

CHAPTER 3

Basics of a Digital Library

OBJECTIVES

- **Understand the concept and purpose of a digital library**
- **Differentiate it with the traditional library and understand its advantages over it.**
- **List key components of digital library**
- **Understand the process: Content creation and data management.**
- **Understand the approach to be taken for digitization**
- **Identify the limitations/problem areas and be aware of strategies to overcome them.**

An Introduction

The term "Digital Library" has a variety of potential meanings, ranging from a digitized collection of material that one might find in a traditional library through to the collection of all digital information along with the services that make that information useful to all possible users.

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.

The digital library is

1. Collection of services,
2. Collection of information objects,
3. Supporting users with information objects,
4. Organization and presentation of those objects,
5. Available directly or indirectly, and
6. Electronic/digital availability,

The collection of services

A digital library is much more than just the collection of material in its repositories. It provides a variety of services to all of its users (both humans and machines, and producers, managers, and consumers of information).

The collection of information objects

The basis for a digital library, however, must be the information objects that provides the content-print and electronic.

Supporting users deal with information objects

The goal of the digital library is to assist users by satisfying their needs and requirements for management, access, storage, and manipulation of the variety of information stored in the library.

Users may be "end" users (those not involved in the management and operation of the library but rather are the customers), library operators, and information "producers" who want their material available through the library.

The organization and presentation of those objects

The key to effective collections management is to structure it in a way that the users can understand easily and use.

Available directly or indirectly

These information objects may be digital objects or they may be in other media (e.g. paper) but represented in the library via digital means (e.g. metadata).

Electronic/digital availability

Although the objects may not be electronic, and may not be available directly over the network, they must be represented electronically in some manner through, e.g., metadata or catalogs. Otherwise, we would not consider the objects to be part of the digital library.

- **Documents, such as articles, preprints, working papers, technical reports, conference papers**
- **Books**
- **Theses**
- **Data sets**
- **Computer programs**
- **Visualizations, simulations, and other models**
- **Multimedia publications**
- **Administrative records**
- **Published books**
- **Overlay journals**
- **Bibliographic datasets**
- **Images**
- **Audio files**
- **Video files**
- **Reformatted digital library collections**
- **Learning objects**
- **Web pages**

Digital library therefore is a combination of

- **Services**
- **An architecture**
- **A set of information resources, databases of text, numbers, graphics, sound, video, etc.**
- **A set of tools and capabilities to locate, retrieve and utilize the information resources available (Borgman, 2000)**

Can we call an automated library as a digital library?

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.

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. Ownership, rights management, and commercial considerations become much more complex in this environment.

Objectives of a Digital Library

1. To expedite the systematic development of the means to collect, store, and organise information and knowledge in digital form. To convert the existing print documents into digital form with the help of information technology.
2. To promote the economical and efficient delivery of information to all sectors of the society.
3. To encourage cooperative efforts which leverage the considerable investment into research resources, computing and communications Network
4. To strengthen communication and collaboration between and among the research, business, government, and educational communities.

DIGITAL FACE OF TRADITIONAL LIBRARIES

Digital library includes both digital collection and traditional fixed media collection, so they encompass both electronic and paper materials. Digital libraries also include digital material that exists outside the physical and administrative bounds of any one digital library. Digital libraries will also include all the processes and services that are the backbone and nervous system of libraries. However, such traditional processes, though forming the basic digital library work, will have to be revised and enhanced to accommodate the differences between new digital media and traditional fixed media.

Is a library consisting of digital collections a digital library in true sense?

Digital collections are “raw content,” while “digital libraries [are] the systems that make digital collections come alive, make it usefully accessible, useful for accomplishing work, and connect them with communities.”

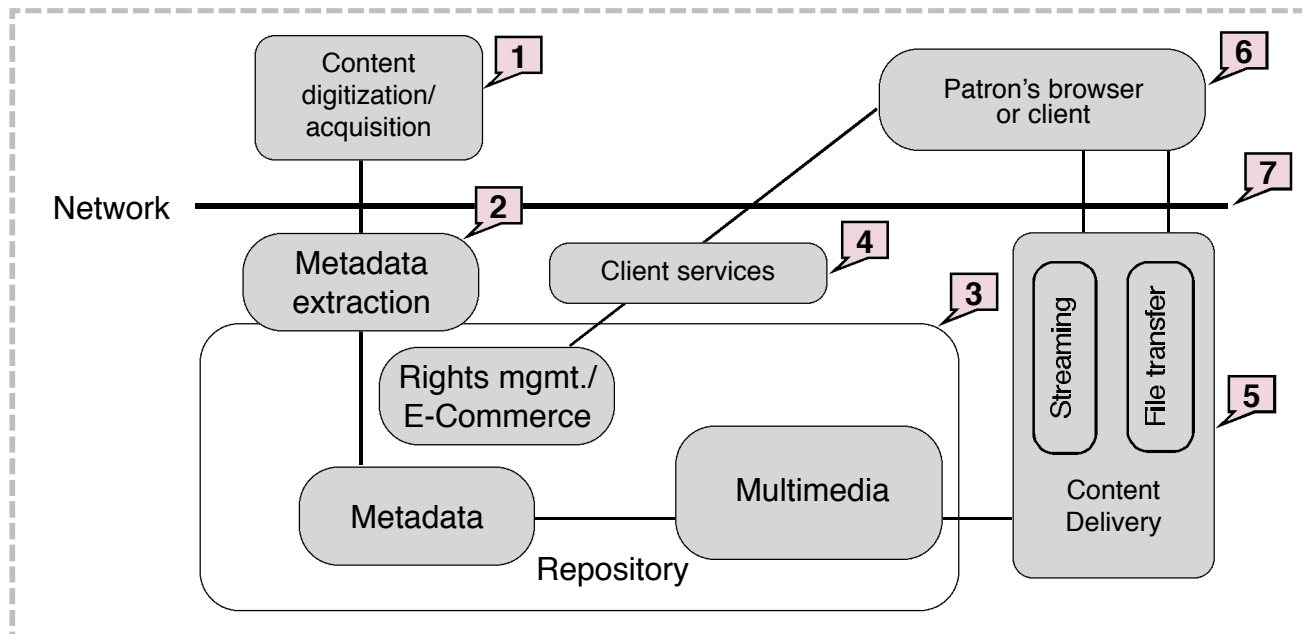
The collections gain value only when these are surrounded by a matrix of content and interpretation that makes them useful. Therefore it should be ascertained that we develop digital libraries, not just digital collections.

Care should be taken to surround collections with appropriate metadata supplying context and interpretation, to develop synergy.

Three general characteristics of the digital library of the future are:

- A comprehensive collection of resources important for Scholarship, teaching, and learning;
- Readily accessible to all types of users
- Managed and maintained by professionals

Building a Digital Library: Getting Started



CREATION OF DIGITAL LIBRARIES

STEP 1: BUILDING DIGITAL COLLECTIONS

There are essentially three methods of building digital collection.

1. Acquisition of original digital works created by publishers, Institutions and other scholars like electronic books, electronic Journals and data set.
2. Access to external materials not held in house by providing pointers to web sites, other collections or publishers' services.
3. Digitisation - converting paper and other media in existing collection to digital form.

Digital resources for a digital library can be categorised under following categories: **Legacy, Transition, Born digital.** Legacy resources are

Some of the important points to be considered in developing a digital library are:

- Acquisition of Digital collections (of various forms and types)
- Conversion of existing print material into digital format: options for conversion
- Archival of digital collections and storage media
- Integrated access interface (Integration with existing databases)

largely non-digital resources including manuscripts, prints, slides, maps, audio and video recordings. Attempts are being made to digitise these resources. **Transition resources** are primarily designed for another medium (mostly print). These are being or have been digitised, making the transition into the digital world.

Born digital/New digital resources are either deliberately created as digital or are created in parallel to print. Publishers are increasingly moving to XML or SGML format. This includes electronic journals, electronic books, etc. Ref: The web for publishers and list of E-journals and books.

THE FOLLOWING IN-HOUSE RESOURCES CAN BE SELECTED FOR THE DIGITAL LIBRARY

The following in-house resources can be selected for the digital library.

1. Internal reports and theses, which require no copyright from the publisher/author.
2. Research papers published/presented by Organisations at various forums.
3. Proceedings of the conferences/seminars/symposia/workshops/tutorials conducted by the Organisation. Invited lectures delivered by eminent speakers.
4. Policies and plan documents.
5. Photographs shot and films produced by the organisation.

2.1 DIGITIZATION: WHAT TO DIGITIZE AND HOW?

The four important steps involved in the process of digitisation are- scanning, indexing, storage and retrieval.

Scanning and capturing the essential components of the original material in digital form is the heart of the process of setting up 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.

The **second stage** of the digitisation 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. 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.

1.3 TOOLS OF DIGITISATION

The important machines and tools needed for digitisation include.

- Computer
- Scanners and scanning software
- Storage system Network
- Display system

a. Computers

There are two components to any modern distributed client/server system;

TIPS

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 then each of the specialist ones will be and the library will be able to grow in the areas where it needs to.

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. This creates the data as an independent resource, which can be accessed (with permission) from any system.

- the server and
- the clients.

Clients are the machines that reside on the user's desks. The library's system can contain a recommended minimum level of equipment (and software) for the user to correctly and efficiently interact with the digital library.

Servers

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.

b. Scanners and Scanning Softwares

Scanners are used to transfer an existing paper image or document into a digital format after which the scanned documents will be manipulated using an imaging software program.

Type of Scanners

- Flatbed scanner
- Slide scanner
- Microfilm scanner
- Drum scanner
- Sheet fed scanner
- Digital camera

Scanner Software

There are two types of software that you will need for most digital imaging jobs.

- Scanning software that comes with the scanner.
- Image editing software, normally applied to the image after it has been scanned.

Optical Character Recognition (OCR) Software

Once the text has been scanned it needs to be run through the OCR program to convert it to a machine-readable encoded form. It allows to scan printed, typewritten or hand written text (numeral, letters or symbols) and/or convert scanned image to a computer readable format, either in the form of a plain text or a word document or an excel spread sheet, which can be used or reused in other documents. (Tip: See tips on indexing also to decide what and what not to OCR)

OCR: Its Limitations

- 1 The problem is that the quoted conversion rates of "better than 95%" is just not true. Proof reading is must for this stage of the work. That is why specialist conversion services exist.
2. Once the OCR has been run then the text exists in machine-readable form and can be indexed by a digital library system and used for retrieval (free text search only).

What to Capture?

Consider Content Capture from a normal printed book or report.

It has both text and images: Ask the following

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?

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 are converted to text. 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 ultimately depend on factors such as:

How the material will be searched for (full text vs. indices and cataloging), How the material will be used (read and transcribed vs. cut and pasted), The functionality of the software (capturing text is of no use without full text searching), The system capacities available (disk space, processing power, network capacity), staff and the time available

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.

1. Name the files by material type and give them some form of sequential number.
2. It makes sense to adopt a simple classification and coding scheme to keep file names.

If the OCRed material is not proofread, the spelling mistakes get into the index. This leads to confusion for the users.

3. Obtaining OCR programs for our regional languages is being tested upon. Organisations like Centre for Development of Advanced Computing (CDAC) Noida have realised this and are working on this.

2.2 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 then store it 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
- If so how many and where do the elements come from

But note: This decision will ultimately depend 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.

2.3 MANAGEMENT OF DIGITAL COLLECTIONS- STORAGE AND RETRIEVAL

Digital Library software

A digital library management system is required to set up to store and retrieve digital collections. functional digital library. This system may be procured from the market. A range of freewares (Open Access Digital Library Management System) also exists for managing digital collections.

Open source digital library software derives its strength from several enabling technology and metadata based inter operability protocols, which have become available recently. Examples of some of these are as follows-

1. Green stone digital library software (GSDL)
2. E-prints
3. DSpace from MIT
4. Site search (OCLC): www.sitesearch.oclc.org
5. PEARS (OCLC): www.oclc.org/research/software/pears
6. Open source software for online journals and conference publishing (e.g. OJS system from the public Knowledge Project, University of British Columbia, Canada)
7. Fedora – Developed by Cornell University & University of Virginia

Organisation of the digital collections for storage and effective retrieval.

As we organise the print collections through classification numbers, publishers, keywords, period, title, author, etc, digital collections also need to be similarly organized in the repository.

Digital Library: Myths and the challenges

Creating effective digital libraries poses serious challenges. An increasingly complex technological, social, legal, and economic environment defines many boundaries within which "digital library" services will evolve. It is worthwhile at this stage to examine some myths and the significant challenges to "digital library" development.

TIPS

1. Refer section on Information Organisation to know more about Indexing and its significance.
2. Refer to the Master Keywords List (or The Thesaurus)

TIPS

1. Refer to the section on Information organisation for tips on HOW TO CLASSIFY and ORGANISE INFORMATION
2. Refer to the section on Information Organisation for tips on HOW TO KEYWORD an information material.

Use of International standards in a digital library management system

It is strongly recommended to ensure that every digital collection is described as per the international norms or in other words the metadata approach is adopted.

Metadata means data about data. It is a description of object, documents, or services, which may contain data about their form and content. This is to improve the possibilities of document retrieval, and to support control and management of collections. Another concept is that it is a machine understandable information for the Web.'

The **Dublin Core metadata standard** is a simple yet effective element set for describing a wide range of networked resources. The Dublin Core standard comprises fifteen elements, the semantics of which have been established through consensus by an international, cross-disciplinary group of professionals from librarianship, computer science, text encoding, the museum community, and other related fields of scholarship.

MYTH 1: The Internet is the digital library.

A global information network, of which the Internet is the seed, has the illusion of promising fingertip access to the world's information. Many would call World wide web, with huge collection of documents a "digital library" because they can read and use whatever they wish by accessing the Web, just as one can use technology to do banking in a "digital bank" or buy compact discs in a "digital record store".

But locating information on the Internet remains highly inefficient compared to traditional library sources, especially for unfamiliar users. Finding information is difficult, the quality of the information is quite variable, and reliable, professional assistance for the confused and lost is lacking. (Ref: Section on Internet for information sourcing)

There remains much work to be done before the Internet will have the coherence and user-friendliness of a library.

MYTH 2: The myth of a single digital library or one-window view of digital library collections.

One can get electronic access to a library without walls where information is accessible anywhere and anytime. [Negro 95]. Digital collections and services will be strongly affected by future copyright and licensing regimes, as well as prohibitive costs for digitization and support of technical infrastructure. "Prime" information resources will probably be locked into proprietary collections essentially "private digital libraries" which are accessible on a subscription or pay-per use basis. Developing interoperability standards for locating and retrieving information in this highly distributed and heterogeneous environment will be a considerable challenge in their own right.

MYTH 3: Digital libraries will provide more equitable access, anywhere, any time.

It is assumed that a global computer network, the Internet or some descendant, will be the primary delivery mechanism for digital information. Limits of network bandwidth and slow transmission speeds may make the effective access to information problematic for many users. The technologies on the desktop, between computers, and for storing and processing information are dynamic variables. What is certain is that the management of technology for digital libraries is becoming more complex as is the administration of licenses and user access. The impact upon equitable access could be considerable.

MYTH 4: Digital libraries will be cheaper than print libraries.

A common assumption about digital library is that it is very cheaper to set up and maintain digital libraries. This is far from established in fact or in practice. the costs of "being digital" are substantive ones. Digital libraries need to invest in hardware and software infrastructure. These expenses will increase, new hardware will be required, more licenses to software, increased infrastructure administration and training. Those institutions that aspire to the development of digital collections and services can expect all of the above plus extensive design, digitization, and implementation costs. So that question that all of us need to ask is how many libraries can afford the effort? And at what cost to the valuable existing services they perform?

SOME CHALLENGES

Resource discovery:

Quality of information retrieved from large heterogeneous databases may be lost in a flood of irrelevant results. Librarians organize knowledge through the processes of subject analysis and cataloguing, creating information about information, or what is known as “metadata”. A major challenge exists to develop methods of consistently and uniquely identifying and retrieving networked information, no matter what format they are or where they reside. **Metadata standards** are still in their infancy. It has been found that given the complexity of metadata issues, a solution to the global resource discovery problem remains distant.

Usability of the digital library

Being digital is not necessarily commensurate with being useable. Considerable study of what users need, how they use information, and whether digital formats serve their needs effectively is still required. Undertaking large digitization initiatives without a fundamental understanding of user needs might result in complete failure of the exercise and lack of support from users and the management.

The reasons this substitution will not easily occur are many: user resistance, limitations on use, poor digital product design, or the medium may not be effective to satisfy the user requirements. The challenge here will be to “span both print and digital materials... [and to] ...provide a coherent view of a very large collection of information.”

Preservation

For example, digital storage media are “fragile”, with a limited shelf life. Worse yet, the digital information on those storage media, even if they do survive will be rendered unreadable by obsolescence of technology. To preserve digital information, digital libraries will continually have to “migrate” information from one digital hardware and software configuration to another.

- Preservation of the storage medium - tapes, hard drives, floppy discs have a short life span when considered in terms of obsolescence.
- Preservation of access to content - this form of preservation involves preserving access to the contents of the document regardless of their format.
- While files can be moved from one storage medium to another, what happens when the formats containing the information become obsolete.

Digital Library Administration

The technical tasks are “the easiest to solve; they will only cost money”. It is the institutional commitments that “will be much more difficult to achieve.”

Internet is the place to find an answer in 3 days for a query that would take 3 hours in a library.

Some of the more serious issues facing the development of digital libraries are Technical architecture Libraries need to enhance and upgrade current technical architecture, such as:

- High speed local network and fast connection to internet,
- Relational database that supports a variety of digital formats,
- Full text search engines to index and provide access to resources,
- A variety of servers such as web services and FTP servers,
- Electronic document management system.

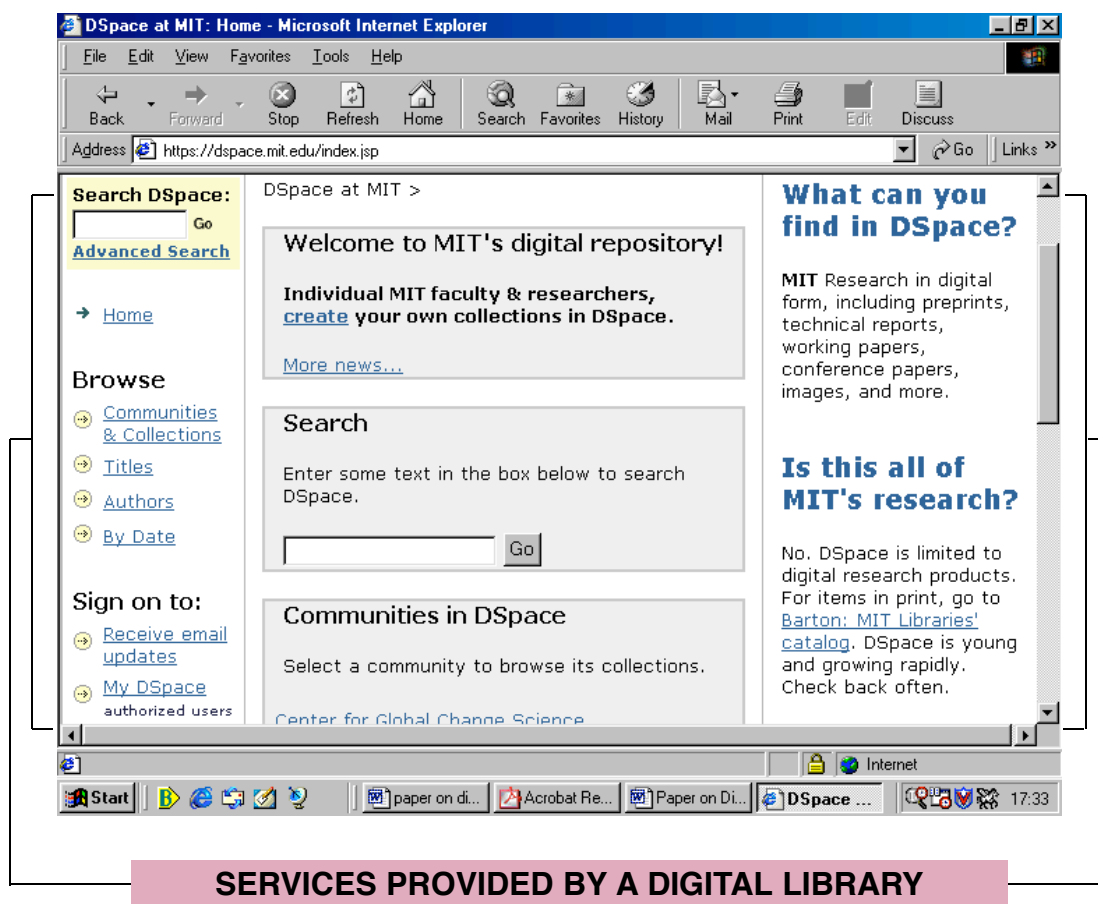
Copyright/rights management

Copyright is one of the most important barriers to digital library development. The current paper based concept of copyright breaks down in the digital environment because the control of copies is lost. Digital objects are less fixed, easily copied and remotely accessible by multiple users simultaneously. It is important to develop mechanism for managing copyright.

SERVICES PROVIDED BY A DIGITAL LIBRARY

The library is not merely a searching tool for people to log onto over the Web or some other access medi-

um. Conventional libraries provide a large number of other services. 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.



Searching facility through Online Public Access Catalogue (OPAC)

Catalogues and their access are the most visible aspect of libraries, particularly when they are on-line. 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. 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. (Ref: Section on Information Organisation)

Free Text Search Facility

Search is not restricted to keywords. It covers the entire content available in the digital library system.

Document delivery

Results of searches are available delivered online. At times, there may be some bulk transfer (such as email attachment, ftp, etc.) if the access is restricted.

Bibliographies

Bibliographies on a subject or a combination of subjects can be created very easily.

Discussion Groups, Forums, News

Since the library is on-line it is possible to run chat rooms or discussions on topics where the library is expert.

Marketing

It is possible to sell the content on a “pay-per-view” basis or we can sell a subscription that gives unlimited access to all, or some section, of your library collection.

Regular E-mail updates on specific subjects

Users can subscribe to e-mail alert services on specific types of collections available in a library and remain updated on their subject areas.

Table 1

Comparative study of some of the open source software

	GSDL	Eprints-II	DSpace	Fedora
Creator	University of Waikato	University of Southampton	MIT libraries & Hewlett-Packard	Cornell University & University of Virginia
Open Source and Free	Yes	Yes	Yes	Yes
Operating System	Unices, Windows	Unices	Unices	Unices, Windows
Web-server	Apache/ IIS	Apache 1.3	Apache 1.3/2.0 and/or Tomcat	Tomcat 1.4
Language	Perl	Mod-Perl 1.0	Java 1.3, JSP	J2SDK v.1.4
Database	Its own	MySQL	Postgre SQL 7.3	McKoi v.0.94 (uses by default) MySQL/ /Oracle 9i (optional)
Resource Identifier	No	OAI Identifiers (similar to URNs)	CNRI Handles	Uses own persistent identifiers (PID)
Dublin Core	Dublin Core	Dublin Core	Qualified Dublin Core	Dublin Core
OAI-PMH v 2.0	No	Yes	Yes	Yes
Subscription	No	Yes	Yes	No
Supported File formats	MS-Word, PDF, HTML, PostScript, JPEG, GIF.	PDF, MS-Word, HTML, JPEG, GIF.	MS-Word, PDF, PPTs, JPEG, GIF.	PDF, MS-Word, PostScript, JPEG.

Note:

These features may change as newer versions of the software are made available.

N O T E P A G E S

Lined area for notes, featuring horizontal ruling lines and a faint pencil drawing of a telescope.