items may vary considerably.

Lists could be post-processed to given a uniform format. This is done by some search engines, but not by all. If results are across different search protocols (such as Z39.50 and http) then the results will come back looking very different. Often these results are displayed separately for each source, which leads to problems identifying duplicates, etc. Post-processing of these lists is a possible service. One of only a few such search engine which addresses all of these issues is produced by the author's company, called Muse.

For the final objects there is little that can be done about the quality, since that is dependent on the quality of the original object. However, re-formatting may be possible to a limited extent. Certainly most times a uniform header for the objects can be extracted from the original data. This can provide a useful description of not only the object, but also where and when and how it was obtained.

Future improvements in encoding may allow more re-formatting into a house-style so that results may be presented almost as a sort of "one issue journal" or report.

Bibliographies

This leads directly to the creation of "bibliographies." This is in quotes because the contents of the bibliography probably will not be books, but some other material.

Bibliographies may be ad hoc just for a single user (who may well have to bear the processing costs), or may be standard ones prepared and run at specified intervals in anticipation of demand.

The extent of coverage and the depth of research and the volume of these bibliographies will depend on their intended use and target audience.

©2003 Sun Microsystems, Inc.

Research

A research or reference service is merely a composite of some of those mentioned above. However, it is a way to add value to the library's products by utilizing the expertise of the library staff.

Merely a more extensive form of searching with more initial contact to determine the user's requirements and probably feedback of initial results, this service may result in a beautifully presented bibliography or merely in a list of references. The results may be annotated or analyzed by the library (or external specialist) staff to add further value and may well utilize resources well beyond those of the library.

Discussion Groups, Fora, News

Since the library is on-line it is possible to run chat rooms or list servers for discussions on topics where the library is expert. These are best run as mediated groups so that they library staff may add value to the discussions.

They can be on any topic, but are most likely to center around the specializations of the library.

These may be reactive in response to user request for discussion on a topic. Or they may be proactive in that the library starts a group and sees if anyone joins in.

Variations on these groups are things like a "video of the week club" or a "top ten xyz" page. These are vehicles for the library to "advertise" its content as well as providing a useful service.

A newsletter or weekly news page or similar notice becomes another way for the library to bring its services (and noteworthy happenings) to the attention of a wider public. These may be subscribed to or may be "pushed" to peoples' desktops depending on library and organizational policy.

Support

There is every reason why the library should consider a "help desk" or support for its users. They are the experts and have an in-depth knowledge of the material in the library.

Legal

The biggest legal issues for libraries of any type are Intellectual Property Rights (IPR) and Copyright. Digital libraries are no exception to this, in fact, they have a worse set of problems than those of a conventional library.

Legal issues do not stop at IPR. Libraries have to arrange contracts for the right to use content which is not their own. Libraries have to contract with their users for supplied services and associated charges. There are the normal legal/contractual matters of hardware and software and other suppliers to deal with.

There are staff and other human resources legal matters to arrange.

Many of these are normal business matters and can adequately be dealt with by a corporate legal department. However, in matters pertaining to the data and the ownership and use thereof, it is a field where specialist advice is really needed. In the resources chapter (Chapter 7) there is a section which lists a number of places where specialist help and advice may be obtained as well as lists of some lawyers and legal associations. In the interests of yourself and your library please seek help from these sources or your advisors if you are at all unsure of you legal position.

This section raises issues and briefly discusses them, it does not offer legal advice or any interpretation of laws in any part of the world. Remember that an Internet based service is available in any part of the world and that your business practices, including the legal ones, must recognize this fact.

Intellectual Property Rights (IPRs)

IPR and copyright issues boil down to one question; "Who owns the content?" Subsidiary to that is the question of who is able to do what with the content. The reason for this is obviously that the owners of the content may place a value on it and do not want to see it distributed indiscriminately without them getting suitable recompense for its use.

This requires that you are aware of every piece of information on the system and keep track of what use is made of it. Even if you (or your organization) own the content for all you material, it is very good practice to know what you have and what has been done with it.

Knowing what you have is mostly a matter of ensuring that your software allows you to make the necessary lists and reports so that the material can be counted in the right way. This may well be according to the requirements of the IPR/Copyright owner and so it may be necessary to incorporate fields for this data in the database structure.

If the situation is at all complicated then it is best to move to a rights management subsystem right from the beginning, rather than struggle along until you find you don't know what is happening. Such systems are very new and are evolving rapidly. However, it will handle matters like:

- Different types of use
- Different "classes" of users
- Dates of use
- Number of times
- Levels of fees and discounts
- Promotional and educational use

These functions tie in very closely with those of measuring access in the next section. It is important to note that for the legal aspects it is important to record certainly a user's "class" even if not the actual user identification itself. The reporting requirements are very similar, but the emphasis here is on making sure that only those users who have permission are able to use an object, and this must be authorized on an individual user/object basis.

One of the requirements of good business (and a legal requirement of many reseller contracts) is that the material is adequately safeguarded. This is particularly interesting in the digital environment where the notion of a "copy" of something is just that much more difficult to control.

In 1993, the European Copyright User Platform was set up by the European Bureau of Library, Information and Documentation Associations (EBLIDA) with funding of the Libraries Programme of the European Commission (DG XIII-E4). Its aim is to encourage discussion of the legal issues in electronic services. It has drawn up a model licensing document. It has set up a Copyright Focal Point on the World Wide Web to host a moderated discussion list on European copyright issues, to provide access to documents on copyright issues and legislation. The Web-site can be found at http://www.kaapeli.fi/eblida/ecup. The ECUP secretariat is at: http://ecup.secr@dial.pipex.com.

Digital Watermarking

Even when a user has the right to view an object, he or she might not have the right to make a copy of that object. They almost certainly will not have the right to freely distribute copies of that file. All of this requires some policing. It is generally impossible to prevent an object displayed on the screen from being saved to file. Thus enforcement must reside in marking the object so that it is clearly identified.

©2003 Sun Microsystems, Inc.

This process is known as digital watermarking and the added data is the watermark. The watermark is both a proof of ownership and an authentication of the origin of the material. Thus, it serves the interests of the owner, the distributor and the reader.

The watermark may only identify the owner of the object or it may contain other information such as a serial number, information about use of the object, rights and fee information, or even a description of the object.

All of this has to be done in a way which cannot be removed without destroying the copy is readily visible to a user without requiring special software and does not destroy the object for the original legitimate copier.

There are a number of techniques for this for text, images, audio and video. All use different techniques (though those for video and audio are similar) and these are continually changing. New techniques are developed and, as old ones are circumvented, they must be replaced.

This is a specialist area and it is one of the subsystems that you will have to add to your ILS or basic retrieval system to achieve a working digital library.

It is an area of much current research and consequently there is an active discussion about it on the Web. See the section in the resources chapter for both discussion and journal information, and also for suppliers.

One particular adjunct to the digital watermark is of interest. The Digital Object Identifier (DOI) is a unique digital ID which can in principle be attached to any digital object. It has been created at the instigation of the American Association of Publishers and others as a means of electronically tagging each digital object so that they may be tracked. Through a remote database the use user may ascertain the owner of the IPR and the type of rights associated with the object.

Measuring and Charging

Whether your library is a commercial venture or not it is important that you are able to measure the use made of it. It would be even nicer to measure the benefit of the library to its users, but that is almost impossible to judge outside of laboratory conditions.

For most libraries the number of users and the amount of use made of the library's services measure success or failure. This section discusses some of the ways these measures may be taken for a Web based system and for general library functions.

The major public access to a digital library is to the catalogue and it is this usage which is the main measure of its effectiveness. Since the catalogue is likely to be Web based the measure of activity becomes very similar to those used by the Web advertising industry and these measures may as well be used. As well as being correctly applicable their use allows libraries (which may or may not carry advertising) to record activity in a consistent fashion and to make use of the sophisticated reporting tools increasingly becoming available.

The basic way of recording access is to record the number of pages retrieved. Web server software logs all hits on a site. A hit is a single request for an object, which may be a page of information, but may also be an image on a page. Most pages have between 5 and 6 objects on them. Reporting software is able to combine these hits into pages either in real time or by analyzing the transaction log of the Web server. Since a page represents a piece of information and its context it is a reasonable measure of information delivery.

The software also extracts and reports “click-throughs” which are when a user navigates from one page to another by means of a link on the first page. The combination of these two measures allows a library to measure the use of its site.

A conventional Web site is considered to consist of a number of pages, each of which is essentially the same in importance. The hits on a page determine which ones users looked at most often (and hence, presumably, found most interesting). Note that the length of time spent on a page (until the next navigation) can also be captured, and this is another measure of interest, although it may only record that the user went for coffee and did not read any of it!

The problem for libraries is that all pages are not of equal value. A query formulation page is not the same as the page of full text results in terms of user satisfaction. Also confusing the issue is that many page designs will involve dynamic elements, which change each time according to what the user has requested. Good examples are a hit set list page and a record display page. Heavy use of the former may not indicate that the user has found a lot of good stuff, but that he or she is frustratedly scrolling through lots of unwanted titles because there appears to be no faster way of getting the desired record.

More is not necessarily better in all cases. However, the numbers do provide a fund of raw data for careful analysis. And they do provide a positive indication of user activity.

For a library that is intent on actually making money out of its site, two possibilities are available. The simplest is to allow advertising in its pages. This may not prove popular with the users or it may be of no account. If this route is followed then access to one of the Web advertising services is probably the best way to go about it. Useful lists and information about what to do are on http://www.iab.org and http://www.cnet.com/Content/Builder/Business/Advertising/.

Remember not only that you can display advertising on your site to help fund it, but that you can advertise on other sites to promote yours. Directory and list sites are the main ones here, along with specialist hosts for “micro-sites” which are effectively a specialized shopping mall concept. Look around.

The other possibility is to sell the content of your site. To do this you must first either own the content or have an agreement with the owner which lets you sell it (see the previous section). After that you will need to decide what and how to sell. The most obvious is the content of the records in the digital library when they represent the full content of the original material. You can either sell the content on a “pay-per-view” basis or you can sell a subscription that gives unlimited access to all, or some section, of your library collection.

The Web server can accommodate an e-commerce application which will link to your library system as its product warehouse and handle all the details of the sale for you. Beware that this is a new area and most of the ILS will not connect to most of the e-commerce systems. Linking these to an accounting system and possibly to on-line credit card payment systems is an area where you will need to put together a specialized set of software, such systems are not currently available off the shelf.

Remember that whenever you sell data that is not your own, you will have to account for that sale to the owner and pay a fee.

If the ILS that you use for the library has a client which can be loaded onto the user’s workstation, then keeping track of user activity is relatively easy and the overall system functionality can be enhanced over that provided from a generic Browser.

©2003 Sun Microsystems, Inc.

If running through a Web Browser then to keep track of the usage of the system it will be necessary to utilize “cookies” (which are small files that the server stores on the user’s computer to keep track of their identity). Since this is an aspect which worries some people your use of and policy on cookies and the information you gain from them should be detailed on your site so users will co-operate.

Currently all metering is done in software however, there are products on the horizon, which may take this function into hardware in the future. For now the major systems are listed in the resources section.

It is worth remembering that the conventional library services such as reprints and even loans are valid services in an electronic medium and can be chargeable services also. In fact, almost all of the services discussed in “Services” on page 78 can be charged for if a way can be found to charge for them easily and efficiently.

Chapter 6

Pitfalls

Things can go wrong and they will! Here we look at some of the things that can go wrong and see what can be done to alleviate the problematic results or even bypass them altogether.

Essential Tools Don’t Arrive, or Are Late, or Don’t Work

Like many problems—once it has happened it is really too late. So the first things are advice on avoidance.

- Choose reputable suppliers
- Choose suppliers committed to the industry
- Choose existing products
- Defer any development to another project

The biggest problem here is that just the tool you want/need is about to be announced by a small start-up company. In fact you have been trying a beta version for a couple of days, and it seems just right for the job. It may be hardware (a faster scanner, a more powerful server, etc.) or it may be software (a database that lets you hyperlink documents, a video player that gives full screen over a 14.4 modem connection, etc.). You decide to use it and re-build your workflow around it.

Here you can see the problems because they are so obvious, but what about a new version of a software package from your existing supplier. It can cut the image color editing time by 33%, but it won’t be ready until next month. This is a calculated risk. If the gain looks big enough then you may decide to trust the supplier; especially if you are an important customer of theirs and can get problems fixed quickly.

 ©2003 Sun Microsystems, Inc.

The safe course is to go for the "bread and butter" solutions. They are the ones that have been around for a long time and everyone is using. This is the good and the bad point. The tools will work, but your collection won't stand out from the others.

This is all a consequence of the rapid rate of change of the whole IT industry where almost everyone is taking a chance on someone else getting things right on time.

Generally the larger the investment the more conservative you should be. Thus for your database software choose Oracle, Informix, or Sybase, for your servers choose Sun. For the exciting frills you can be more adventurous. Just make sure there is a fall back position so that when (not if) some of the tools fail you can still deliver a product, even if it is without all the balloons you would wish.

What to do when the inevitable happens? As in all crises—don't panic. Undoubtedly someone has been here before and there will be a way to salvage some, if not all, of the project. Unless the original material has been damaged (see the next section) you are no worse off than if you hadn't yet started. Of course, you are some months down the line with nothing to show.

Unless this tool is absolutely unique there are probably others which have been around for a while, which do essentially the same job. It is unlikely that you are doing anything that new that others have not tried it before so there must be a way to do it. Look around for an alternative. If you did a rigorous selection of the tools in the first place there will be the second and third place candidates. If not look at the Web shopping sites and journals for advertisements and ask colleagues or your consultant. Make sure your new selection does actually have a working version of what you want and get to try it if you can (of course you should have done this for the first choice, but the "newness" of it may have overcome better judgment—now is feet-on-the-ground time). If you have any significant proportion of your data processed (as may be the case if your server proves to be inadequately powerful once you start real volume loading) then make sure the new tool can allow you to use that without having to start from scratch again.

If you have multiple problems all occurring at once, attempt to solve them one at a time. Part of the failure may be a knock-on effect from one to the other. When you are now late, now is the time to lengthen your time line and do some things serially rather than attempt them in parallel.

You need to be in control and allow things to proceed at a pace, and in a fashion, where you can show some small progress over a small time. This is essential for everybody's morale and for the project's credibility. It is essential that you are honest about the problems and realistic about their effect. Virtually all projects slip a couple of times, but when a project is seen to slip at every progress meeting management support soon evaporates. Rather bite the bullet and exaggerate the extent of the delay so you have time in hand later, than minimize the slippage and have to ask for more time next week.

In some cases a complete re-think of the work (or design) may be needed. This doesn't mean that what has gone before is necessarily wasted, just that it can be used in a different way. One that doesn't need the late/missing tool!

Data Capture/Conversion is Late, or Wrong, or Incomplete

The answer here is prepare and plan then test and then do it all again.

Here are some of the safeguards to put in place to see that the time and money spent on data conversion isn't wasted.

General Problems and Safeguards

Divide the material into batches of the same type.
Put all the color 35mm mounted transparencies together for processing. Make a batch of all the newspaper clippings. This is the first step. To do it successfully you must know exactly what material types you have to deal with. Do a thorough inventory of the types of material and the volumes. Make sure the materials actually are the same. It is worth having a second person or group do the same inventory check independently of the first.

> A client thought that they had 6500 audio tapes in boxes which would need capturing with only a tape number on the outside of the box for identification. A chance opening during a project meeting showed that some of the boxes had a descriptive sheet inside the box. So, now there were two types of material where there had been one. Fortunately, this was discovered before the tapes were sent off for processing where all the information on the sheets would have been ignored or lost.

Divide the material within a type into small batches and test rigorously.
Small batches allow you to test all of the facets of conversion against the original. Thus if only three maps are to be scanned in parts it is possible to ensure that all of each of the maps is covered and that there are no gaps (overlap or no depends on the type of capture decided on). The purpose of this is to determine that the method decided on for this particular material type actually captures the required information.

It is also easier to load the small numbers of records into the final system and to evaluate if the information is captured adequately. Use of the map images may show that an overview image of the whole map at reduced resolution is required. It is then possible to revise the procedure and resubmit the batch without having to wait weeks for the data to be returned.

Insist test batches be returned exactly as for the production.
This enables you to test loading times and procedures as for the production. It also enables you to determine that a delivery mechanism works. For instance a devised scheme of returning one image on one floppy disk may sound nice at first (it makes the idea of physical backup sound attractively easy). However, when the first batch of ten images comes back on 15 disks it soon becomes clear that there will be a handling problem and probably a data corruption problem if there are thousands of images to handle.

You also need to test things like the postal delay for batches of material being shipped to the converter and being returned. These times can be significant and the problem of out of sequence returns of data must be considered. For most material it does not present a problem, unless sequence numbers of some type are being assigned on return. What it does affect is the process of batch control to ensure material and/or data is not missing.

After the whole batch conversion is working—send one more.
It is too easy for the project committee to decide that it is all working perfectly "with this one small change that won't affect anything else." The time wasted is small compared to that if the "small change" does affect something.

 ©2003 Sun Microsystems, Inc.

Whatever the timescale for the conversion initially—double it!
With a number of organizations involved things will go wrong. Even if everyone is an expert in what they do they are not in what the other people are doing and even in how to put them all together.

Since the capture/conversion is a labor, and hence time, consuming operation, it is a natural for parallel operation and it does work that way. Except it does not work at all when people are on holiday. Be sure you calculate working days and not elapsed days when setting up schedules. Make sure you know where the material is going.

It only takes one missed piece of data to be discovered halfway through a big material type for the whole schedule to be thrown badly off. Check all the initial batches on a random sampling basis and then randomly sample the later ones at a rate consistent with the throughput. This means the early days will be at a lower throughput than the later. This is sensible for the people doing the work as well as they have to learn the material and what is required. This is particularly true if the material or techniques or information required is new to them.

On the other side, do not assume that the early production rates will speed up dramatically and allow you to "catch up" later. Use the test batches to time production as well as get the procedures right.

Thoroughly document all the procedures.
This will happen automatically if you use a professional conversion company. They will insist on it as well as acceptance criteria for the returned data. You must do it even if it is your own staff undertaking the work. Set acceptance criteria (permitted number of errors, number of missed images, number of video clips without sound, etc.) and devise ways to test for them. Make sure that everybody knows the procedures and the tests and that if a test has failed then the whole of that batch will have to be re-done.

Staff will get sick and even leave and will be rotated. If procedures and standards are not written down as they are to be done then the new staff will have no chance of getting things right and quality will suffer.

Remember also that this will probably be an ongoing process so someone will have to digitize an annual report once a year. Even if it is the same person they will have forgotten the decisions that were made last time.

The documented procedures must be the ones used not the ones the consultant suggested which are mainly bypassed. Change the documentation as the procedures change.

The procedural documents and the testing documents are the key to the quality and timeliness of the data. This where an iterative process is required in the early stages to refine the procedures and the amount of testing (and hence rejection) so that an acceptable quality is achieved within an acceptable time. The two values here are ones that only you as the head of the project can decide on. Time to market and market acceptance and cost of production are the factors to juggle.

Data-Specific Problems

Different material types have different special areas where they can go wrong. Here are some of them.

Images

- Covers of reports or books, what about the spine?
- Both sides of pages
- All the edges of pictures
- All the required formats and resolutions have been produced
- Color has not been lost
- Optical corrections have been applied (or not depending on what is required)
- Size and aspect ratio has been maintained

Video

- Both beginning and end are there—it is easy to stop converting at a black frame near the end
- Check the content for a short clip—file names are easily mixed up especially if they are sequential numbers
- Check color match with original
- Check for sound
- Check edges of image have not been lost due to differences in movie, TV and computer screens
- Check that not too many frames have been dropped

Audio

- Check for the wrong sampling rate—high frequencies will be lost
- Check stereo vs. mono vs. 3D vs. surround sound according to what is required and the original
- Check for end of recording—easy to stop at quiet period near end
- Check for audio processing (such as bass enhancement) if required

These are the main material types for digital libraries. Other material can be included but is currently much less frequent. Some of these tests are very technical and may not be required, some are just checks against sloppiness.

Is Anybody There? Communications Problems

In this era when almost all access to a digital library is over the Internet or through a local network, the issue of communications problems takes on a whole new importance. Not only do you have to master capturing your material and loading it, now you have to become a network engineer as well. Well, not quite. What you do have to have is a basic understanding of the issues concerned with letting the great public access your material.

What can go wrong with communications? Well it breaks down into four main areas

Servers

Servers are only computers and they do fail from time to time. They are built to a higher standard than desktop PCs and often have redundant components which can fail, and the server just keeps on going. But sometimes the redundant parts fail as well, or one of the servers in the chain doesn't have them. So machines do fail.

All you can do is look out for the machine on which your digital library is running. It is very unlikely that you will have any say in how the other servers in the communications chain are configured. They will belong to the telecommunications company, your university, or even a foreign company.

©2003 Sun Microsystems, Inc.

Your compromise is, as almost always, one of money spent or saved against features. For your server you can decide to have a simple PC. It will work well for months or possibly even years, but it will fail and have to be repaired at some point. When it fails your Web site and hence digital library is "off the air" If this happens for a couple of days it may not matter. If that is unacceptable then look higher.

Buy a server with redundant components. As long as it is maintained it should be up for all except a few minutes in a year.

Don't buy a server at all, pay to have your Web site hosted on a commercial hosting service. They can afford all the things you can't—backup power supplies, air conditioning, redundant communications lines, even spare computers to move your Web site to if the worst happens. And you can increasingly get a quality of service agreement from them, guaranteeing a certain up time—but it will cost you.

A simple solution to any level of server is to have a "mirror site." This is a computer which does exactly what you server does, situated in another location. This may well be halfway round the world. If you have colleagues who can accommodate your Web site on their server (maybe asking that you reciprocate, so be careful what you ask for), then this can be a good way to keep up and running. You have to arrange to update both sites at once, and you may wish to have the two or more sites take over from each other automatically if one fails. This is perfectly possible, but puts the cost up.

Local Network

If you are in an organization, then you are probably connected through a Local Area Network (LAN) to the Internet. This may be a weak point. You will have very little control over what is set up, but make sure the IT department and the networking staff know you are running a server and that it is accessed all the time and from outside the organization.

Once you tell them this, they may well insist that your digital library resides on one of the servers in the IT department. This is not altruism on their part, but usually to protect the integrity of the LAN and its security measures. Talk to them in the early stages of setting a Web site up. It could save lots of heartache later.

Internet

The Internet does have problems. The beauty of it is that they are generally local and the Internet is so vast and connected that any local problems will not affect most users. Whether the problems are as mundane as a broken cable (catastrophic if you are on the wrong side of the break), or a major denial of service attack by hackers, or massive bottlenecks due to popular sporting or news events, they are unlikely to last long or affect you very much.

Having a mirror server at a remote location obviously removes many of the worries about local problems.

The best advice for an Internet problem is to not worry about it. Just sit and wait it out. First make sure it is an Internet problem, of course, and not a change your IT people made last night, or an unplugged computer.

Administration

The last area is not a technical problem at all. Most access to the Internet is through a commercial Internet Service Provider. They are commercial organizations, So, if you can't access the Internet, check that you have paid your bill this month!

Part 3

Resources and the Future

...Or "Help! Where Are We Going?"

Here you will find a large number of references to other sources. Some are paper, but the vast majority are electronic and available on the Web. This is deliberate as they are likely to be more current and easier to get at. This does presume that you have Internet access. If you don't then seriously consider it for the sake of your sanity during this project.

The references are, necessarily, a mixed bag and you will find both beginner's texts and advanced discussion papers. Use the navigational facilities of the Web and your librarian skills to move from the starting points we give to those areas which you really need to read.

Some resources contain useful information, others are places where you can start to look for what you need. Many of them supplement and extend what is contained in this document. Some are commercial supplier sites and they will be, understandably, touting their wares, not those of the competition. There is much useful information to be gleaned from these sites (particularly about technology advances), but glean with care.

In addition to the resources listed here you might like to utilize the search engine at the MuseGlobal site. This is a metasearch engine which will allow you to enter advanced searches (including Boolean logic, brackets, and proximity operators) and have the search applied to about 3 dozen search engines of your choice. These include both Web search engines and libraries, so the results are a mixed bag. They are re-formatted and interfiled so they are easy to peruse.

©2003 Sun Microsystems, Inc.

A free (Web search only) service is available or you can register and use the full power of the system—still for no charge at the moment. The resources for Chapters 7 and 8 were partly culled using early versions of this engine.

Chapter 8 is not really about resources in the same sense as Chapter 7. It contains places (on the Web) where you can visit digital library projects or related research work. Some are "complete," most admit that they are "works in progress." Whatever their state, except for those that have been abandoned and deleted, you will find ideas you can adopt and modify. You will find people behind the projects you can talk to and ask for help or compare notes.

Chapter 7

Resources

This is the reference section where all the pointers to other resources are located.

Standards, Formats, Protocols

These are the rules by which objects are described, their data is stored and the systems communicate. This is an eclectic mix of such rules. It is not exhaustive, the members are chosen because they are useful, fundamental or illustrative.

There is nothing in this section which you can buy (except the definitions and many books). The items are here because you will have to make some decisions or ask some questions of your suppliers and these are the names that will come up.

Some are international standards set by bodies like ISO (International Standards Organization), IETF (Internet Engineering Task Force). Some are national standards set by bodies like NISO (National Information Standards Organization) in the U.S. or BSA (British Standards Authority) in the U.K. Some are industry standards set by industry bodies. Some are corporate standards produced by a single company and accepted by widespread usage. Some are not standards in any sense except they are in widespread use.

In all cases the function is to try to unify the representation, manipulation or transmission of some piece of information so that two or more different systems can “understand” it the same way. They are the basis of interoperability, portability, modularity, building blocks, objects, and all the other names invented to describe how two pieces of software should be able to simply work together. Be very wary of claims to easy and complete conformance and ask to see examples or a demonstration.

 ©2003 Sun Microsystems, Inc.

Generally standards change fairly slowly (some would say extremely slowly at the official National and International level) and a superseding version will (usually) require conformance to the earlier version. This is not always the case, but the exceptions are rare and are documented in the standard. It is well worth the effort to access these documents on-line to make sure ones you are interested in are not about to become obsolete or to change their function radically.

So called "commercial" or de facto standards do change rapidly and often in a way that is deliberately designed to block systems built to an earlier version. These "standards" are often devised and promulgated in an attempt to gain market advantage and so it is in the interest of the promoting organization to "lock in" other users and to "lock out" their competitors. The Microsoft/AOL "war" in late 1999 over the "standard" used for access to on-line chat rooms is a classic case. The de facto standard (really a protocol) used was changed every day over a period of about a week to a form which was incompatible with the earlier version so the "competition" could not access the service. This is currently a rare occurrence, but may become more common as the Internet is commercialized more and more. You can't do much about it except to be aware of the possibility of it happening with proprietary ways of doing things.

Bibliographic

These are concerned with the description of the material, both as to its content and to its physical and descriptive attributes. They are generally very complex (MARC has some 800 field definitions) and cover the most difficult, intellectual, part of the object definition. These definitions are necessary for processing the material and also for searching for it.

These definitions are all a form of metadata in that they are information about the basic record (the "data").

MARC

lcweb.loc.gov/marc

A standard for recording bibliographic data at the logical level. Contains elements for both content and physical and process description. This is not a single standard, but rather a framework within which each country has developed an individual standard. USMARC is the standard from the United States and is used as is by many other countries. Generally maintained by an office within the relevant National Library. USMARC is maintained by the USMARC office within the Library of Congress.

A new MARC format, MARC21, has become a standard within the last couple of years and is being adopted as a common format by the Australian National Library, the British Library, the Canadian National Library and the Library of Congress. It is hoped this will be adopted by other national libraries (many of whom base their existing standard on those of these four libraries) and become a de facto "world MARC."

Dublin Core

purl.oclc.org/metadata/dublin_core

This is another standard for record content and descriptive data. Much simpler than MARC (only 15 elements) and devised for use across the Internet to allow a common description for professionally catalogued material in libraries and for "amateur" material at other Web sites. MARC and Dublin Core data elements can be interchanged according to a prescribed scheme for user display purposes.

Dublin Core is undergoing an evolution from its simple 15 field form to a more comprehensive "qualified fields" form. Within this the original fields are qualified for more precision (so "Creator" could become "Creator.personal"). This allows a better mapping to the more comprehensive and specific MARC records. It also allows flexible extension so that other data elements (fields) may be added as needed, even for private use through the idea of "NameSpaces" where a particular set of fields and qualifiers is given a name and deposited in a well known place so any application can find out about them and their syntax and rules of use. This is a similar idea to the use of Schema in XML rather than a DTD in each document.

BIB-1

lcweb.loc.gov/z3950/agency

A simplified record structure for online transmission. It is essentially a sub-set of MARC. It is the original format for transmission of records within a Z39.50 dialogue between two systems. It has elements that can be mapped to both MARC and the Dublin Core.

During 2002 a number of organizations have started to develop and promote a BIB-2 standard which is simpler than the original BIB-1, in the hopes it will become more widely acceptable.

Text Encoding Initiative (TEI)

The initiative provides a scheme to encoded text so that parts of it such as the start and end of lines, paragraphs, pages, chapters, acts, and so on can be marked. Thus such text can be processed to produce accurate indexes for searching. Other features of the text both grammatical and linguistic and also content indicating such as the actors in a play can be identified allowing for a rich analysis. These rules require the actual text be marked up with SGML encoding.

A brief introduction can be found at **www.dsu.edu/~johnsone/tei.html**.

Electronic Archive description (EAD)

lcweb.loc.gov/ead

An encoding scheme devised within the SGML framework to define the content designation of documents and other archival objects. It is defined with a minimum number of descriptive elements, but in an extensible fashion. It is designed to create descriptive records which will assist in searching for the original material in a number of ways.

Federal Geographic Data Committee (FGDC)

www.fgdc.gov

A metadata standard for the description of the elements of maps and other cartographic objects, including such attributes as scale, projection, co-ordinates (and co-ordinate scheme), etc. This is just an example of a number of specialist descriptive schemes for different objects and material types.

Metadata

This is not a standard or even a description. It is a currently popular word which correctly describes the class to which all the above descriptions belong. It is literally "data about data" and thus is the description of the structure of the record which actually holds the data. Since it is a class of things it can have no single useful description. However, it is being used by a number of vendors as a flashy buzzword as in the (made up) phrase Our system incorporates the latest bibliographic meta-data paradigms for internal object representation." A translation for this is "Our system uses MARC for its record structure."

©2003 Sun Microsystems, Inc.

Anglo-American Cataloguing Rules

These are a set of rules which define how an object is to be described. They are entirely intellectual and are concerned with such things as the consistency of people's names and subject descriptions. They are not absolutely tied to any standard format, though they are developed in conjunction with MARC records. They (or a close derivative) are in use in libraries around the world. Russian and German libraries developed their own similar rules for the same purpose (GOST and RAK), but AACR2 is by far the most common.

Classification Schemes (Dewey, UDC, BSO, etc.)

These are intellectual schemes for the ordering of knowledge and are used for assigning a work to a class along with works of similar content. They are also used as the basis for the physical ordering of physical objects (books, tapes, etc.) on shelves. They could be used for assigning a "location" to an electronic object—most likely as a way of deriving a unique file name for the object. Dewey and UDC are far and away the most common schemes.

Addressing and Directories

URLs

These are Uniform (or Universal) Resource Locators and are the addresses of objects within the Internet. As such they satisfy the requirement for Uniform, but because they are limited to Internet (and more generally World Wide Web) use they are not Universal. They are the links between sites or pages on the Web, which allow the linking (or Hyperlinking) which provides the navigational functionality of the Web. They are not bibliographic in nature but can be used to provide a non-linear logical structure to documents on the Internet or Intranets.

RDF

www.w3.org/RDF

Resource Description Framework is an attempt to define resources (databases, search engines, library catalogs) available across the Internet (more correctly within the Web) in a way that allows applications to identify them and their properties. The actual RDF records about a resource are formatted in XML for wide interoperability.

It is a framework and actual collections of fields within the framework have yet to emerge for most applications (including bibliographic), so it is too early yet to expect to see RDF data around.

UDDI

www.uddi.org

The Uniform Description, Discovery and Integration standard is a method for businesses (primarily) to define what products and services they offer in a consistent manner so that other systems can find method to interwork with these services. Library systems are just starting to think about applying UDDI to their world, but it will become more important quickly in dealing with booksellers and publishers.

WSDL

http://www.w3.org/TR/wsdl

The Web Service Definition Language is an XML format for describing network services to be used as part of a larger communication where two (or more) systems interact across the Web. It provides a way of defining what the systems can do. In reality it is very similar in function to UDDI, but they are proposed and supported by different groups and will co-exists for a while.

Record Structure

These define the physical and logical structure of the record which holds the data. The very simplest of them hold only a single type of data (such as an image) and they are listed later in the section on formats. The records considered here are complex in that they contain multiple fields of variable length and which may occur more than once. Except for proprietary structures there is really only one structure used for bibliographic data of any complexity. These formats are for exchange of data between systems and are not intended for human consumption.

ISO 2709/Z39.2

This defines a very flexible structure for individual records and wholes batches of them (originally for tape storage) which is exceptionally well suited to handling the MARC format. (They were developed together, but 2709 can be the structure of almost any type of record.) The main strength of 2709 is its ability to simply handle variable length fields and records where the occurrence of fields is also variable. It is much too complex and time-consuming for standard business records where all data fields have fixed lengths and positions and always are there (even if they have no data).

SUTRS

Simple Unstructured Text Representation Scheme is a format for structuring bibliographic records. It is, in essence, a method of displaying MARC (q.v.) records (in an expanded form with English text tags and no structure other than that of a piece of text. It is human readable, can be edited with a simple word processor and is easily parsed by receiving systems to extract the information content. It can be used with other forms of record (such as Bib-1 or Dublin Core), but usually is used for MARC records.

GRS-1

Generalized Record Structure (-1) is exactly what its name implies and is a very flexible format for representing bibliographic and associated data, such as holdings. information and museum and archive information. It is a fielded structure which is human readable and easily parsed by receiving systems.

XML

Extensible Markup Language is a general purpose scheme for defining the content of documents and other record type. It is derived from SGML and is specialized for defining content rather than layout or design (see HTML). Each record structure is defined in a DTD (Document Type Definition) which says what fields may be present and any rules about their use. The DTD may be contained within the record or it may be externally referenced. XML is the current great white hope for exchanging information between applications across the Internet.

There exist XML versions of MARC and other record descriptions, which Z39.50 servers and clients should be able to deliver and receive.

©2003 Sun Microsystems, Inc.

HTML

HyperText Markup Language is a scheme for describing the layout and presentation of a document (the "look and feel"). It is a derivative of SGML specialized for its layout function. It is the format in which Web pages are defined and transmitted. With minimal exceptions (<TITLE>, <KEYWORDS> and <DESCRIPTION>) it does not define content, merely layout and decorative aspects of the document.

It should be a complement to XML, but the two are something of rivals, though for must users purposes it is not obvious. Browsers will display XML documents as well as HTML ones and both have the ability to define a document in the others terms. It is possible that XML (the newer format) will replace HTML as the de facto standard for Web record formatting as more and more applications wish to exchange data.

Encoding

This section concerns the way individual characters are represented in the files and records. It is concerned with text within records almost exclusively. Theoretically, other material objects (such as images or sounds) could be encoded within a complex record (where there could be a text field, an image field and a sound field), but it is very rarely done. These combinations are done either in a meta-record (such as the bibliographic record) or within the text record itself. The other records are referenced (as in an HTML page with a reference to an image file). However, not only is text data combined in complex records, it is also represented in different ways. The most obvious example is a record that contains two pieces of text in different languages. To make this worse they could be in different character sets. This is what character encoding is about.

Characters have long been represented by one byte (composed of 8 bits and capable of representing 256 different values) per character. This is quite sufficient for all Western European (Indic) languages taken one at a time. However, when wishing to encode characters of two (or more) different character sets there are not enough numbers (called code points) to represent all the characters needed. Thus for completeness in a multi-lingual world it was necessary to enlarge the number of available code points.

A number of attempts were made at a national and international level. Various Double Byte Sets (DBS) were developed, particularly in Asia where the problem is acute. Various schemes for adding bytes where needed by "shifting" from one character set to another were developed (such as ISO-2022). These had the problems that they were local and/or cumbersome to manipulate and hence slow.

Unicode

www.unicode.org

This is a universal encoding scheme using 16 bits to represent each character. It has the advantages of being simple, complete, and is being widely adopted. Its disadvantage is that all characters take twice as much space even for single language data. However, disk storage is getting cheaper (text is very small compared to images and video) and there are ways of simply and speedily compressing the data in storage. Unicode is controlled by the Unicode consortium and is the operational equivalent of the ISO-10646 standard. Note that 10646 also defines 32-bit characters, but these are not in any general use.

Although Unicode is a 16-bit encoding there are a number of variants of this value which are used for compression or, more importantly for Unicode acceptance, for compatibility with existing 8-bit character) to a variable width encoding. UTF-8 in particular is useful because the ASCII characters are encoded in one 8-bit byte—exactly as in the ASCII encoding. Thus software designed to handle UTF-8 encoding can handle (read and write) old fashioned ASCII files from legacy systems.

ASCII

There are a wide variety of 8-bit character encodings in use round the world, but the most common is that of the American Standard Code for Information Interchange (ASCII). This defines all the characters necessary for English and many special characters. This code has been used as the basis for most other 8-bit codes. The "lower" 128 are left alone (they contain the Latin alphabet, numbers, control codes and some special characters) and the "top" 128 characters are used for a second language. Thus there is almost universal compatibility at the "low 128" level and almost none for the rest. IBM/Microsoft produced a number of "National variants" for the PC DOS operating system and these have a large measure of acceptance through wide distribution—however, they are only a manufacturer's standard.

Communications

There are many layers of communications (seven if you consider the "OSI seven layer model"), most of which do not concern us here. The one that is important is the level at which the computer systems connect to each other to create a connection that will pass our messages back and forth. There is one protocol for this level that is by far the most common. It is also the protocol of the Internet.

TCP/IP

This protocol called Transmission Control Protocol/Internet Protocol (TCP/IP) is for controlling the creation of transmission paths between computers on a single network and of connecting between different networks. It is in almost universal use for public networks and many in house local area networks. It is the protocol of choice for all UNIX servers (Sun uses it universally) and most workstations. The only reason not to specify it is if your system will be entirely in-house and the existing network uses something else.

Protocols

These are the "language" of the messages passing between the systems connected via a TCP/IP (or other) protocol. There are a variety of protocols for different purposes, which may be used at different times by the same two systems or by one system "talking" to two others.

http

This protocol (HyperText Transfer Protocol) is the protocol of the Web. It is used for carrying requests to the Web server and returning pages to the user client. It is also used for requests from one server to another. It is limited to a fairly simple "request–supply" structure; though this has been extended by encoding search and processing data within the "request" (and sometimes the reply). This protocol is usually not directly supported by ILS or DBMS and needs a Web server connected to the "background" system to provide an Internet presence for the library.

 ©2003 Sun Microsystems, Inc.

ftp

This protocol (File Transfer Protocol) is used for exactly what its name suggests. It is a file transfer protocol and is universally used across the Internet for shipping files be they large program downloads or small emails. Note that your email actually is sent from your client machine to the postmaster using either Post Office Protocol (POP) or Simple Mail Transfer Protocol (SMTP), it is from there to the destination where it may be batched and sent from "post office" to "post office" by ftp.

Z39.50, ISO 23950 (ISO-10162/3)

http://lcweb.loc.gov/z3950/agency

This is a NISO and ISO standard for searching (and retrieving) across more than one library system. It is primarily a protocol for between library and information retrieval systems. It is not used by the Internet search engines (they use http). It is more complex and more comprehensive and powerful than searching through http. It has been extended to allow system feedback and intersystem dialogue. Thus it is being talked about for non-search functions such as interlibrary loans.

It is a continually evolving standard and the current version is the third. However, most ILS vendors support only the second version features or a very patchy sub-set of version 3. As it is a protocol for searching between systems and is not used by humans directly, it is syntactically complex. It also attempts to be wide ranging and, in particular, support a number of search languages (such as RPN—support for which is mandatory in all Z39.50 targets (the search engines))and a number of data formats for the retrieved records (such as BIB-1, SUTRS, GRS-1). It is currently being debated how the protocol may be extended to accommodate XML structured records.

Z.39.63 (ISO-10160/1)

http://www.nlc-bnc.ca/iso/ill/

http://www.niso.org

These are the NISO and ISO interlibrary loans protocols. Unlike the Z39.50 (ISO10162/3) protocols they are not identical, merely functionally equivalent. Most ILSs support these functions through a stand-alone module and some through access to third party services.

In North America (and much of the rest of the world) the OCLC ILL protocol is used, and virtually all the major library systems support it. It is a "commercial" standard, but is in widespread use.

http://www.oclc.org

An example of a third party ILL service is **http://www.cps-us.com**.

An EC funded document delivery project with a good bibliography and list is AIDA at **http://liber.cib.unibo.it/aida/AIDA_letter0.html**.

Formats

These are not standards in the sense that they are formally ratified as are those described above. They are the forms in which the information of the digital library is held. Many are set by commercial organizations and come into common use through the success of their parent organization in sponsoring them and encouraging their use. They are for low level physical data organization and, as such, deal with only one type of data each.

They are listed by their file name "extensions." No attempt has been made to describe them as this is very adequately done in many books and a search on the Web for the three letter groups will bring up many document describing the formats and their use and restrictions. Even some of the programs used for manipulating the files contain descriptions and references.

Images
.BMP .TIF .GIF .PNG .WMF .PICT .PCD .EPS .EMF .CGM .TGA .JPG .PNC ...

Animation
.ANI .FLI .FLC

Video
.AVI .MOV .MPG .QT

Audio
.WAV .MID .SND .AUD .MP3

Web Pages
.HTM .HTML .DHTML .HTMLS .XML

Text
.DOC .TXT .RTF .PDF

Programs
.COM .EXE

Associations

These are groups of various degrees of formality that are active in the library, computing, digital library and associated areas. Most of them are American simply because they are the most visible. Virtually all Nations have a library association and a computing association. Organizations like IFLA contain such directories within their Web sites. All these organizations have Web sites that are extensive and contain a number of sections that may be of interest.

For more complete and updated information and directories visit the Web site for this paper at http://www.edulib.com/ or link through the sun site at http://www.sun.com/edu/.

Library

These are associations that are concerned with libraries and librarians in general.

IFLA	International Federation of Library Associations	http://www.ifla.org
ALA	American Library Association	http://www.ala.org
SLA	Special Libraries Association	http://www.sla.org
LA	Library Association	http://www/la-hq.org.uk
RLG	Research Libraries Group	http://www.rlg.org
CNI	Confederation for Networked Information	http://www.cni.org

©2003 Sun Microsystems, Inc.

Digital Library

These organizations are either associated with the computing and information science aspects of digital libraries or are funding research into this area.

NSF	National Science Foundation	www.nsf.gov
DLI	Digital Library Initiative	www.dli2.nsf.gov
SIGIR	Special Interest Group—Information Retrieval (ACM, BCS, etc.) q.v.	
ASIS	American Society for Information Science	www.asis.org
DLF	Digital Library Federation	www.clir.org/diglib/dlfhomepage.htm
UKOLN	UK Office for Libraries and Networks	www.ukoln.ac.uk/services/elib

Computing

These are general computing associations.

ACM	Association for Computing Machinery	http://www.acm.org
BCS	British Computer Society	http://www.bcs.org.uk
IEEE	Institute of Electrical and Electronic Engineers	http://www.ieee.org

Standards

ISO	International Standards Organization	http://www.iso.org
NISO	National Information Standards Organization	http://www.niso.org
IETF	Internet Engineering Task Force	http://www.ietf.org
W3C	World Wide Web Committee	http://www.w3.org

User Groups

These are informal groups of users (and researchers and even vendors) for discussing digital library topics. They are ever changing and can best be reached by contact through a particular vendor's Web site or a particular library or via one of the organizations listed at the start of this section. They can often provide the real world addition to a vendor's rosy view of their products. They are also a source of experience and a forum for raising questions.

Many of these groups have active participation from their respective vendor. Often other vendors will participate (such as a hardware vendor like Sun participating in the user group of one of its partners such as Geac) and many will monitor what is said. This adds benefit to the users in that their suppliers remain aware of their needs and grievances.

Publications

These are publications that have content of interest to digital library projects. There are many more which deal with libraries and information science and digitization and computing which from time to time deal with digital library topics. Note that Highwire Press is included as it is actually a digital publishing house.

A search of any library catalogue will reveal many journals that are suitable in whole or in part for information and ideas on digital libraries.

D-Lib	http://www.dlib.org
Highwire Press	http://highwire.stanford.edu
Journal of Electronic Publishing	http://www.press.umich.edu/jep
Digital Library Newsletter (of IEEE)	http://cimic.rutgers.edu/~ieeedln/
Current Cites (Berkeley)	http://sunsite.berkeley.edu/CurrentCites/
Marketing Library services	http://www.infotoday.com/mls/mls.htm
Journal of Digital Information	http://jodi.ecs.soton.ac.uk/
Biblio Tech Review	http://www.biblio-tech.com
Library Hi Tech	http://www.mcb.com/
LJ Digital	www.ljdigital.com
Scholarly Electronic Publishing Bibliography	http://info.lib.uh.edu/sepb/sepb.html

Vendors

These are organizations which sell all or part of what is needed for a digital library. With the possible exception of the “ILS” and “Digital Library” sections there is no attempt to be complete. There are many hundreds of vendors in some sections. These are merely either the dominant ones or a representative selection. There will be local manufacturers in many countries and many of the companies mentioned will have offices or distributors world-wide.

Hardware

Servers

These are the larger computers needed to hold the database of material, run the searching and processing applications and handle communications with the users.

The users are always considered to be on remote clients even if they are running on Sun Rays in the next room. In computing terms they are remote and distinct from the servers.

A server needs to be computationally powerful, have adequate main memory (RAM) to handle the expected work load for the given software, have large amounts of secure disk storage for the database(s), have features to allow them to perform non-stop for very long periods and have good communications capability. In the totality of a digital library design a number of specialized servers are needed. However, for a smaller library many of these logically distinct servers may be combined into one machine. Thus, the ability to grow and to distribute the

©2003 Sun Microsystems, Inc.

processing, storage and communications load across more than one computer is important. A family of servers such as Sun's Enterprise™ server range is what you should look for to ensure there is sufficient depth to allow for growth without pain.

Sun Microsystems	http://www.sun.com

Desktop Workstations

These are the personal computers used as the clients for the users to work on. They are strong in display and communications and relatively powerful computers. They may be regular PCs running a graphical or text interface through special software loaded on each PC, or they maybe network computers (for example Sun's Sun Ray workstations) which automatically download software from a local server to run as required, or even run it on the server. The requirements for different digital libraries will vary enormously and some libraries will have no requirement for workstations, as they will expect all their users to connect remotely using their own computer and a standard interface like a Web Browser.

Capture Devices

This is a small selection of suppliers chosen more to show the range of capture devices available than as a sensible shopping list.

Umax	Flatbed Scanners
Logitech	Single Sheet Scanners
Creative	Video Cards, Audio Cards
Diamond	Video Cards
Matrox	Video Cards
ATI	Video Cards
Turtle	Audio Cards
Kodak	Digital Cameras
Polaroid	Digital Cameras
Olympus	Digital Cameras
Xerox Imaging Systems	Scanners, Cameras

There are the normal large amounts of peripheral hardware necessary for a digital library as for any other installation. This includes items such as:

Communications	LAN
	Router/Switch/Hub
	Modem/ISDN/Terminator
	DSL “Modem”
	Wireless Access Point
Storage	RAID Array
	Tape/Disk Back Up
	Uninterruptable Power Supply (UPS)
	Printers
	Consoles and Test Computers
	Fax Machines
	Telephone System

The size and amount of such devices depend on the size of the operation and the types of activities undertaken.

Software

This section only lists some of the available suppliers. Some sections contain essentially all the suppliers (as measured by market share) in the section, others are merely a representative sample and a search of any software catalogue or on the Web will yield many more.

Remember that in a number of the categories there are programs available either free or as shareware which equal the capabilities of some of the much more expensive offerings. It is well worth spending some time reading the product descriptions and even trying some of them from the download sites listed. Again, this listing is not complete.

The software is broken down into two groups; that which you will need to digitize your data, and that which you will need to run your Digital Library. The digitization group comes first.

Capture

This software allows you to capture either audio or video from a suitable peripheral device (microphone, tape player, VHS camera, VCR, etc.) and store the result in a computer file.

Image capture from documents is usually performed by the software that comes with the scanner used for the purpose. This is produced by the scanner manufacturer and is specific to that piece of equipment. Thus you really don’t have much choice. However, for a large scanning job it is as well to consider the functionality and ease of use of the scanner capture software as the optical characteristics of the scanner.

Bear in mind that digital cameras can be used for image capture and single frames can be captured from a played video. Digital cameras nearly have sufficient resolution to capture documents to process them to extract the text. However, if you want pictures of people and places and actual objects they will probably be the best solution.

©2003 Sun Microsystems, Inc.

Modern video cards are capable of capturing images and video from an input source. Thus a special video capture card is not necessary. These cards (see above) come with their own control software just like the scanners. However, the existence of standardized systems interfaces to the cards means that you can sensibly consider using a piece of third party software to run the video capture process. All the companies below make video capture and editing software programs.

Creative http://www.creative.com	Video Blaster	Audio, Video
Matrox http://www.matrox.com	Rainbow Runner Studio	Audio, Video
Microsoft http://microsoft.com	Video for Windows	Audio, Video

Manipulation

These are just a few of the programs which will allow you to manipulate your images, sound files and videos after they have been captured. This manipulation may be as simple as cropping unwanted edges off an image or producing a thumbnail version. It may be as complicated as editing a number of video and audio elements into a video presentation.

There are many other programs which can perform these functions, and some of the best are shareware. A good place to look and download and try out software is at http://www.download.com. This is the download site of http://www.cnet.com/.

The audio studio software is usually bundled with a sound card, which it controls. The other image and video may be bundled with hardware or may be generic.

Adobe http://www.adobe.com/	Photoshop Acrobat FrameMaker	Images Documents Web sites
Jasc Inc. http://www.jasc.com/	Paintshop Pro Media Center	Images Media Management
Corel http://www.corel.com/	Bryce 3D Draw	3D Worlds Images
Creative http://www.creative.com/	Ensemble-- Wave studio	Audio
Macromedia http://www.macromedia.com/	Director Shockwave	Video Animation
Matrox http://www.matrox.com/	Rainbow Runner Studio	Video
Coffeecup http://www.coffeecup.com	Coffeecup HTML editor	Web Sites

Integrated Library Systems

The following organizations represent the major suppliers of Integrated Library Systems. They are not the only suppliers; the majority of smaller (and possibly more innovative) suppliers have not been included in this printed listing.

All of these companies provide systems which will automate all the major functions of a library whether conventional or digital. Many have special modules for non-book material and even for digital material. Many also have Web access facilities to make the catalogue (and sometimes other library services) directly available on the Web.

The systems are all extensively described in their Web sites. A more extensive listing and a detailed comparison of their architecture and more specialized features can be found on a number of the library journal sites (listed in "Publications" on page 103) or through the associations listed in "Associations" on page 101.

Ameritech Library Systems	See epixtech
Data Research Associates	See SIRSI
Endeavor Information Systems http://www.endinfosys.com	Voyager
EOSi http://eosintl.com	Q Series, T Series, GLAS
Epixtech (formerly Ameritech Library Systems) http:///www.epixtech.com	Dynix, Horizon, Sunrise
Ex Libris http://www.aleph.co.il	Aleph
Geac http://www.library.geac.com	Advance, Plus, Geo-, Vubis
Innovative Interfaces http://www.iii.com	Innopac, Millenium
Sirsi http://www.sirsi.com	Unicorn
The Library Corporation http://TLCDelivers.com	Library.Solution, CARL, Kids Catalog
VTLS http://www.vtls.com	Virtua

Delivery

This is generally software added to the client (Browser in the form of a "plug-in") which supports the playing of the particular data format.

Real Networks http://www.real.com	Real Player G6 Real Audio Real Video	Audio, video delivery and browser plug-ins
Cartesian Inc. http://www.cartesianinc.com	CPC View, Tools	Text compression and delivery server software with delivery to a Browser plug-in

©2003 Sun Microsystems, Inc.

Web Servers

These servers and other programs provide the essential connection between your ILS and the rest of the world on the Web. Until ILSs have built in http capability (and it is doubtful that they ever will—it is not a cost-effective development) you will need a Web server. The other programs are examples of utilities that add either to the Web site creation process or to running it.

Sun http://www.sun.com	Netra™	Server

This is the most widely utilized commercially available Web server product. It has a very comprehensive range of features and is maintained and developed at the forefront of Web standards and functionality by Sun.

Apache Project http://www.apache.org	Apache 1.4	Server

This is the most popular public domain server. It is supported by a co-operative group of its users.

Web Publishing

There are some hundreds of Web page authoring tools available. All of them include some form of publishing capability. Possibly just to upload your completed Web pages to a server, or they may contain a server in their own right.

Web Access

These pieces of software can prevent browsers accessing undesired sites or can positively direct access to sites to create a resource universe for searching.

Symantic www.symanticom/sabu/igear	IGear	Access control, filtering
EduLib www.edulib.com	STOPit	Access control, filtering

Information Retrieval

These search engines may be used on their own or be connected to an Integrated Library System or DBMS to provide a fully searchable collection. All the systems are basically free text search engines—that is they index each and every word in a document. The Web search engines are so named as they are less user controllable in that they do not allow for segment searching (other than possibly "Title") and generally do not provide the advanced search facilities of the free text search engines. They are much easier to use and are more familiar to users, but some times their retrieval performance is terrible.

These systems generally run on UNIX servers.

They usually are accessed via another piece of software (such as a Web server or a Z39.50 server) which makes their answer lists available to the user and also handles the users eventual selection of the material to view. These systems do not store the data themselves; they are only indexing mechanisms.

ILSs often have their indexing built in. If not then they will probably handle all the necessary interaction with the user for the external search engine(s) they support.

Convera http://www.convera.com	Retrievalware (formerly Excalibur)	Free text search engine
Google http://www.google.com	Google	Web search engine
Yahoo http://www.yahoo.com	Personal Yahoo	Web search engine
Excite http://www.excite.com	Excite	Web search engine
Alta Vista http://www.altavista.com/	Alta Vista	Web search engine
Teoma http://www.teoma.com	Teoma	Web search engine
Verity http://www.verity.com	Verity	Free text search engine
Fast www.fast.no	Fast Search and Transfer	Web search engine
Open Text http://opentext.net	Open Text	Free text search engine
Hummingbird http://hummingbird.com	Fulcrum	Free text search engine
CNIDR http://www.cnidr.org	Isite	Web search engine

Web indexer and search engine including Z39.50 access to other engines and as a client to itself from other engines.

University of California http://www.swish-e.org	SWISH-E	Web search engine

A Web site containing details of the major Web search engines and information about search tools you can include within your site is at http://www.searchtools.com. An invaluable place for information about search engines is the Web site at http://www.searchenginewatch.com. There is a free e-zine as well.

A new class of search engine is becoming more available now—the meta search engine. This allows a library to broadcast searches to one or more search engines at the same time. The vast majority are single protocol search engines and are designed to search multiple Web search engines or multiple Z39.50 catalogues. These search engines may have their own indexes or not. They all return results from more than one search engine, but the results may not be combined and usually are not in a consistent format.

Web meta search engines reside on a server on the Internet and act as a service which can be connected to by any (authorized) user with a browser. Personal search engines are software loaded onto the user's personal computer and run from there.

©2003 Sun Microsystems, Inc.

The multi-protocol search engines can search both Web search engines (http) and library catalogues (Z39.50) simultaneously and re-formats and combines the results lists, as well as removing duplicates. They are run as a Web service for libraries and other portals to connect to a number of Sources and provide a service for the portal users.

Meta-crawler www.metacrawler.com	Metacrawler	Web meta search engine
Copernic www.copernic.com	Copernic 2000	Personal meta search engine
MuseGlobal www.museglobal.com	Muse	Multi-protocol search engine
Webfeat www.webfeat.com	Webfeat	Multi-protocol search engine
Innovative Interfaces http://www.iii.com	Metafind	Broadcast searching OPAC
Lib-IT http://www.libit.de	Libero	Meta search catalog

Conversion

These programs are for conversion of scanned text into machine readable encoded text. They are the essential second step before indexing the text with a search engine to make it retrievable.

These programs run either on PCs or on UNIX servers. The PC versions are designed for "one at a time" operation rather than large batches. Batch operation is undertaken on a central server. General error rates of about 0.5%–2% mean that anywhere from 10–40 characters will be wrong per page (2000 characters). Thus, it is usually necessary to manually edit the results and this is best done immediately after the batch conversion while the original page or image is at hand.

Scansoft www.scansoft.com/textbridge/	Text Bridge	OCR

Database Management System

These provide the basic storage and retrieval functions for the rest of the system. All library and information retrieval system have a DBMS underneath them. Sometimes they are standard ones as those listed below, sometimes they are specialized for the particular function. The Relational DBMSs listed below can all be accessed by using SQL (Standard (or Structured) Query Language) so the data is accessible to programs written by the library's own staff or third parties (such as Decision Support System suppliers).

Oracle http://www.oracle.com	Oracle 9i	Relational database
Informix http://www.informix.com	Informix Universal Server	Relational database Multimedia object database
Sybase http://www.sybase.com	Sybase	Relational database

Digital Library/Multimedia Modules

These may be stand alone systems or may be an integrated part of an ILS or DBMS. If they are stand alone they may still only operate with the particular vendor's other software.

Support for other vendors systems usually comes in the form of a protocol interface (Z39.50 or http usually) or as an Application Programming Interface (API) which means that programs will have to be written to interface the two pieces of software. This is not an absolute prohibition, but time and resources must be allowed for in planning the project, or a third party interface tool must be sought.

Sirsi http://www.sirsi.com/sirsiproducts/hyperion.html	Hyperion DMA	Digital media archive
Endeavor www.endinfosys.com	Image server ENCompass	Digital documents Digital library system
Epixtech www.epixtech.com	iLibrary	ASP library modules

Works stand-alone and with other vendors' systems through an external API. This is a multimedia module from an established ILS supplier.

ex Libris http://www.aleph.co.il	digilib	Digital library system
Mediaway http://www.mediaway.com	MediaAsset 2.0	

This system has an asset repository which stores multimedia objects and uses the Verity search engine to allow full text based access to them. Searching is against notes or memos added to the object and includes topic, date or location. Data is stored and transmitted in compressed form for speed.

Project Management

These programs allow projects to be defined and a timeline to be created and then progress to be tracked. They run either on PCs or on servers and most allow more than one user to at least view the project at any time. They allow the project to be re-scheduled when needed and can produce time and resource reports.

Computer Associates http://www.cai.com/products/bas/spj4.htm	Superproject V4.0

©2003 Sun Microsystems, Inc.

Metering

This software measures the traffic on a Web site. Most of them do this by analyzing the log that all Web servers keep of all traffic. Some analyze the traffic in real time and thus allow absolutely up-to-date figures. All have a report generator so that many views of the traffic can be generated and some allow custom reports to be produced. Others connect to browsers for display or to spreadsheets for further analysis.

Web-stat http://www/web-stat.com	Web-stat
Boutell.com http://www.boutell.com/wusage	Wusage
Marketwave Corp. http://marketwave.com	HitList Enterprise Live
Webtrends http://www.webtrends.com	Webtrends

Another method of metering usage is to have the traffic recorded at a more logical level, according to the site or pages visited and to record this as part of an access control and monitoring system. These are usually installed at a gateway, which may be the entrance to the digital library or a service it provides to link users to other library sites or to content providers for electronic distribution of material. The access control software mentioned above will generally have a metering component as will some search engines (meta search engines will record by the individual search engines they use, not specific sites).

Rights Management

This software offers ways of controlling access to content and keeping track of who uses it for what. It also provides ways of protecting the content from theft or misuse.

InterTrust, Inc. http://www.intertrust.com

The InterTrust System Developer's Kit addresses many of these issues and provides the foundation for the creation of practical digital content distribution systems.

DigiMarc http://www.digimarc.com

Provides digital watermarking software for all types of media.

Content Delivery

Electronic delivery of content is becoming more possible and commercially feasible. The technology has become reliable both for delivery and for metering usage for rights management. As well as traditional publishers, who make the full text of their journals available to subscribing libraries or directly to clients, there are special sites which deliver the electronic content of monographs either to computers (usually to be read through a proprietary piece of software) or directly to e-book hardware.

OCLC http://www.netlibrary.com	netLibrary	Electronic books
ebrary http://www.ebrary.com	ebrary	Electronic books and reading tools
Infotrieve http://www.infotrieve.com	Infotrieve	Electronic document delivery

There are a large number of digital libraries and publishers and projects which make content available online. These are often free, but are limited to the material they have in their collection, which is often very specialized and may be limited.

Portals

Portals are a recent phenomenon and range from merely a Web site builder with a number of easily added "widgets," through to industrial strength software which provides everything the current state of technology will allow.

Most are general purpose, though some concentrate on library applications, others on knowledge management, and so on.

Sun http://www.sun.com	Sun™ ONE Portal Server (formerly iPlanet™ Portal Server)	Full function portal system
BEA http://www.bea.com	WebLogic	Full function—business oriented
Museglobal http://www.museglobal.com	Muse	Library oriented
Imanage http://www.imanage.com	WorkSite	Knowledge management oriented
Plumtree http://www.plumtree.com	Collaboration server	Enterprise documentation oriented

©2003 Sun Microsystems, Inc.

Conferences

Conferences come and go by their very nature, but a number of organizations and conferences repeat year after year with some regularity and stability. Below are listed some of those which touch on, or are centered on, digital libraries. A few are very general, but most are specific enough to have relevant papers. Most of the sponsoring organizations have Web site for the conference series, and many of these have past proceedings available in electronic form.

GENERAL LIBRARY	
American Library Association http://www.ala.org	June in Toronto, January in Philadelphia
National Online/Info today	May in New York
International Online http://www.informationtoday.com	December in London
Special Libraries Association http://www.slal.org	Annual, June Los Angeles
World Library Summit http://www.wls.com.sg	May in Singapore
IFLA http://www.ifla.org/IV/ifla68/	July in Glasgow
LIBRARY AUTOMATION	
ELAG http://www.elag.org	April in Bern
DIGITAL LIBRARY	
Joint Conference on Digital Libraries http://www.acm.org/jcdl	July in Portland
International workshop on Digital Libraries (Dlib2002) http://www.ifs.tuwien.ac.at/ifs/events/dlib2001/	September in Munich
China Digital Libraries Conference http://www.nlc.gov.cn/dloc	July in Beijing
European Conference on Research and Advanced Technology for Digital Libraries http://www.ecdl2002.org	September in Rome
All-Russian Scientific Conference http://rcdl2002.jinr.ru/news.html	October in Dubna
International Conference on Asian Digital Libraries http://pipe.cais.ntu.edu.sg:8000/icadl02	December in Singapore
Canadian Digital libraries Symposium http://www.digital-libraries.net	November in Toronto
Libraries in the Digital Age http://www.ffzg.hr/infoz/lida	May in Dubrovnik

Many of the above conference Web sites contain lists of related conferences and other information.

Help

This section contains references to useful resources which don't fit into any of the other categories. Some are resource lists and others are discussion groups, some are "just" sites devoted to digital library matters. They will all repay browsing through them for information and ideas.

Remember also that the journals listed in "Publications" on page 103 contain many articles (many of them referenced in the resources in this section) which discuss the issues surrounding digital libraries.

Digital Library Federation

www.clir.org/diglib/dlfhomepage.htm

Fifteen of the United States' largest research libraries and archives have agreed to co-operate on defining what must be done to bring together—from across the nation and beyond—digitized materials that will be made accessible to students, scholars, and citizens everywhere, and that document the building and dynamics of United States' heritage and culture.

This site, which contains both some resources and discussion papers and committee activities, is now an independent Web site (formerly hosted by the Library of Congress).

Digital Library Resources and Projects

http://lcweb.loc.gov/loc/ndlf/digital.html

One of the pages of the Digital Library Federation (still available through the Library of Congress Web site, if this fails try the DLF immediately above) which links to a large variety of projects and to other resources including papers and software.

Beyond the Beginning: The Global Digital Library

http://www.cni.org/regconfs/1997/ukoln-content/repor~t5.html

This is a report of a very prestigious and useful international conference held in 1997 in London. It contains papers which address many of the issues of digital libraries as well as discussions of experience.

This HTML version is hosted by the site of the Confederation for Networked Information, itself a very useful organization whose aim is to make the interconnection of information systems more practicable.

Berkeley Digital Library SunSITE

sunsite.berkeley.edu

This is a site sponsored by Sun Microsystems™ and hosted and maintained by the University of California at Berkeley Library. It was created as part of their Digital Library project and contains a large number of collections and special library projects as well as information for digital library developers. It is a full digital library and has a number of components, a couple of which are highlighted below.

The site is a gateway to the digital collections of UC Berkeley maintained by departments or the library for special collections. It shows the potential for single access and collaborative organization. The collections are not "unionized" in any way, they are just available through a common access point.

©2003 Sun Microsystems, Inc.

Librarians Index to the Internet

http://sunsite.berkeley.edu/InternetIndex

An index to items on the Internet of interest to librarians. It is not limited to technology or digital libraries. This is, in fact, a Web catalogue of material collected from the Internet.

Current Cites

http://sunsite.berkeley.edu/CurrentCites

A monthly bibliography of articles, books, and electronic documents on information technology. The material is selected and reviewed. Specific "virtual bibliographies" can be created from the site (cite?) for given subjects.

Index Morganagus

http://sunsite.berkeley.edu/~emorgan/morganagus

A full text index to about 70 library related electronic journals.

Image Database Information

http://sunsite.berkeley.edu/Imaging/Databases

This site contains a large amount of useful image information and links to a large number of other image sites. These may contain articles or actual collections of images.

The Clearinghouse of Image Databases and IMAGELIB Listserv Archives

http://dizzy.library.arizona.edu:/images/image_projects.html

The clearinghouse and listserv are hosted by the University of Arizona Library. This is a directory of image collection sites (including films and video). The collections are described in technical detail (scanning methodology , server computer, etc.) as well as their content.

The IMAGELIB is a listserv for all aspects of imaging from the technical to the legal.

Geographical Information Center (GIC)

http://fisher.lib.virginia.edu

Run as part of the University of Virginia Library, the GIC provides a view of some of the future directions of a digital library. As well as a catalogue of its maps and spatial data, the GIC provides an interactive mapper and a references desk online.

The interactive mapper allows users to specify a county map and what features they wish to see on it. The map is then "built to order" for them. The Reference Desk builds lists of links to articles, Web sites and databases which answer (or at least refer to) the users' query.

Both of these are specialist facilities which the library can offer because it is digital. They are indicators of the future of library services in the post-processing of information for the user in an automated environment.

Online Catalogs with "Webbed" Interfaces

http://sunsite.berkeley.edu/libweb

This is a site listing catalogues with World Wide Web interfaces and contains links to other useful sites.

Scholarly Electronic Publishing Bibliography

http://info.lib.uh.edu/sepb/sepb.html

This selective bibliography has over 1600 articles, books and electronic documents about publishing on the Internet and in other electronic media. It is itself an interactive electronic document. It is updated regularly, and is now in its 42nd volume.

UNESCO Libraries Portal

http://www.unesco.org/webworld/portal_bib

This site contains general information for libraries and an extensive list of conferences.

IFLANET Digital Libraries: Resources and Projects

http://www.ifla.org/II/diglib.htm

This site contains information about digital collections (only one at present—add yours?), a bibliography, and another extensive conference list. It also list digital library projects from around the world.

Chapter 8

Future Trends and Research

This is a rapidly moving field where today's developments are gone (and often forgotten) tomorrow. However, it is worth looking at what is being done by researchers at the moment as they will be pointers to the resources and requirements of tomorrow.

There are a number of places where appropriate research or development. or just forward thinking, maybe being done. Universities are an obvious choice, however, even they have to pay for what they do and thus many are part of "research initiatives" or "projects" funded by government agencies. As well as the Library and Information Science Department, the Computer Department or the Communication Department are all candidates for suitable research. Also, look at the individual departments as hosts for the development rather than the perpetrators of it. Thus, an Art History or Music Department may have a digital library since they have the material and the Library Science and Computing Departments are the ones developing the tools and techniques.

Do not ignore the commercial sector. Many of the innovations that you will be able to use come from them. Some are originated there through their own research. Others are developed by them into viable product offerings from university research. At the end of the day it is likely that most of the software you use for your digital library project will come from commercial suppliers. The alternative is to take a research idea and develop it yourself. This is a long and dangerous (setbacks, money, false starts, money, delays, money) route which is only very infrequently crowned with success.

As far as major systems thinking and software is concerned the above are the most likely places for the future to emerge. However, in the area of organization, workflow, marketing, administration and management, the place to look is at those who are practicing. These are areas where incremental "good ideas" and "best practices," rather than developmental breakthroughs,

©2003 Sun Microsystems, Inc.

are the way to the future. Here you can mix and match from what you see from those who are doing. Of course, it will be so much the better if you combine ideas with a unique twist of your own which can start someone else off on the next circuit of the spiral.

Listed in these sections are some of the major research efforts and some of the more interesting library systems.

Digital Libraries Initiative

The Digital Libraries Initiative is funding exercise of various agencies of the U.S. government to promote research into the technologies and implementation strategies underlying future digital libraries.

Phase I was funded by the National Science Foundation (NSF), the Defense Advanced Research Projects Agency (DARPA), and the National Aeronautical and Space Administration (NASA).

Phase II is funded by the National Science Foundation (NSF), the Library of Congress (LoC), the Defense Advanced Research Projects Agency (DARPA), the National Library of Medicine (NLM), The National Aeronautics and Space Administration (NASA), and the National Endowment for the Humanities.

Phase I was started in 1994 and ran for four years. Phase II started in 1999 and will run for a maximum of 5 years. Some of the DLI-2 projects will use and co-operate with Internet2 research projects.

The remainder of this section gives a brief description of each of the projects. The descriptions are all taken from the individual project sites. The description of the aim of Phase II comes from the Library of Congress site and gives some context to both phases as well as the intent of the second phase.

From these descriptions it is clear that the intent is for the first phase to concentrate on the development (or at least investigation) of underlying technologies, and the second phase to look more at applying those technologies (and others) in real life library situations.

Phase I

Six projects were funded and were led by major universities. This section lists the research direction of those projects in descriptions from the individual project Web sites. For more details on each project visit the appropriate Web site, if they still exist, or else look through NSF DLI Web site given below.

The following description of the first phase goals is from the NSF Web site.

> Six research projects developing new technologies for digital libraries—storehouses of information available through the Internet—have been funded through a joint initiative of the National Science Foundation (NSF), the Department of Defense Advanced Research Projects Agency (DARPA), and the National Aeronautics and Space Administration (NASA).
>
> The project's focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks—all in user-friendly ways.
>
> A common strategy in all of these projects is to emphasize research partnerships. We view building partnerships between researchers, applications developers and users as essential to achieving success in generating new knowledge, promoting innovative thinking, and accelerating the technology transfer process.

The initiative will both capitalize on advancements made to date as well as promote research to further develop the tools and technologies needed to make vast amounts of useful information available to large numbers of people with diverse information needs.

Further information can be found at the following Web sites:

NSF Digital Libraries Home Page	www.dli1.nsf.gov www.dli2.nsf.gov
National DLI Synchronization Page	http://dli.grainger.uiuc.edu/national.htm

Many of these projects utilize Sun hardware in a wide variety of configurations. Most of them have undertaken application development utilizing the architectural and programming advantages of Java as their language of choice.

Since these projects are now completed, many of the Web sites will be moving or will disappear as the project information becomes obsolete or other work is started. If this is true for the sites then it is even more true for the contact personnel. Probably half of the people mentioned in the following section will no longer be there. However they will be the authors of reports and papers and, as such, they provide a valuable access point to the literature generated by the projects. Project descriptions were taken from the project Web sites.

University of California, Berkeley

The Environmental Electronic Library

http://elib.cs.berkeley.edu/

Principal Investigator: Robert Wilensky (wilensky@cs.berkeley.edu)

Contact: Charlene Ryan (charlene@cs.berkeley.edu), (510) 642-0930

The UC Berkeley Digital Library project is part of the NSF/ARPA/NASA Digital Library Initiative and part of the California Environmental Resource Evaluation System. Research at Berkeley includes faculty, staff, and students in the Computer Science Division, the School of Information Management and Systems, and the Research Program in Environmental Planning and Geographic Information Systems, as well as participation from government agencies and industrial partners. The project's goal is to develop the technologies for intelligent access to massive, distributed collections of photographs, satellite images, maps, full text documents, and "multivalent" documents.

University of California, Santa Barbara

The Alexandria Project

http://alexandria.ucsb.edu

Principal Investigator: Terrance R. Smith (smithtr@cs.ucsb.edu)

Contact: Mary-Ann Rae (mrae@alexandria.sdc.ucsb.edu), (805) 897-0639

Welcome to the home page of the Alexandria Project. We are a consortium of researchers, developers, and educators, spanning the academic, public, and private sectors, exploring a variety of problems related to a distributed digital library for geographically-referenced information.

Distributed means the library's components may be spread across the Internet, as well as coexisting on a single desktop. Geographically-referenced means that all the objects in the library will be associated with one or more regions ("footprints") on the surface of the Earth.

 ©2003 Sun Microsystems, Inc.

The centerpiece of the Alexandria Project is the Alexandria Digital Library (ADL), an online information system inspired by the Map and Imagery Laboratory (MIL) in the Davidson Library at the University of California, Santa Barbara. The ADL currently provides access over the World Wide Web to a subset of the MIL's holdings, as well as other geographic datasets

Carnegie Mellon University

Informedia

http://www.informedia.cs.cmu.edu

Principal Investigators: Howard Wactler (wactlar@cs.cmu.edu)

Contact: Colleen Everett (cae@cs.cmu.edu), (412) 268-7674

The Informedia Digital Video Library is a research initiative at Carnegie Mellon University funded by the NSF, DARPA, NASA and others that studies how multimedia digital libraries can be established and used. Informedia is building a multimedia library that will consist of over one thousand hours of digital video, audio, images, text and other related materials.

Informedia's digital video library is populated by automatically encoding, segmenting and indexing data. Research in the areas of speech recognition, image understanding and natural language processing supports the automatic preparation of diverse media for full-content and knowledge-based search and retrieval. Informedia is one of six Digital Libraries Initiative projects.

University of Illinois, Urbana Champaign

Federated Repositories of Scientific Literature

http://dli.grainger.uiuc.edu

Principal Investigator: Bruce Schatz(schatz@uiuc.edu)

Contact: Susan Harum (dli@uiuc.edu), (217) 244-8984

The Digital Libraries Initiative (DLI) project at the University of Illinois at Urbana-Champaign is developing the information infrastructure to effectively search technical documents on the Internet. We are constructing a large testbed of scientific literature, evaluating its effectiveness under significant use, and researching enhanced search technology. We are building repositories (organized collections) of indexed multiple-source collections and federating (merging and mapping) them by searching the material via multiple views of a single virtual collection.

Our testbed of Engineering and Physics journals is based in the Grainger Engineering Library. We are placing article files into the digital library on a production basis in Standard Generalized Markup Language (SGML) from engineering and science publishers. The National Center for Supercomputing Applications (NCSA) is developing software for the Internet version in an attempt to make server-side repository search widely available. The Research section of the project is using NCSA supercomputers to compute indexes for new search techniques on large collections, to simulate the future world, and to provide new technology for the Testbed section.

University of Michigan

Intelligent Agents for Information Location

http://www.si.umich.edu/UMDL

Principal Investigator: Daniel Atkins (atkins@umich.edu)

Contact: JoAnne Kerr (jmkerr@umich.edu), (313) 763-6414

Much digital library work has begun from the centralized, structured view of a library and sought to provide access to the library through digital means. In the University of Michigan Digital Library Project (UMDL) we believe that this approach loses the advantages of decentralization

(geographic, administrative), rapid evolution, and flexibility that are hallmarks of the Web. In UMDL, we are instead embracing the open, evolving, decentralized advantages of the Web and introducing computational mechanisms to temper its inherent chaos. However, we are also embracing the traditional values of service, organization, and access that have made libraries powerful intellectual institutions.

The challenges we face are providing an infrastructure that lets patrons (and publishers) feel like they are working within a library, with the traditional emphasis on providing service and organized content, when in fact the underlying space of goods and services is volatile, administratively decentralized, and constantly evolving. Moreover, the decentralized and flexible infrastructure can be exploited to allow information goods and services to evolve in a much more rapid, diverse, and opportunistic way than was ever possible in traditional libraries, for the good of consumers and providers.

In the UMDL we are meeting these challenges by defining and incrementally developing interfaces and infrastructures for users and providers such that intellectual work (finding, creating, and disseminating knowledge) is embedded in a persistent, structured context even though the underlying networked system is evolving. The infrastructure supports extensible ontologies (meta descriptions of collections and services) for allowing components in the digital library to self-organize, dynamically teaming to form structures and services that users need. Principles from economics are also being used to efficiently allocate resources and provide incentives for continual improvement to networked goods and services. This approach enables third parties to join or use UMDL technologies to define and manipulate agents, facilities, and ontologies so that the Web of resources grows in an orderly but decentralized way.

Stanford University

Infobus

http://www-diglib.stanford.edu

Principal Investigator: Hector Garcia-Molina (hector@cs.stanford.edu)

Contact: Marianne Siroker (siroker@cs.stanford.edu), (415) 723-0872

The Stanford Digital Libraries project is one participant in the 4-year, $24 million Digital Library Initiative, started in 1994 and supported by the NSF, DARPA, and NASA. In addition to the ties with the five other universities that are part of the project, Stanford also has a large number of partners. Each university project has a different angle of the total project, with Stanford focusing on interoperability.

Our collection is primarily computing literature. However, we also have a strong focus on networked information sources, meaning that the vast array of topics found on the World Wide Web are accessible through our project as well. At the heart of the project is the testbed running the “InfoBus” protocol, which provides a uniform way to access a variety of services and information sources through “proxies” acting as interpreters between the InfoBus protocol and the native protocol.

With the InfoBus protocol running under the hood, a variety of user level applications provide powerful ways to find information, using cutting-edge user interfaces for direct manipulation or through Agent technology. A second area of focus for the Stanford Digital Library Project is the legal and economic issues of a networked environment.

 ©2003 Sun Microsystems, Inc.

Phase II

The following description of Phase II is taken directly from the Digital Library Initiative Phase 2 Web site at www.dli2.nsf.gov.

Note that it supplies context for Phase I as well and also contains (in the original) a number of links.

> Digital Libraries Initiative Phase Two is a multiagency initiative which seeks to provide leadership in research fundamental to the development of the next generation of digital libraries, to advance the use and usability of globally distributed, networked information resources, and to encourage existing and new communities to focus on innovative applications areas.
>
> Since digital libraries can serve as intellectual infrastructure, this Initiative looks to stimulate partnering arrangements necessary to create next-generation operational systems in such areas as education, engineering and design, earth and space sciences, biosciences, geography, economics, and the arts and humanities. It will address the digital libraries life cycle from information creation, access and use, to archiving and preservation.
>
> Research to gain a better understanding of the long term social, behavioral and economic implications of and effects of new digital libraries capabilities in such areas of human activity as research, education, commerce, defense, health services and recreation is an important part of this initiative.

Projects within phase 2 have been going for nearly three years, and some have already completed. The projects are listed below with the project Web site link for further information (or look at the DLI2 page referenced above). Because of the large number of projects only a one line description of the objective has been given.

The following funded projects are ordered alphabetically by institution.

University of Arizona
High-Performance Digital Library Classification Systems: From Information Retrieval to Knowledge Management
http://ai.bpa.Arizona.edu/go/dl

University of California Berkeley
Re-inventing Scholarly Information Dissemination and Use
http://elib.cs.Berkeley.edu

University of California Davis
A Multimedia Digital Library of Folk Literature
http://philo.ucdavis.edu/SEFARAD

University of California Santa Barbara
Alexandria Digital Earth Prototype
http://www.alexandria.ucsb.edu/adept

Carnegie Mellon University
Informedia-II: Auto-Summarization and Visualization Over Multiple Video Documents and Libraries
http://www.informedia.cs.cmu.edu

Carnegie Mellon University
Simplifying Interactive Layout and Video Editing and Reuse
http://www.cs.cmu.edu/%7Esilver/About%20SILVER

Columbia University
A Patient Care Digital Library: Personalized Search and Summarization over Multimedia Information
http://www.columbia.edu

Cornell University
Project Prism at Cornell University: Information Integrity in Digital Libraries
http://www.prism.cornell.edu

Eckerd College
Digital Analysis and Recognition of Whale Images on a Network (DARWIN)
http://darwin.echerd.deu

Harvard University
An Operational Social Science Digital Data Library
http://www.thedata.org

University of Hawaii at Manoa
Shuhai Wenyan classical Chinese Digital Database and Interactive Internet Worktable
http://www.uhm.hawaii.edu

University of Illinois, Chicago
Digital Library for Human Movement
http://arik.uic.edu/vdsearch.cgi

Indiana University Indianapolis/Bloomington
A Distributed Information Filtering System for Digital Libraries
http://sifter.Indiana.edu

Indiana University
Creating a Digital Music Library
http://dml.indiana.edu

Johns Hopkins University
Digital Workflow Management: The Lester S. Levy Digitized Collection of Sheet Music, Phase Two
http://levysheetmusic.mse.jhu.edu

University of Kentucky
The Digital Atheneum: New Techniques for Restoring, Searching, and Editing Humanities Collections
http://www.digitalatheneum.org

©2003 Sun Microsystems, Inc.

Universuty of Massachusetts, Amherst
Word Spotting: Indexing handwritten manuscripts
http://ciir.umass.edu/wordspotting

Michigan State University
Founding a National Gallery of the Spoken Word
http://www.ngsw.org

Oregon Health Sciences University
Oregon Graduate Institute of Science and Technology—Tracking Footprints through an Information Space: Leveraging the Document Selections of Expert Problem Solvers
http://www.cse.ogi.edu/dot/research/footprints

University of Pennsylvania
Data Provenance
http://db.cis.upenn.edu/research/provenance.htm

University of South Carolina
A Software and Data Library for Experiments, Simulations, and Archiving
http://econ.badm.sc.edu/beam

Stanford University
Stanford Interlib Technologies
http://www-diglib.Stanford.edu

Stanford University
The Stanford Encyclopedia of Philosophy
http://plato.stanford.edu

Stanford University
Image Filtering for Secure Distribution of Medical Information
http://www-db.stanford.edu/pub/gio/TIHI/TID.htm

University of Texas at Austin
A Digital Library of Vertebrate Morphology, Using High-Resolution X-ray CT
http://www-ctlab.geo.utexas.edu/dmg

Tufts University
A Digital Library for the Humanities
http://www.perseus.tufts.edu

University of Washington
Automatic Reference Librarians for the World Wide Web
http://www.cs.washington.edu/research/diglib

The following projects with undergraduate emphasis are ordered alphabetically by institution.

University of California Berkeley
Using the National Engineering Education Delivery System as the Foundation for Building a Test-Bed Digital Library for Science, Mathematics, Engineering and Technology Education
http://www.needs.org

Columbia University
Columbia Earthscape: A Model for a Sustainable Online Educational Resource in Earth Sciences
https://wwwc.cc.columbia.edu/earthscape

Georgia State University
Research on a Digital Library for Graphics and Visualization Education
http://canute.cs.gsu.edu/secdl

University of Maryland
Digital Libraries for Children: Computational Tools that Support Children as Researchers
http://www.cs.umd.edu/hcil/kiddiglib

University of North Carolina, Wilmington
A Digital Library of Reusable Science and Math Resources for Undergraduate Education
http://www.uncwil.edu/nccl.htm

Old Dominion University
Planning Grant for the Use of Digital Libraries in Undergraduate Learning in Science
http://dlib.cs.odu.edu

Swarthmore College
The JOMA applet project: Applet support for the Undergraduate mathematics Curriculum
http://www.mathforum.org/joma_applet.htm

University of Texas at Austin
Virtual Skeletons in Three Dimensions: The Digital Library as a Platform for Studying Anatomical Form and Function
http://uts.cc.utexas.edu/%7Evskel

International Collaborative Projects

In addition to the domestic DLI2 projects there is an international co-operative series of projects. All these can be accessed through the NSF DLI2 Web site.

- NSF-JISC (US-UK) International Digital Libraries Collaborative Research and Application Testbeds (NSF02-085)
- NSF-DFG (US-Germany) International Digital Libraries Research
- DELOS/NSF Working Group—Reference Models for Digital Libraries: Actors and Roles
- NSF/EU Digital Libraries: Future Directions for a European Research Programme

©2003 Sun Microsystems, Inc.

As well as these collaborative frameworks, a number of the DLI2 projects are international in scope and involve co-operation between US academic institutions and those from the UK, Australia, Germany, Africa, Japan.

European Projects

The European Union has funded a number of research and "proof of concept" projects through its various Telematics projects. Its funding has tended to be rather more widely spread both in time and in the number of projects supported, than the US DLI. Projects such ONE to provide a single European wide virtual union catalogue, and MALVINE to provide a similar facility for libraries and archives, are multi-organization collaborations by their nature. EU funding also requires a multi-national make-up for its projects.

Its projects have ranged from infrastructure to user studies and collaborative delivery (as in the two mentioned above).

Currently European research is funded through the remainder of the Fourth Framework programme and the beginning of the Fifth framework programme. These are co-ordinated in an IST—Information Sciences and Technologies—initiative and information about the whole operation can viewed at the cordis site at http://www.cordis.lu .

"Working" Library Systems

This section lists a selection of library sites and projects. It is not comprehensive and will be continually out of date. Use the sites listed in "Help" on page 115 to view lists of libraries on the Web and browse them, or look at http://www.edulib.com for the online version of this document where a more complete and up-to-date list is kept.

Remember that many "general" Web sites will show innovative interface and presentation features and a number such as the news services (BBC, CNN, ABC, DW, etc.) and the newspapers (times, nytimes, tribune, ft, etc.) offer extensive searching of their back issues and archives. Most of them offer images and some video (particularly the TV services) for display. Even the Web search engines offer a view of the current state of the art for general Web searching. These are the systems most users are familiar with and anything which differs radically may meet with resistance in a mass market.

The descriptions of the projects or library sites primarily come from the sites themselves.

Many of the sites contain references to other sites and thus the circle may be completed. The section on "resources—help" (see page 120, "Help," in Chapter 7) contains URLs for some sites which themselves contain lists of digital libraries among other useful places to visit.

American Memory

http://lcweb2.loc.gov/ammem

A Library of Congress project storing digital versions of documents, photographs, movies, and sound recordings that tell America's story. This site does have audio and video digital recordings which can be downloaded and played. Some of the audio can be streamed (played across the Internet).

Making of America

http://cdl.cornell.edu/moa

Materials accessible here are Cornell University's contributions to Making of America (MOA), a digital library of primary sources in American social history from the antebellum period through reconstruction. The collection is particularly strong in the subject areas of education, psychology, American history, sociology, religion, and science and technology. This site provides access to 267 monographs and over 100,000 journal articles with 19th century imprints. The project represents a major collaborative endeavor in preservation and electronic access to historical texts.

The Making of America collection is comprised of the digitized pages of books and journals. This system allows you to view scanned images of the actual pages of the 19th century texts. Optical Character Recognition (OCR) has been performed on the images to enhance searching and accessing the texts.

America is made possible by a grant from The Andrew W. Mellon Foundation.

Current online holdings: Pages: 907,750; Monographs: 267; Serial Volumes: 955

Cline Library

http://dizzy.library.arizona.edu

A Web-accessible image database, set up as a number of exhibits which link images into documentary text. This collection is hosted by Northern Arizona University.

Corbis Image Catalog

http://www.corbis.com

A commercial site offering many of the most famous images in the world. The broad subject (or thematic) categories lead to pages containing multiple thumbnails which in turn lead to the "full size" image. There are no descriptions beyond title and attribution.

Since this is a commercial site and the images are for sale, it is interesting to compare the "terms and conditions" for these images with those for the university and national library sites.

UC Berkeley Earth Sciences Library Digital Map Collection

http://www.lib.berkeley.edu/EART/digital/tour.html

A tour through the collections of this map library. It contains maps and map fragments of various kinds, including general, topographic, thematic, facsimile, and nautical charts. The tour is an interesting introduction to the topic for casual users and leads to many resources.

National Library of Australia Pictorial Collection

http://www.nla.gov.au/catalogue/pictures

Contains historical images. The bibliographic record contains a thumbnail image which links to the "full size" (medium resolution) image. Searchable in a variety of ways.

Tokyo University Digital Museum

http://www.um.u-tokyo.ac.jp

Contains descriptive records with thumbnails and full size images. This site is in Japanese and loses something if your Browser does not support Japanese characters (ISO-2022-jp is used). However, some of the images are worth just clicking at random to find.

©2003 Sun Microsystems, Inc.

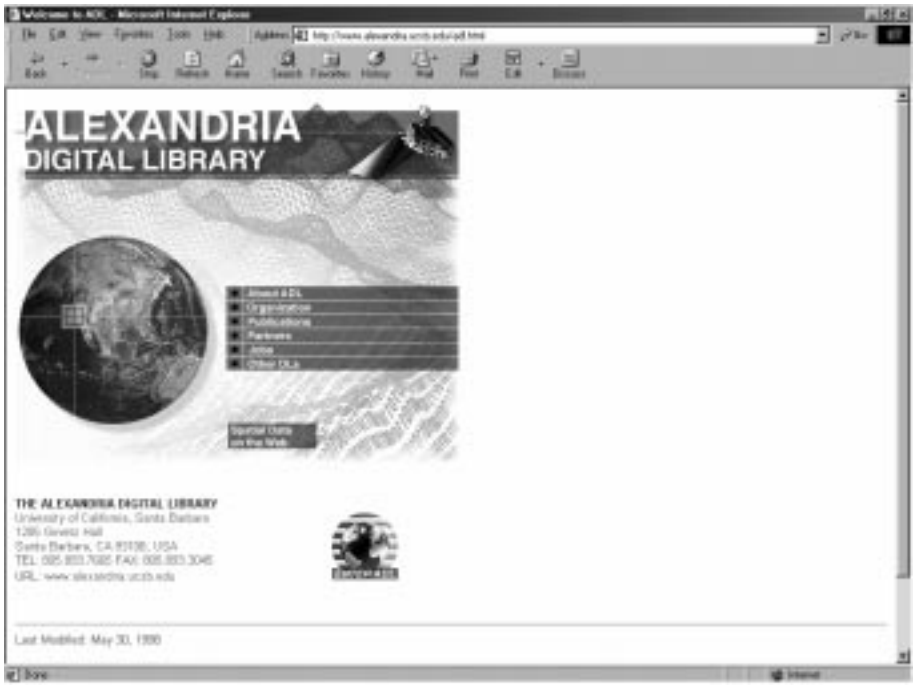
Some Sites in Pictures

Some of the above sites and some others captured for a presentation of this work at the VALA 2000 (www.vala.org.au) conference in Melbourne, Australia.

Images of these slides are presented below. The originals can be obtained from the author (email: peter.noerr@edilub.com) or from the VALA site above.

http://www.acm.org/dl/

ACM's digital library is a typical text and digitized pages library. It is 'closed' (fee paying), and a small 'corporate' collection.

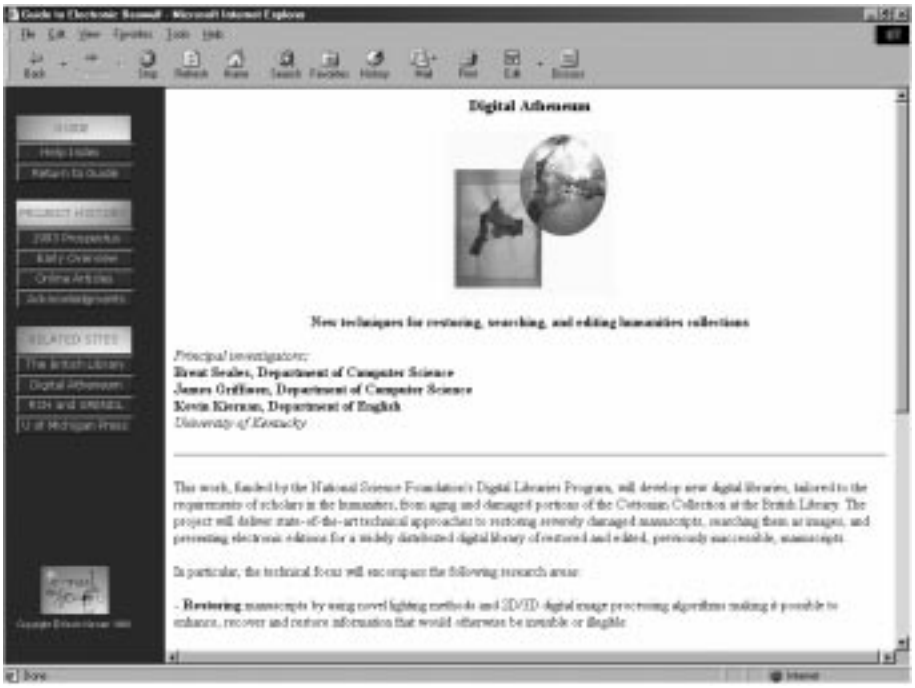


http://www.alexandria.ucsb.edu/adl.html

A project from DLI-1 with geospatial data and a useful 'workspace' and specialized searching.

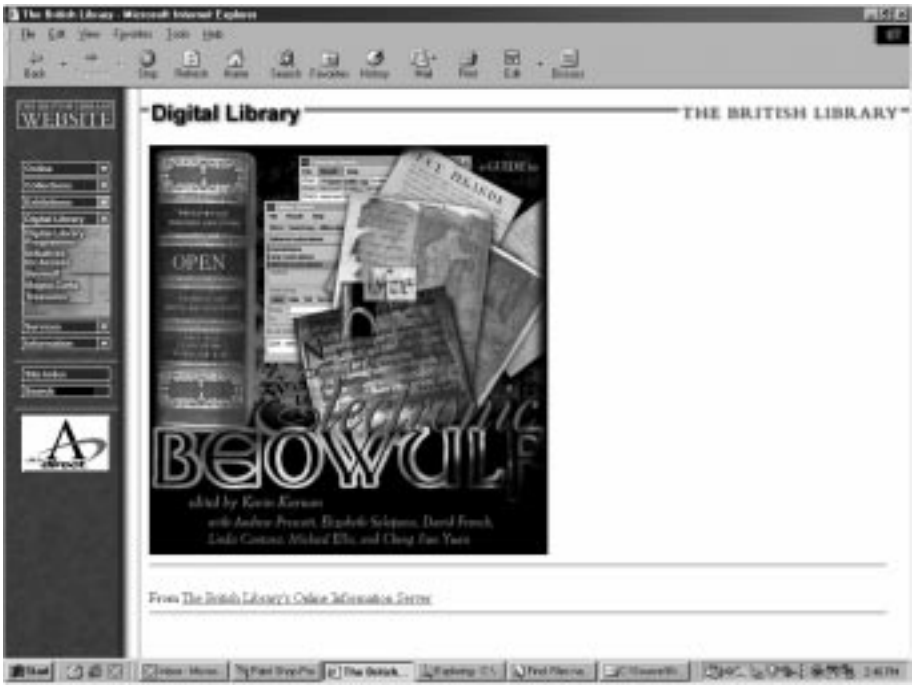
http://www.digitalatheneum.org/

A project digital library concerned with the preservation of the original material. Thus it is likely to concentrate on physical description (missing pieces, etc.) more than most and have a wealth of unusual images—not just the digitized 'pages'.



http://www.bl.uk/

A long term site which is actually not a digital library at all—it is additional information and help for the single document that is Beowulf—the CD-ROM.



©2003 Sun Microsystems, Inc.

http://www.melvyl.ucop.edu

A very typical digital library where that is defined as “a web accessible library”. It is the web OPAC to the University’s traditional catalog.

http://www.clir.org/diglib/dlfhomepage.htm

Not really digital library examples at all, but a few useful resources to use for information of a general nature about the advancement of digital libraries and research.

http://mhtml.ulis.ac.jp

A really unique digital library of multi-lingual renditions of Japanese folk tales. It is slowly growing and they are asking for volunteers to help translate.

http://imagelib.ncsa.uiuc.edu/imagelib.html

No digital library collection would be complete without pictures from NASA. This has them but, if you are prepared to wait you can also download 3-D simulations and VR walk-throughs of galaxies.

©2003 Sun Microsystems, Inc.

http://www.nla.gov.au/images1

A long lasting digital library. It is well organized by collection, even if some of the 'collections' are quite small. The bird paintings are beautiful.

http://www.digital.nypl.org/

New York Public has a quite extensive set of images in its collection. They are also collection oriented and searching is through this route, more browsing than searching.

http://timeframes1.natlib.govt.nz/

Two from New Zealand. Timeframes showing historical documents in a variety of collection contexts.

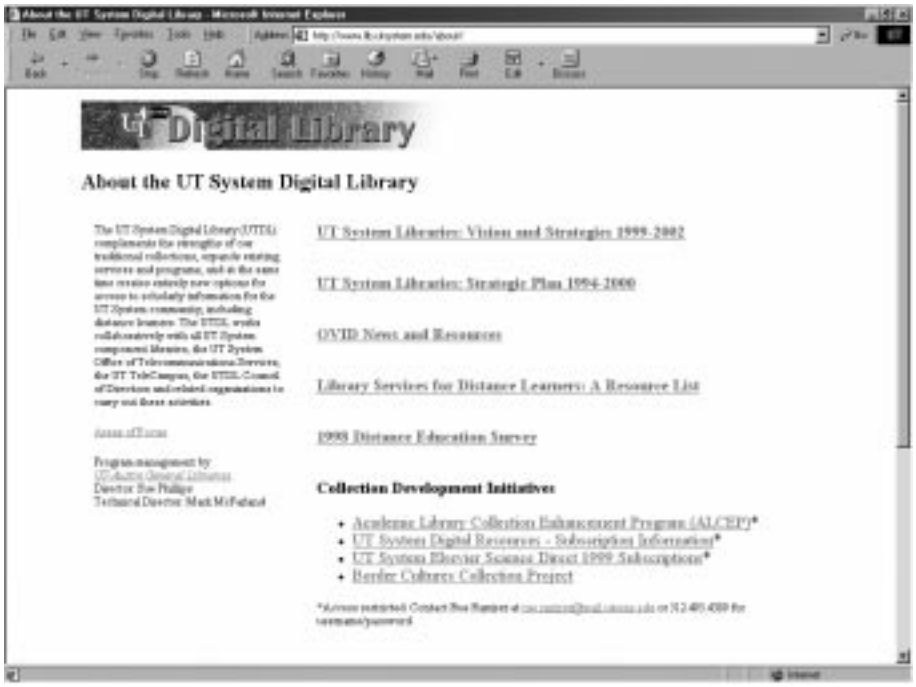
http://www.nzdl.org/fast-cgi-bin/library

The NZ digital library project is one of the most eclectic around and its collections are thematic and cover a range of topics and sources. It includes news service offerings as well as static material collections. It delivers full text documents.

©2003 Sun Microsystems, Inc.

http://www.trumanlibrary.org/

A thematic digital library (within a thematic library) with material of diverse types. It includes audio and some video as well as text and image.

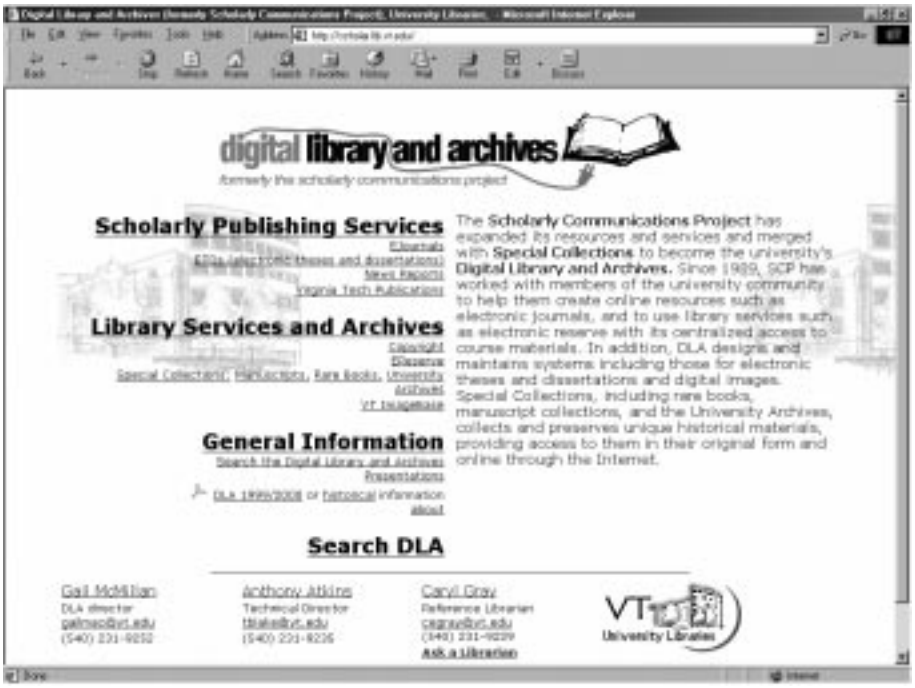


http://www.lib.utsystem.edu/about/

A very standard digital library (read 'web accessible library') with a dedicated audience in mind. It is designed for the university and is primarily an effective delivery mechanism for them.

http://scholar.lib.vt.edu/

Virginia Tech gets two examples as they have a good electronic publishing digital library which provides full documents directly oriented to student use.



http://scholar.lib.vt.edu/imagebase/

Virginia Tech also has a very nice clean image library with simple and advanced keyword searching (as are all image libraries) and very thorough descriptions of the images as objects in their own right.

©2003 Sun Microsystems, Inc.

Muse is not a digital library, but it is the sort of tool which may become very important for their exploitation. It is a multi-protocol meta-search engine which allows searching through web search engines and the library's main catalogue (Z39.50) engine simultaneously. It provides a method of dynamically combining digital libraries for user searching.

It was used to collect the examples used here.

Nice Things Just Around the Corner

These are topics that will become important for libraries and digital libraries in the future. Some will arrive very soon, some are longer term, and some are just things that I think are (or should be) important and should be addressed.

Document Management Systems

The worlds of document description (libraries) and document management (filing systems, version control, routing and recording, warehousing and review) are slowly drawing closer. The digital library is very close to a document management system. The near future will see these systems draw even closer and document management functions (such as co-operative creation, file tracking, version control, etc.) will become adjunct functions to the descriptive capabilities of the library. The management functions will complement (and replace some of) the circulation modules of today's ILS.

Third Edition Update

Its not here yet, but is getting closer. The advent of portals and the move towards integrated EIP (Enterprise Information Portals) and KM (Knowledge Management—which is basically the document management system described above by a fancy name) means that the corporate world is starting to see this happening before the library world.

Multimedia Systems

More and more media types are being added to today's libraries and they need different handling from the print material. This is specially true if the material is in digital form. The handling and processing of these will become a special module of an ILS just like any other and the idea of a "multimedia" library as being anything special will disappear—all libraries will be multimedia.

The libraries will have access to remote resources to complement their own. Thus a library could "rent space" on a video-on-demand server to allow display of its video material while still providing the catalogue searching functions and controlling all access itself.

Third Edition Update

This is here in some libraries now. Many libraries offer access to multiple material types through their conventional OPACs. More ILS vendors are adding "multimedia" capabilities and almost all are expecting to co-operate with delivery services for Web based delivery. The Library Web portals allow basic catalog records to be enhanced with images from off site vendors, and also allow direct document and e-book delivery.

Metadata

The concept of metadata has been around since the first catalogue. It is only now with the advent of the Internet and the amount of material being published on it that the problems are becoming urgent. Metadata (the description of how an object it described, i.e. a list of the attributes used to describe it) is moving in two opposite directions.

It is being simplified (as in the Dublin Core) so that a "minimum" standard exists which can be applied to as many objects as possible and which stands some chance of being applied economically and reasonable consistently. This makes more material available and allows the prospect of automatic extraction of the descriptive information from either HTML "metatags," or XML tags, or Dublin Core elements.

It is being enhanced to provide more comprehensive descriptions. The advent of new material types (mostly digital) are requiring new attributes for both technical reasons and descriptive. Attempts to search across multiple databases hinge on the availability of accurate metadata about the content of those databases. Further enhancement of this to allow more resources to be included in searches at a greater level of abstraction requires a second level of metadata. The descriptive pyramids will continue to grow as more and more powerful ways of combining the data are developed, and they need to know more details of the similarities and differences of the underlying data, information, or knowledge.

Third Edition Update

Another, nearer, but not there yet. XML structured MARC records are around, but are not routinely used. Dublin Core is used by many search engines in the information field. A number of the DLI2 projects are looking at this very problem, but it will be a while until their research findings are translated into products. Meanwhile systems such as the author's Muse have been produced to allow various ILS and other publishing and information systems to "talk" to each other so the practical benefits of interoperability are available before we have a common standard.

 ©2003 Sun Microsystems, Inc.

Distributed Databases, Systems, Libraries, Services

The trend to distribution is well established and with the increasing number of "libraries without walls" there is a strong need for it. Databases will be located (logically) closer to the creators of the information and the libraries will become processing centers drawing on a variety of resources to provide their answers.

The universal availability of libraries through the Internet means that users will chose a "favorite" and use it wherever it is located as long as it provides the services and resources they need. Libraries may become distributed "virtual" entities which are the logical extension of today's consortia in providing a single access point to the aggregated resources of the member libraries and their partners.

At the most technical level there are products emerging today (such as Sun's Intelligent Storage Network) which will enable databases and portions of databases to be clustered and partitioned as need be for the different services the library wishes to offer. Here we are considering the distribution at a level below the searching application so that, unlike today where a number of Z39.50 applications have to communicate to get a consolidated answer, the single application of the library can query physically and logically distributed databases as if they were its own single database.

This leads to more controlled operation of the application and better performance since the distribution is where it is needed, not where it is possible. The libraries will become processing nodes dedicated to creating specialist information of a high quality. They will be providing specialized (or general) services to a particular market.

Third Edition Update

This is an area whose time has come. The meta search engines and portals now coming into production allow for virtual libraries of almost any size. The Internet allows users to search all over the world if they so desire. Single Search Interfaces are now being requested by the biggest libraries and smaller ones will follow to allow their patrons to more easily search the world.

Better Bibliographic Models

The current bibliographic models are coming under strain from the advent of new media objects and the new ways they can be related or changed. These current models are designed for static and long life material with relatively simple and fixed features.

New material is becoming very time sensitive and is related in many more ways than before. It is also becoming important to enable relationships to be inferred, created and recorded so that the user is provided with a more rounded set of results to their query.

The recent IFLA sponsored work on a more theoretically satisfying bibliographic model, which culminated in the Book "Functional Requirements for the Bibliographic Record" (see the IFAL site www.ifla.org for details) provided a major advance. It should be capitalized on and systems should, at least, move towards the structures it recommends.

Third Edition Update

The FRBR model is here and is widely recognized as a theoretical improvement over previous methods of dealing with bibliographical material. However there is no clamor for it outside academic circles, and so the major ILS vendors are not putting any efforts into producing systems which support it, despite the better retrieval it would offer to users. A case of economic need being the driving force.

Active Clients

Most current library applications are client server architecture with a Web Browser interface added. This will change with the more widespread use of components and particularly of Java. The beauty of Java is its Virtual Machine (the JVM) which allows the same code (program) to run on any machine. This means that a processing program can be written once and then made to operate on all the users' individual computers. This allows a controlled way for the systems designer (the vendor) and the systems librarian (in the library) to create a situation where processing is moved to the user's computer in a controlled manner.

This is the premise behind the network computer, such as Sun's SunRay. The application will be downloaded and run on it. Thus, the Browser interface will disappear and Java functions will be used to give all the functionality and control of a normal PC interface (such as a Windows program), but operating over the Internet. Browsers will exist as the lowest common denominator means of communicating to sites (or applications) that have no active clients to provide the extra functionality and local user control.

Java will also become increasingly used on the server side of the total system to allow for different databases and search and processing servers to be tied together by common processing software communicating through the Internet. As well as the flexibility of this evolution, there will be the improvement in performance which will come from the three tiers of the total application interacting and generating results and displays where they are needed. After all, it is faster for almost any PC to generate a nice screen layout from a stream of text, than to have the layout generated at the server side and sent in formatted form. It also helps conserve that scarce commodity—network bandwidth.

Third Edition Update

Java is certainly making serious inroads into the processing of bibliographic data. However it is residing on the server, not on the user's PC. Recent security concerns and worries about performance have lead to a virtual block on downloaded software in general situations. Management of the software has proved difficult and expensive.

Processing is being devolved however, but to distributed servers situated all round the Internet or a campus LAN, rather than to the user's workstation. The intriguing advent of peer-to-peer computing (where a number of similar computers join to share the processing of a task) means that it is quite likely we will see co-operative load sharing between libraries in the near future. Whether this is through a formal Grid system, or as a more informal arrangement remains to be seen.

Integrating IR and MM into the ILS

As stated above the systems will become more modular and it will be possible for the library to choose the IR or MM components of their ILS. Different libraries have different search and material handling requirements and that flexibility will become part of future systems.

The whole idea of Application Service Providers is one which would fit very well in the library market. It should be a direction to move in the near future.

©2003 Sun Microsystems, Inc.

Third Edition Update

ASP has made no really big inroads into library computing. The concept of buying individual processing modules still remains a dream of most library managers, and a nightmare of most systems librarians. However there is slow progress. Some systems environments exist which will allow mix-and-match of ILSs, but the individual components are not there, so no savings.

E-commerce

Even though libraries may not be commercial entities and charge users for the content they supply, they will have to operate in the world of electronic commerce. Users will be used to online service provision from commercial sites and will expect it from their library. They will expect to be able to pay fines electronically—or even automatically—and to order services and even receive goods over the Internet.

The library's suppliers will similarly become more electronic and the purchase of goods and services from journal subscriptions to paper clips will become more network-oriented. The library's ILS will have to be capable of keeping up.

This electronic trade will lead to more commercial transactions being performed by two programs talking to each other, with no human intervention except in the case of problems. This could lead in one of two directions. The library and its suppliers could have their systems become more closely connected and provide a more efficient and tailored service. The alternative is that virtually every purchase could become an electronic auction in a search for the best service and lowest price. Look out, the electronic agents are coming!

Third Edition Update

E-commerce is making grounds. Recent infrastructure systems allow libraries to interact with vendors of electronic material over a network, and recent virtual systems (such the author's Muse) allow for very full real time inquiry and order and request placement. Not here yet, but systems and standards are moving towards it.

Components and CORBA, Etc.

ILSs will become more componentized as pieces of software. They are already some of the more complex of programs. Future development will become more complex and costly as users demand support for more material types, more functionality and more methods of interaction and delivery.

The only way to undertake these developments will be to split them into components and undertake development and upgrade one component at a time. However, the transition to a component capable system is not an easy one and will take time and resources.

Fortunately, this work will be aided by languages such as Java that allow a component to be re-used wherever it is needed.

Third Edition Update

No real progress here. The major ILS vendors are sticking to traditional software models. It will take a few small companies to make things move, and then watch out.

Bandwidth

New technologies and research projects (such as Internet-2 and the Next Generation Internet (NGI)) mean that bandwidth will be increasing. User side technology improvements (such as xDSL—x Digital Subscriber Line—and Gigabit) will mean that common telephone connections will become orders of magnitude faster.

Alternative delivery mechanisms such as cable and satellite will also offer bandwidth improvements. All of this means both bandwidth and Quality of Service (QoS) will improve and more information can be sent to users faster and more reliably. These improvements will be felt first at academic institutions and then in the commercial and individual world. However, the pace of change is such that the time lag is likely to be months rather than years.

One good effect of this headlong pace of communications development is that countries that do not have an existing infrastructure can leapfrog and acquire the latest, often for a cost of less than yesterday's technology.

Third Edition Update

There is plenty of bandwidth in parts of the world and almost none in others. But connections and bandwidth are growing rapidly everywhere. Even the US has some areas with poor connections, and its competitive environment means often that existing infrastructure is not being well used.

Searching

Catalogue searching is a major function and, in the last two decades of library automation, it has changed considerably. Currently a catalogue should allow the user to either directly search for items by an user entered question, or to browse pre-compiled lists of useful access points as a way to find what they want.

Direct searching generally involves a Boolean search query (which is often disguised behind graphical input screens) which is applied to either specially indexed parts of the bibliographic record or to the keywords extracted from the full text of a document. Non-textual material is handled by creating a text record (a surrogate) for it and then indexing that record. Some systems allow searching of non-text objects by their features, but this is currently a specialized operation.

The systems usually return a list of potential answers (hits) to the user for consideration and often will allow "more like this" searches to be started, or the original searches to be refined or expanded to obtain a reasonable number of hits (reasonable is defined by each user for themselves).

The systems generally do not attempt to rank the results and when they do—as in Internet search engines—the results are so bizarre that users generally have little confidence in them. Research is being conducted into more meaningful interaction between users and the search systems in many universities, but it has not found its way into the mainstream ILS products.

The search advances will be incorporated into the ILS (possibly as optional modules) and specialized searching (for images, sound, video clips, etc.) and thematic or conceptual searching will become available.

New forms of user interaction will be provided with some truly graphical interfaces to better visualize and manipulate the information. Users will get the ability to customize the interfaces to their requirements.

Meta search engines will become more important as the number of libraries accessible via the Internet grows. They will provide convenient access points to an individual library and to the wider world of other libraries, Web sites, content providers, commercial indexing services and such like to provide a "one stop" information access through the library.

©2003 Sun Microsystems, Inc.

An area still to be adequately addressed is the searching of the multimedia material which is a digital library. Reducing searching to a textual surrogate record is less than satisfactory, but there is little alternative at the moment. Certainly there is little support for direct image or audio searching within the existing ILS systems. It is an area with a fair amount of research work, but few available results.

Third Edition Update

Meta search engines are now starting to become an accepted part of the searching scene in libraries. One stop searching is being required in ILS bids. More processing of the returned results is here, and enrichment and onward linking from results much improves the information returned.

However the era of natural language questions and actual answers (rather then lists of possibilities) is still some ways off.

Portals

With the expanding Internet and, particularly, the increasing availability of high bandwidth connections, the age of portals is upon us. Portals are merely aggregations of services on one Web site so the user does not have to hunt in many different sites for the things s/he wants whether these things are information or games or shopping. Portals are really department stores or shopping malls on the Web.

Portals are already becoming specialized and a serious commercial operation. One of the most defined niches is the educational portal where the portal service provides all the computing services a university or school will need via its Web site. Each university using the service has individualized access to the functions and features from student administration to the development and delivery of remote learning courses. This sort of service portal is not yet available in the information world, but it cannot be far away.

Libraries need to establish their information portals with a broad spectrum of information providing services so they can retain their position as the one stop shop for information in their community. If they don't do this then the community portals (geocities, etc.) will provide local information, the business portals will take that role, the direct delivery shops and sites (amazon.com and netlibrary.com) will supply the reading material and the library will be left as a series of meeting rooms and an archive.

Portals may possibly represent the biggest threat to libraries in the medium and long term, but they also offer an opportunity for libraries to regain a position they haven't held since possibly the middle ages.

Third Edition Update

Portals are being embraced by libraries and they may be their salvation in the battle against the Web search engines as the place to look for answers. The libraries' online presence is fighting back with the portals and meta search engines and enriched content and additional services, all from a source the users trust.

Summary

This paper considers in some depth the issues surrounding the creation of a digital library. The initial part (Chapters 1 and 2) raises a large number of questions which have to be addressed before the decision to proceed with the creation of a digital library is taken. The second part (Chapters 3–6) contains a discussion of the decisions, the management, the techniques, and the methods that need to be understood and used in the design, creation, implementation and maintenance of a digital library. The last part (Chapters 7–8) contains lists of references and resources (and a glimpse at what others are doing and are planning) that will help make the vision of a digital library a reality.

The paper addresses important issues and directs the reader to consider them and decide on their unique solution. There are numerous resources and further reading for all these issues to assist the reader in obtaining a well-rounded view of the issue and to show what others have done, and what the current thinking is.

Many sections enumerate and detail the techniques needed to build the digital library. Step-by-step procedures and calculations guide the novice reader through the stages of planning, designing, and building the digital library.

The resources section is, perhaps, the most important in the whole paper. It provides an invaluable reference to the tools and equipment that are available to help in all aspects of dealing with a digital library. As well as associations and publications dealing directly and indirectly with the issues, it contains lists of vendors of the software and hardware that forms a necessary foundation for creating and running a digital library.

The vendors are vital both for present offerings and for future potential. They are the partners who will make the digital library possible. Taking an overview of all the offerings it becomes clear that there are companies strong in individual specializations, as would be expected, but there are very few who can offer a complete range of products and services from one source. These may not all be internal offerings. But a wide alliance of partnerships is a strength in itself.

Digital libraries are here to stay. However, their form will evolve rapidly as the external world evolves and offers opportunities and makes demands. It is vital that the partner(s) you choose for your digital library project will stay with you for the long haul. They must be capable of offering and guiding the changes your digital library will want, and have, to go through.

The future will be with us tomorrow and a digital library needs to be prepared for it more than most. Here good basic design with an eye on the immediate and longer term future is the foundation. The rest of the structure must be built using tools and techniques which provide solid reliable operation with the flexibility to change and adapt and innovate in the future. Good basic design is language independent and depends on the expertise and innovative flexibility of the designers. However, being able to design with the knowledge that the structures can be built is important, even vital. What is needed is a universal and secure environment for programmers so that what they produce does not interfere with other programs. This limits the chance for unauthorized access (commonly called “hacking”) as well as preventing programs crashing each other. It provides an architecture with the promise of running on almost all computers and promotes a design which allows individual components to work together at the user’s desk and at the various servers in the (possibly virtual) digital library. The future will see the demise of the “all function encompassing” monolithic systems in favor of aggregates of compatible interworking parts from which the purchaser will be able to “mix-and-match” a system with the exact functionality they need.

Building a digital library is a big undertaking, but it is a very exciting one and one that will ensure a library a place in the hearts and minds of the users of the 21st Century.

©2003 Sun Microsystems, Inc.

About the Author

Dr. Peter Noerr has degrees in Physics and Computer Science and a PhD in Information Science from The City University, London. In the 1970s he worked for The British Library for 6 years, four of them as Head of Systems Development. He then spent three years consulting for academic, national library and IGO clients in all parts of the world. During this time he started his one library automation systems company (Information Management and Engineering Ltd. (IME)) in the UK.

This company produced the Tinlib/The Information Navigator software and sold it through distributors in 25 countries to over 2,750 customers in 38 countries. The company ranked in the top three in either special libraries or academic libraries in the annual Library Journal survey from 1994 onwards.

Just as important as the company's market success, was the advanced design of the software. It utilized an entity-relationship database for accurate bibliographic modeling. It introduced a multi-user DOS based system in 1985, incorporating a client/server architecture. In 1987 the system was migrated to UNIX and it incorporated Unicode support as early as 1990. The system operated in an isolated environment (very similar to a Java VM) allowing portability across multiple operating systems. The entire functionality of the ILS was controlled via profiles, and the system introduced browsing and inter-record navigation long before the Web was available.

Dr. Noerr was the chief systems architect of this and the company's other products. He created systems designs for all levels of library from small special libraries to national infrastructure plans for government ministries.

In 1996 IME was sold and merged with Data Trek to form EOSi. Dr. Noerr stayed with the company through a transition period and left in August 1997. Dr. Noerr is currently employed by MuseGlobal, a company he co-founded, as Chief Technical Officer. He is engaged in consultancy and system design work in the areas of digital libraries, multiple character set handling, information modeling, search system and interfaces, and systems architecture and design.

MuseGlobal currently produces and markets its multi-protocol meta search environment for the Web (Muse) through a number of library and other partners and directly.

He may be contacted through the company at peter.noerr@museglobal.com or www.museglobal.com.

SUN™ Copyright 2003 Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, California 95054 U.S.A. All rights reserved.

RESTRICTED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (c)(1)(ii) of the Rights in Technical Data and Computer Software clause at DFARS 252.227-7013 and FAR 52.227-19.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

TRADEMARKS

Sun, Sun Microsystems, the Sun logo, Solaris, SunPlex, Sun Enterprise, Sun Management Center, Java, Java 2 Enterprise Edition, J2EE, Sun StorEdge, SunSITE, StarOffice, Sun Ray, Intelligent Storage Network, iPlanet, and Sun Fire are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and or other countries.

All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the United States and other countries. Products bearing SPARC trademarks are based upon an architecture developed by Sun Microsystems, Inc.

UNIX is a registered trademark in the United States and other countries, exclusively licensed through X/Open Company, Ltd.

WebCT, WebCT Campus Edition, and WebCT Vista are trademarks of WebCT, Inc.

THIS PUBLICATION IS PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NON-INFRINGEMENT.

THIS PUBLICATION COULD INCLUDE TECHNICAL INACCURACIES OR TYPOGRAPHICAL ERRORS. CHANGES ARE PERIODICALLY ADDED TO THE INFORMATION HEREIN; THESE CHANGES WILL BE INCORPORATED IN NEW EDITIONS OF THE PUBLICATION. SUN MICROSYSTEMS, INC. MAY MAKE IMPROVEMENTS AND/OR CHANGES IN THE PRODUCT(S) AND/OR THE PROGRAM(S) DESCRIBED IN THIS PUBLICATION AT ANY TIME.

COPYRIGHT AND IPRS

Throughout this paper many trade names, trademarks, registered trademarks, service marks, registered service marks, and other names or phrases belonging to a particular corporate entity are used. This is particularly true in the resources section. All of these names, marks, etc. are recognized as belonging to their respective owners.

For the sake of clarity of presentation and reading copyright and/or trademark and other symbols have not been placed within the text to mark these. The reader is reminded that if any sections of the paper are used in other contexts, then an acknowledgement of these IPRs must be made either by explicit symbols or by a rights acknowledgement such as this.

Please
Recycle

Adobe PostScript

Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 USA Phone 650-960-1300 or 1-800-555-9SUN Web sun.com

We make the net work.

Sun Worldwide Sales Offices: Africa (North, West and Central) +33-13-067-4680, Argentina +5411-4317-5600, Australia +61-2-9844-5000, Austria +43-1-60563-0, Belgium +32-2-704-8000, Brazil +55-11-5187-2100, Canada +905-477-6745, Chile +56-2-3724500, Colombia +571-629-2323, Commonwealth of Independent States +7-502-935-8411, Czech Republic +420-2-3300-9311, Denmark +45 4556 5000, Egypt +202-570-9442, Estonia +372-6-308-900, Finland +358-9-525-561, France +33-134-03-00-00, Germany +49-89-46008-0, Greece +30-1-618-8111, Hungary +36-1-489-8900, Iceland +354-563-3010, India–Bangalore +91-80-2298989/2295454; New Delhi +91-11-6106000; Mumbai +91-22-697-8111, Ireland +353-1-8055-666, Israel +972-9-9710500, Italy +39-02-641511, Japan +81-3-5717-5000, Kazakhstan +7-3272-466774, Korea +822-2193-5114, Latvia +371-750-3700, Lithuania +370-729-8468, Luxembourg +352-49 11 33 1, Malaysia +603-21161888, Mexico +52-5-258-6100, The Netherlands +00-31-33-45-15-000, New Zealand–Auckland +64-9-976-6800; Wellington +64-4-462-0780, Norway +47 23 36 96 00, People's Republic of China–Beijing +86-10-6803-5588; Chengdu +86-28-619-9333; Guangzhou +86-20-8755-5900; Shanghai +86-21-6466-1228; Hong Kong +852-2202-6688, Poland +48-22-8747800, Portugal +351-21-4134000, Russia +7-502-935-8411, Singapore +65-6438-1888, Slovak Republic +421-2-4342-94-85, South Africa +27 11 256-6300, Spain +34-91-596-9900, Sweden +46-8-631-10-00, Switzerland–German 41-1-908-90-00; French 41-22-999-0444, Taiwan +886-2-8732-9933, Thailand +662-344-6888, Turkey +90-212-335-22-00, United Arab Emirates +9714-3366333, United Kingdom +44-1-276-20444, United States +1-800-555-9SUN or +1-650-960-1300, Venezuela +58-2-905-3800

LFC 3.2 Printed in USA 01/03 FE1320-1/2K