DESIDOC Bulletin of Information Technology, Vol. 17, No. 6, November 1997, pp. 7-14

Digital Library: Content Preservation in a Digital World

Richard P Hulser*

Abstract

There is a growing need and worldwide trend to collect, organise, manage, protect and distribute information in digital form. In this paper, a brief discussion about information services trends and directions expands on this point and is followed by a description of IBM Digital Library whose components enable institutions to addresses digital content management issues. The description includes the integrated functions of create and capture, search and access, distribution, storage and management, and rights management. Guidelines and considerations which are important when contemplating implementation of a digital library project conclude this discussion.

1. INTRODUCTION

The increasing use of the Internet and World Wide Web (Web) has developed awareness and concerns about access and retrieval of information across networks. What about the content travelling across those networks? Where does it come from? Where is it stored, and how is it managed? Are there ways to ensure quality and protection of the information being transmitted and made available to others? Fortunately, there are now technology solutions which can address these and other questions.

1.1 Information Services Trends

Institutions, libraries and end-users are all affected by, and at the same time, influence information industry trends. People are becoming increasingly aware of the usefulness and capabilities of computer technology for work, research and recreation. Users are demanding a broader variety of information services which can be tailored to their wants and needs. They also are demanding better quality of such services and a single point of access which requires a minimum of training to use them effectively.

In order to address some of these requirements, libraries, archives, publishers and businesses are providing electronic access to information to augment or substitute for information previously provided only in print and other physical forms, such as films and videotapes. Institutions are converting such analogue information collections to digital form to enable faster and easier access to them. This requires robust networks in order to provide adequate bandwidth, or space, for information to flow through the world-wide networks. Fig. 1 depicts the emerging trends.

In addition to the technical challenges these requirements present, there is the challenge of copyright and ownership verification which will be referred to here under the general term of rights management. Laws and regulations related to rights management can affect the availability of content regardless of format, but is particularly evident in the digital arena.

*Digital Library Consultant
IBM Corporation
rhulser@vnet.ibm.com

1.2 Why Digital Libraries?

Why is there a need for digital libraries? Just as a room full of books requires some kind of organisation in order to effectively find and access them, the same is true in the digital world. One reason for using digital library technology is to manage large amounts of digital content such as thousands of images or hundreds of audio clips. Advances in storage technologies have enabled large amounts of content to be made available locally at increasingly affordable costs. While discussions previously centred on megabytes (MB) of data, now it is not unusual to consider gigabytes (GB) and even terabytes (TB) of storage. The scanned pages of one million books as images, for instance, would require 50TB of storage if those books are assumed to average 500 pages per volume and one side of each page required 50KB (kilobytes) of storage. In contrast, 635MB is needed for a typical one-hour, stereo-quality audio compact disc and 720MB for a standard one-hour videotape.

Documents and other materials housed in collections are deteriorating at a rapid rate. While much work is done to conserve and preserve these collections in their original form, digitising this content enables it to be preserved in an additional way. Libraries and archives with unique collections are digitising these materials to not only preserve them, but also provide more extensive access to their content. This is difficult with traditional printed materials in many cases due to physical limitations of library facilities and the fragility of the materials in the collections.

Another reason to use digital libraries is the need to be able to perform searches that are impractical manually. The vast amounts of information being created and stored each day makes it that much more difficult to find specific information later. Adding to this are images and multimedia content which present even more challenges for retrieval. For instance, finding all the images with blue skies and white sandy beaches can be done manually if the images were indexed with colour and other parameters in mind, and given enough time and patience. There are electronic tools which enable this to be done quickly and easily by text search and

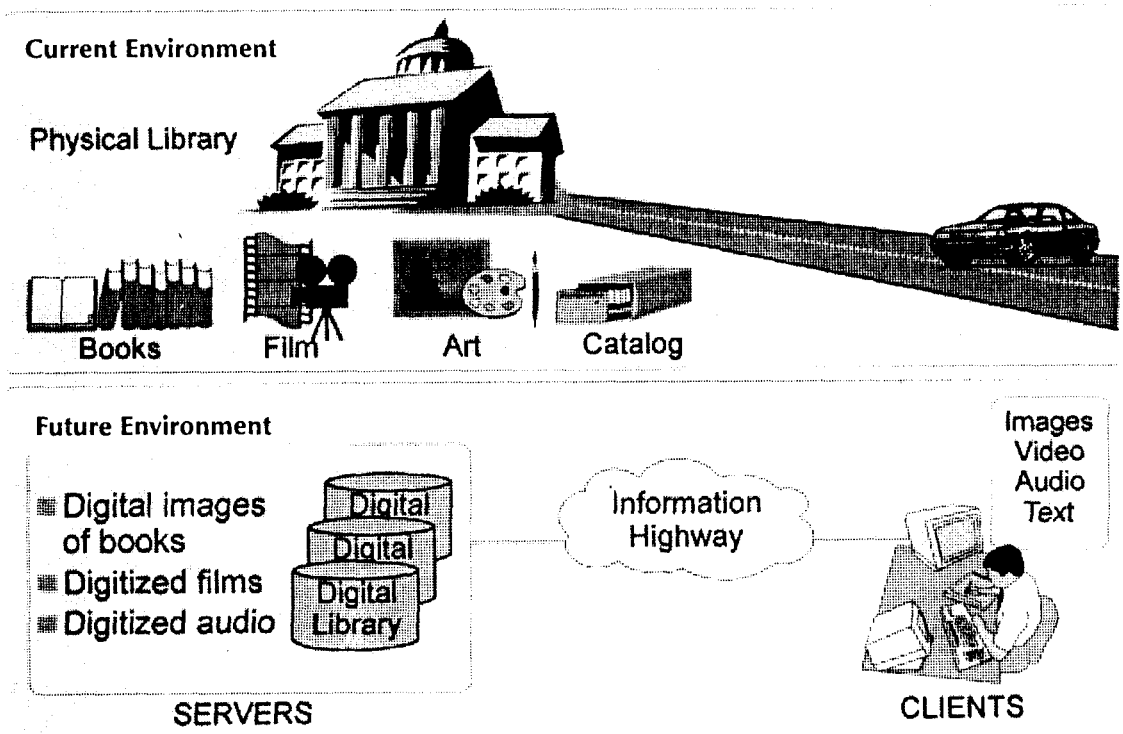


Fig. 1 From Physical Library to Digital Library

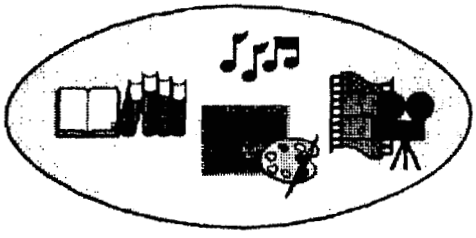
also by actually searching on the colours, textures and even shapes in images.

While local ownership of materials has been the practice for centuries in libraries, this is becoming impractical and unaffordable. Digital libraries enable greater access to content, are capable of managing content from multiple locations, and provide a way to enrich the teaching and learning environment. As a result, there is less need to own everything and more emphasis on providing access to it, whether available locally, from a distant source, or even from multiple sources.

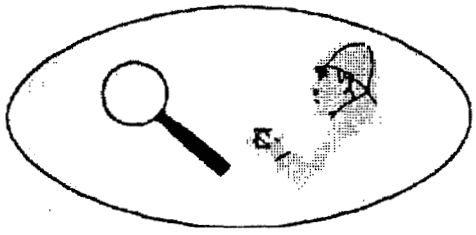
There is already a great deal of information available on the world-wide networks, though still only a small portion of what exists in print. There have been instances where some of that information was proven to be false. In other cases, information has been changed without an owner's permission or even reproduced and claimed by others as their information. Thus, protection of the content owners' information is important and skills in information literacy become imperative. Whether in analogue or digital form, identification of original, unmodified information is critical. In addition, the owners of information may want to be financially compensated for the use of their ideas and work which have been described. Rights management tools as part of digital library technologies provide ways to electronically identify an owner and deter misuse of digital content. Fig. 2 shows the four key features of a digital library.

1.3 Digital Libraries: A Definition

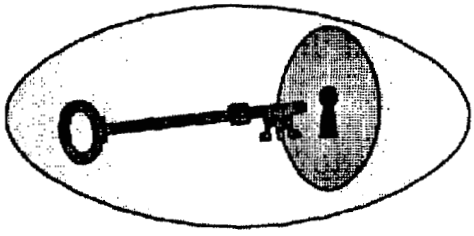
There is currently no consistent, world-wide agreement on the definition of a digital library. There are many descriptions, definitions and varieties of digital libraries. The size of files, the format of content housed, and the access restrictions (or lack of restrictions) are some of



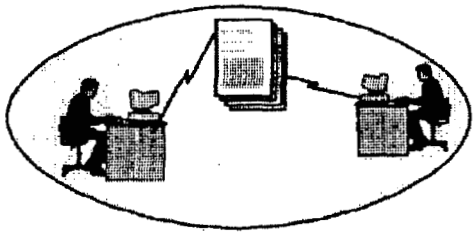
Manage very large amounts of data
Preserve unique collections



Perform searches that are impractical manually



Protect content owners' information



Improve access to information
Deal with data from multiple location
Enhance the distributed learning environment

Fig. 2 Digital Library : Key Features

software contained on a CD-ROM is described as a digital library as well. A series of digital audio or video clips or full-length movies can also be thought of as digital libraries.

One kind of digital library often discussed is that of content managed and accessed on a Web server. Browsers which enable access to the Web are growing in popularity due to their relative ease of use, thus contributing to Web servers as a typical warehouse for digital content. Information which is for broad distribution, has little or no rights management requirements, and is relatively small in size (no more than a few megabytes) can be stored, managed, and accessed using Web servers. Work is underway by many companies to enable protection of sensitive information transmitted over the Web, such as credit card data, so that electronic commerce can be conducted in a secure way.

There is often a need to collaborate on development of content and to share information during that process. Whether it is a group working on a research project or developing a product or discussing a topic, collaboration is taking place. Groupware such as Lotus Domino combine an ability to do this while also taking advantage of the ease of use of Web browsers. Digital libraries manage this information. There is a limit on the size of files which such software can handle, typically several megabytes.

It is a challenge when there is a need to limit access to a variety of content, perhaps charge for access and delivery of such content, and manage content which is many megabytes, gigabytes or larger in size. Scalability for later growth is also important. Digital library is a relatively new technology and one of the ways to ensure success using this technology is through development of a technology strategy plan and implementation of a proof of concept. From this proof of concept the implementation can be enlarged to incorporate more content and application of requirements as needed. Products such as IBM Digital Library can provide management of content with these requirements and characteristics.

It is important to note that any and all of these can be considered digital libraries which can be implemented individually or in combination with each other.

1.4 IBM Digital Library: Content Access with Protection

IBM Digital Library is a solution for storage, management, and distribution of all types of digital content, including text, images, audio and video. It is the incorporation of creating & capturing, storage & management, search & access, distribution, and rights management of digital intellectual content in an open, scaleable, multi-platform environment such as AIX and Windows NT. The four functional units of the IBM Digital Library are shown in Fig. 3.

The key is the integration of these functions into a single management system. Sophisticated enhancements of the individual functional areas give further utility to such systems. For instance, IBM Digital Library has the capability to incorporate search technologies such as text mining, federated search, and image content search seamlessly with the base parametric and natural language text search. It also has multi-language development tools, integrated multi- search capability, and Java-based capabilities as well.

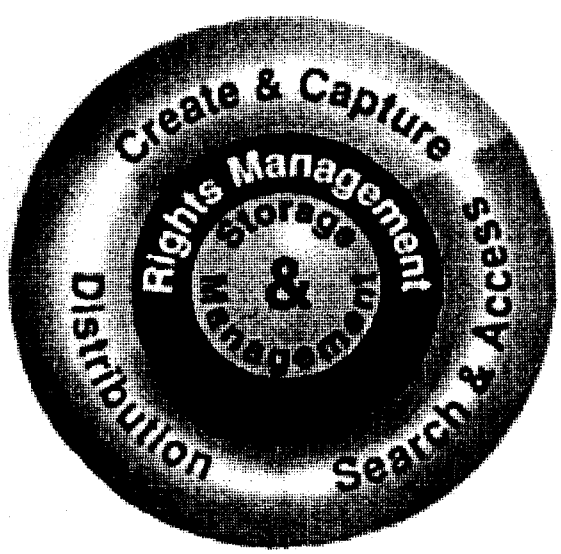


Fig.3 IBM Digital Library : Key Functions

1.5 IBM Digital Library Architecture

The IBM DL triangular architecture (see Fig. 4) enables flexible and scaleable storage and management. It is based on a triangular client/server model comprised of a Library Server, one or more Object Servers, and one or more clients.

The Library Server is the core of the infrastructure and manages the catalogue information, locates stored objects using a variety of search technologies, provides secured access to the objects held in the collection, and communicates with the Object Servers. Client communication through the Library Server helps ensure the integrity of the data objects.

Digital content is stored in the Object Servers, which can contain a wide range of digital formats such as text, image, audio and video content. These can be available n a single location or distributed across an institution's network to provide convenient user access. This is important for delivery of large multimedia objects such as audio and video.

The end user (or client application) receives direct access to the information requested from a collection regardless of where the data is stored. Many clients can be connected to the IBM Digital Library system, including supported Web browsers. A multi-search Internet application is included with the system, along with sample content.

The IBM Digital Library architecture is scaleable and can be implemented across multiple platforms allowing both integration of additional Object Servers on the same or different platforms, and scaling up from one operating environment to another. The system can grow from a single workstation to a networked enterprise-wide implementation. Collections can be maintained on a single personal computer hard drive, then grow to be managed by storage servers that can contain terabytes or even petabytes of content. This becomes increasingly important as more and more content is added and the number of users of the system expands.

1.6 Digital Library Positioning with Library and Collections Management Systems

While digital libraries are being used as standalone systems, the focus is on integrating them with existing systems such as on-line catalogue systems (integrated library systems and collection management systems) so that the user can access the content through a single interface.

Support of the Z39.50 communications protocol is one step in this process and will be included in the future release of IBM Digital Library. This will enable search of IBM Digital Library using a Z39.50 compatible system.

IBM Digital Library is not a replacement for automated library or collection management systems, but rather a complement to them. Automated library and collections management systems were built and optimised for managing physical objects and they do this well. While many of the vendors of these systems have been working on developing systems which also manage digital objects, such work is still in the early stages.

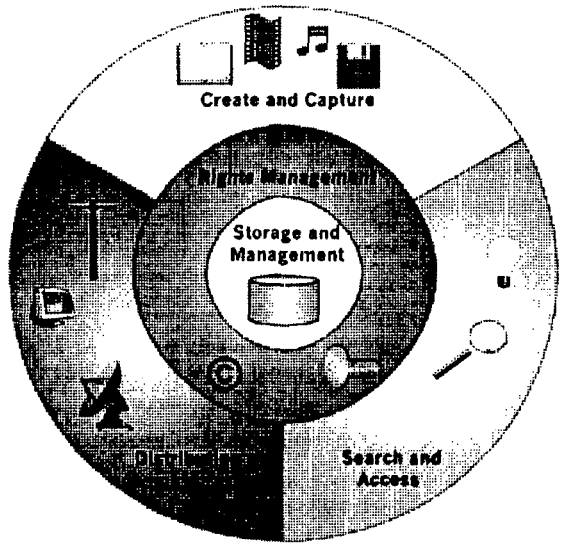


Fig. 4 IBM Digital Library : Architecture

1.7 IBM Digital Library Collection Treasury

There are institutions with archival collections that have either already been captured in electronic form or have plans to do so in the near term. For various reasons, these collections may not initially require a sophisticated management system yet there is a need to store and access their content quickly and easily. The IBM Digital Library Collection Treasury (DLCT) solution was developed in conjunction with several customers to meet this need.

The IBM Digital Library Collection Treasury solution enhances and works in conjunction with the base IBM Digital Library. This solution extends world-wide access to special collections by providing high quality image storage and management, batch load capability, and the ability to create and store derivative images. It takes advantage of the IBM Digital Library triangular architecture, scalability and other components while providing additional tools for management of digital collections.

Archive materials are very delicate, to be accessed with a minimum of handling, and may only be able to be used a few times before they crumble. Such materials require digital capturing with the highest resolution possible as that may be the only time to do so ever again. In addition, the high resolution enables careful study of the image not otherwise possible, sometimes even better than examining the original. There are an increasing variety of capture devices, including digital cameras, which provide a way to digitise images with high resolution.

Once an image is captured, DLCT provides tools to create derivative images of the original scanned item. Usually three or four derivative images are created. One is a thumbnail, a small image with low resolution typically used in Web pages. The second is what is known as a screen size image, one that fits the viewing screen and suggested to have a digital watermark integrated with it to designate ownership. The third is a full image, one that is higher in resolution and enables the user to look more closely at the details of the image and spills over beyond the size of the screen. Sometimes there is a need to provide users with a very high resolution image which is a duplicate of the original scanned image which is not made generally accessible.

If a collection of images has already been captured in electronic form, these can be batch loaded with associated metadata which is defined by the individual institution into the DLCT. This saves a great deal of time and work.

1.8 Rights Management

Rights management has a variety of characteristics and is an important issue, that is why it is treated here at length.. It can include ownership designation, billing requirements, or royalty management. There are technological, sociological and cultural issues that affect rights management and these will be reviewed before addressing some of the technologies enabling rights management.

There are a number of non-technical factors affecting rights management. Legislation, laws, regulations and licenses are developed to protect intellectual property rights of their owners. The strength of their enforcement deters improper access and use of intellectual property. Societal behaviour and culture influence the creation of such laws and regulations, but also determine how closely they are followed. While much work is underway to have global agreement on intellectual property rights, there continue to be many differences among countries in interpretation and enforcement of these agreements. This is becoming increasingly important as world-wide access to information occurs through networks such as the Internet and World Wide Web.

Billing costs and pricing policies are additional factors to be considered. In some cases, use of intellectual content may not require any fees to be imposed. However, there is high interest in being able to efficiently and securely manage access to contents for a fee. The fee may be a pay-per-use, pay-per a pre-set number of uses, or perhaps a site license for unlimited use by a specific group for a specified period of time. Alternatively, it may be more expensive to collect the money for use of a

particular item or piece of that item (such as a diagram or image within a larger item) than to simply grant permission to use it. Pricing policies are developed by institutions which address these concerns and then implemented using various methods, including technology tools.

There may be risk trade-offs to consider when providing access to digital content. Is it worth time, effort and money to implement safeguards around a collection of information, or should the information simply be provided globally and challenges made on ownership when violations of use have been detected? The user is expected to accept and adhere to policies and procedures to protect intellectual property as set up by an owner or manager of that information. The user may be unwilling or unable to pay for access to a particular piece of information. Related to this is the perceived or marketplace value of information as a product. This may be intrinsic, but may also exist based on its timeliness or longevity. The modularity of information for distribution may also affect its value.

1.9 Technology Factors for Rights Management

Technology which addresses some issues related to rights management has been developed. Marking technologies enable designation of ownership of digital content and deter misuse. IBM has developed visible and hidden marking technologies as part of IBM Digital Library. Visible watermarks utilise a patented process called variable random brightness alteration (VRBA). The brightness of the picture elements, or pixels, of an image is increased or decreased in a random way to create a watermark based on a black and white bitmap. The bitmap can be a logo of an institution, a person's name, an accession number, or even a machine readable barcode. It can be positioned across the entire image or just a portion of the image. In addition, it can be adjusted to be clearly visible or almost transparent.

Another marking technology is data hiding. Data hiding is done by adjusting one of the eight bits representing a byte of information of an image. This is called LBC (Least Significant Bit Coding). Text or bitmaps can be hidden to the least significant bit of each pixel on every colour plane (blue, green or red). The user only sees the image and does not have access to the information encoded within the image.

Packaging information in secure containers for transmission across networks is another way to protect intellectual property. Cryptolope technology enables such secure distribution of content over networks. A cryptolope is an envelope or package which uses cryptography to protect its contents from unauthorised access or alteration. It can have the content description, content usage rules and authorised actions such as view, print, copy, etc.; as well as price per usage type. This technology has been further enhanced to bind code and data together into a single object which can be programmed to govern how the enclosed information is used and other capabilities.

1.10 Digital Library Planning And Development

Before embarking on implementation of a digital library, it is important to develop a strategic technology plan. This plan should address current and future needs and align with the mission and goals of the institution and department. Such a plan is usually developed by a team representing a cross-section of the institution and includes people responsible for information management at the institution, such as librarians and computer services personnel.

Considerations should include issues related to physical resources, personnel and organisational structures, as well as policies and procedures in place for management of intellectual content. Some initial points to address include: examination of current and future information services and resource requirements of people throughout the institution; the quality and procedures of current information services to meet institutional needs; and the priorities for current and future information services.

Physical resources issues should include space, application systems currently installed

such as an automated library system, appropriateness of centralised and decentralised technology placement and support, as well as identification of areas where outsourcing services may be appropriate.

Personnel issues can include the training needs of current and future staff as well as identification of the changing role of information professionals within the organisation and skills assessments. Just as there is a significant change in technology capabilities, so too is a need for additional or different skills within the staff responsible for information services within an institution. Plans need to include an understanding of skills and talents that are readily available across the institution, regardless of department, so that there is a maximum use of these skills. This will enable identification of needed skills and prompt review of providing training to current staff or hiring new staff with those required skills.

Co-operation within and across departments becomes imperative in order to maximise use of skills and available technology. Organisational structures and management styles may need to be evaluat ed and possibly changed to enable such co-operation to occur. Creation of formal and informal teams to tackle projects is one example of how this could be done.

Careful assessment of current policies and procedures on information access and sharing within and beyond an institution should be done. Technology provides means of access to information previously not possible and may require adjustment to the current ways of doing business. In addition, budget planning for long-term maintenance and replacement of systems should be included throughout the planning process.

2. SUMMARY

Digital library technologies can be applied broadly across all kinds of institutions. These technologies were almost unheard of just a few short years ago and now the pace of development and application of these continues to accelerate. Advances such as high resolution capture devices, sophisticated search engines, and affordable large storage of digital content all contribute to the ability to implement a digital library. Some of the resulting benefits include access to intellectual content not easily obtainable previously and an ability to create content and distribute it to many people or have it accessed by many people at the same time.

Regardless of the features that may be possible with digital technologies, it is important to ensure implementation through the development of a strategic technology plan which is tied to the needs of users and fits within the mission and goals of the institution. The strategic technology plan sets the blueprint from which the institution can implement an appropriate and successful digital library project.

3. RESOURCES FOR FURTHER INVESTIGATION

Additional information about some of the topics presented in this article can be found on the Web as follows:

IBM Digital Library
www.software.ibm.com/is/dig-lib/

IBM Digital Library Version 2
www.software.ibm.com/is/dig-lib/v2factsheet

IBM Digital Library Collection Treasury
www.software.ibm.com/is/dig-lib/dlctprod.html

Cryptolopes
www.cryptolope.ibm.com

Safeguarding Digital Library Contents and Users : Digital Images of Treasured Antiquities
www.dlib.org/dlib/july97/vatican/07gladney.html

Watermark Technology Overview
www.research.ibm.com/image_apps/waterma rk.html