MANAGING THE DIGITAL CONTENTS CHALLENGES IN SERVICING A DIGITAL LIBRARY

By

Shripad S. Hardas*
Wilson K. Cherukulath**
Narendra Gupta***

ABSTRCT

Managing the digital contents is a generic problem. Digital contents can include anything from business logic (rules, procedures, standards, specifications, communications etc.) and business transactions to digital books, video and images (digital library) gathered for the purposes of reading, viewing, listening, study and reference. Thus digital library can be called a subset of digital contents. The complexity of managing them is directly proportional to the volume, location and type of contents. Storage technology is advancing and access technology is undergoing changes in terms of hardware and software. The technical challenges in managing and servicing a digital library, have been presented here. Numerous examples have also been given to elaborate the ideas.

* Scientist, Institute of Armament Technology, Girinagar, Pune-411025, India
** Head - Information Center and Library, Institute of Armament Technology, Girinagar, Pune-411025, India
*** Chairman - Faculty of Computer Engineering., Institute of Armament Technology, Girinagar, Pune-411025, India

0. Introduction

In today's business environment, the time taken by a new product - from concept to reality - is decreasing rapidly. This needs data to flow speedily across various departments. To make this happen, the key to success is the digitalization of products and processes. According to Morgan Stanley Dean Witter Internet business-to-business 2000 report, "Around 60% of all product life cycle support information (e.g. technical manuals, fault diagnostic information, parts catalogue), resides in paper, which is in poor physical condition, often outdated, not easily accessible and not integrated into the product business" [4]. This typically represents a library using print media. Academicians and research community also needs access to global information resources to remain at the cutting edge of the technology. In the age of IT, luckily no one has to wait for information.

Libraries of tomorrow may not physically have wall shelves, book stacks and a reading room, but logically the concepts will exist. They will still provide all the technical services that today's library provides viz. acquisition, cataloging, classification and services to users. New ways to access the contents may emerge, which will allow the reader to be anywhere and still enforce copyright laws. The term 'servicing' spans storage management, and providing easy access to digital resources.

1. Infrastructure for Data Processing and Data Transport

To make the digital library into a reality, the basic infrastructure like computer, network and some peripherals to acquire and display data are essential. Following paragraphs describe the same, with reference to their characteristics.

1.1 Server

A network data server is a class of computer, which is optimized for input/output (I/O) rather than processing. It has built-in features for redundancy and integrity. Thus it may have redundant power supply in case first fails, may have redundant array of inexpensive disks (RAID) to cater for mass storage. To make it crash-proof, the future servers will have a "switched-fabric I/O architecture" in which the data transfer between memory, processor and peripherals may be directed by a switch that is outside the processor. When that happens, the CPU may be able to reboot a switch locked up due to a malfunctioned device. It will also eliminate the need for peripherals to share a bus for data transfer, because a separate data channel will exist for each one.

1.2 Storage management

Everyone depend heavily on information for gaining competitive advantage. The challenge is to protect the huge data gathered, make it available to all users and applications, and maintain minimum downtime and maximum performance at the same time. The storage media needed depends on the aim of the library – either to provide *on-line* access or to provide *standalone* access or both, to the digital data. At best, CDROM based contents may suffice to provide standalone access but not the on-line.

Currently, hard disk manufacturing technology uses the giant magneto resistive (GMR) heads in place of MR heads. GMR provides nearly three times storage density compared to MR. Other techniques that affect the storage capacity are increasing the number of platters, maintaining a constant linear storage density etc. But a single disk is no answer to 'terabyte-storage' problem.

This has led to reconsider the storage management. In the first place the stress was given on increasing the storage capacity – as mentioned above - and secondly such disks were made to work together, providing terabytes of storage space, popularly known as redundant array of inexpensive disks (RAID). Such inexpensive disks are made to work together, but only inside a cabinet space, i.e. over a short distance of say 70 mtr. RAID allows a single large file (e.g. library database) to be stripped across multiple disks, or different databases to be located on separate disks, for faster access. In addition, it can also provide redundancy, for fail-safe operations. Five such techniques are described by RAID standard. Main point to note here is that the storage is attached directly to the server.

Imagine a situation in which server crashes and needs a reboot. This can probably happen when multiple users will start accessing video library, stored directly on a server. The data attached to this server will be unavailable to the clients for 30-90 seconds and there can be no business!. Imagine how the backing up the data can be a headache for network administrator, when many servers exist in a network with local storage. Lastly can terabytes of data (say video or satellite imagery) be stored on single server?

Storage area network (SAN) was evolved as a solution to the problems mentioned above. SAN is a separate network, which connects storage devices regardless of server platform or storage vendor, through fiber channel and switches. This allows the storage to be added without disrupting the server availability. SAN incorporates fiber channel arbitrated loop, (FC-AL) which allows 126 devices per loop. FC allows muticast, permitting addition and removal of multicast groups. FC can integrate diverse protocols, permitting storage to be shared amongst all servers (may be Unix or NT). SAN using single mode fiber can extend over 10 km, with multimode fiber it can extend over 500 mtr, while using copper cable it can extend over 30 mtr.

Network attached storage (NAS) provide a simple, reliable and cost effective way to add shared storage to the network. It uses thin server technology, which performs reduced set of server functions. So it does not need a separate file server, but still has an interface to the network. The key advantages of these technologies are…
* Scalability provided in storage capacity and access performance
* Freeing up the server CPU from storage management tasks
* Centralized storage management
* Free up LAN channels for application's use while separate optical network caters to the data access from SAN.
* Manage heterogeneity e.g. different operating systems like Unix and Windows

SAN and NAS will co-exist. Companies like HP and IBM are providing the needed optical components and necessary advice (preservation and integration of existing resources) to install SANs. (`www-1.ibm.com/services/its/us/sanoverview2.html`).

To access the data stored on CD ROMs, a CD Server (e.g. Hewlett Packard SureStore™) be attached to the network, although it is not an effective alternative to on-line library access. The typical problem is the dependence of the user on someone to change the CD ROMs at the CD Server. This option is useful in cases where Internet access is not easy in terms of bandwidth and reliability. IEEE has an option for subscription to 'standalone access' or 'on-line access for n users', to their library.

1.3 Transport network

For information dissemination to be effective (i.e. fast and reliable), a suitable network infrastructure (e.g. media, switch/hub, gateway etc.) is needed.

1.3.1 Bandwidth

Networks can be classified as *wired* and *wireless*. Laying and maintaining a wired network is time consuming and administratively difficult task. Wired networks like public switched telephone networks already exist, while some private Indian companies (belonging to Reliance Ltd. and BSES Ltd.) are planning optical fiber network having a high bandwidth suitable for information superhighway.

Wired networks include telephone networks and optical fiber based networks. Telephone networks have existed since long. Techniques like digital subscriber line (DSL) and asymmetric DSL (ADSL) enhance the bandwidth of such networks. ADSL has maximum upstream speed of 1 Mbps and downstream speed of 8 Mbps, and supports voice and data simultaneously. Optical fibers have high bandwidth compared to copper cables, because they carry signals in the form of light.

Wireless networks include very small aperture terminal (VSAT), wireless in local loop (WLL) and mobile networks. VSAT networks use KU band, but have transmission delay of 500 ms. WLL networks are useful in a congested city where laying copper lines is difficult, but work in a small area. 2.5G networks for mobile phones, which use global system for mobile communication (GSM) enhanced with general radio packet switching (GPRS), provide 100 Kbps. In future, 3G networks which use wavelength code division multiple access (WCDMA) will provide a bandwidth of 2.4Mbps, sufficient to display multimedia on mobile phone screen.

Unfortunately, a real life network is a mix of wired and wireless networks, which leads to complexities.

1.3.2 Protocols

Protocol design characterizes the parameters like congestion control, flow control and reliability. It takes into consideration the needs of the application at hand, and the characteristics of the network infrastructure available to it. The ultimate aim is to deliver end-to-end Qos in such a way, as to allow the end user to negotiate QoS.

Connection oriented protocols, like transmission control protocol (TCP) and hypertext transfer protocol (HTTP), establish connection before communication starts. So they guarantee delivery of packets, needed by text messages, but not necessarily in real time. *Connectionless protocols*, like user datagram protocol (UDP) and real time protocol (RTP), send the packets fast. They give fast response, needed typically by video streaming, but do not guarantee delivery of packets, hence some degradation in video.

Asynchronous transfer mode (ATM) network carries the data in packets, each of 54 bits. IP carries data in packets of variable size. So IP over ATM needs protocol conversion. Protocol conversion consumes time.

1.3.3 Switches

Switches are the devices that are used en-route to store and forward data packets. Switches have been made intelligent to detect the type of packet. The ones that do not require integrity check (e.g. video data) or security check may pass through, saving time.

The features like dynamic routing and caching implemented in the switches, will be useful in the presence of simultaneous video streams. ATM switches analyze the packets using hardware, which is fast compared to software based analysis. Optical switches in future may not need to convert data from optical to electrical and back, for switching. This will enhance switching speed.

1.3.4 Next generation Internet

Networks can also be classified as *public* and *private*. The public network Internet2 (www.internet2.edu) [1] is likely to replace today's Internet. Internet2 had identified four technical innovations. They are…

(i) Multicast messages across heterogeneous networks, for which protocols like PIM-SM, MBGP, and MSDP have been designed.

(ii) Quality of Service (QoS)

(iii) Internet Protocol version 6 (IPv6) to provide security and more address space.

(iv) Support for large delay-bandwidth products, for advanced networking applications, which need high bandwidth, low latency and low jitter, which cannot be implemented on today's Internet. This will help in adding more users on the Internet, and also add QoS and bandwidth.

2. Access provision

Sorting, searching and retrieval techniques are of paramount importance to the user, if the digital data is to be gainfully utilized. In addition the issues like authentication of data, authorization of user and security of data also play a vital role in a digital library.

2.1 Metadata for searching

Metadata is data about data. Achieving significant capture throughput is directly related to the quality and abundance of metadata available to the search tool. As per RLG Diginews (www.rlg.org/preserv/diginews/diginews4-3.html), metadata can be classified as…

(i) Administrative : who owns the data, when and how was it created.

(ii) Structural : required by computer programs. It supports preservation of complex objects by representing relationship between components such as sequence of image.

(iii) Descriptive : supports discovery through search and browse functions.

The current areas of research are directed towards how to collect sufficient metadata, automatically, and some success in that direction has been reported in (http://tamu.edu/DL95/papers/ kacmar/kacmar.html).

Libraries use mainly three schemes of classification. They are Universal Decimal Classification (UDC), Dewey Decimal Classification (DDC), and Colon Classification (CC). If the originator can create the classification code and pack it along with the data, then classification task can be automated. In addition, the administrative metadata can be extracted automatically.

Full automation of indexing the digital contents, will need some time to achieve success. For example, query on images can be pixel based or descriptor based. If heuristics can be used on pixel based queries it can deliver quicker results. Other criterions used are normalized correlation coefficient and sum of squared differences. But one must also note that the satellite image of the same region in the presence and absence of clouds can vastly differ, making indexing difficult. Invariant parameters of an image (also called as 'features') may also be stored as metadata, but may not yield optimal results. The image may possess too many significant and non-significant objects and events, not anticipated in advance. As an example, a human face can have a low resolution image at pixel level, flesh tone, skin or hair texture and face shape at descriptor level and name of the person at the highest level - all acting as search keys. A postal stamp can have the text embedded in the picture, objects in the picture itself and the cost, country of origin, occasion for publishing – all acting as keys. [5]

Another area of active research is about querying a video database. Generally 10-20 text keywords may be stored with the video to reduce initial search. For example, certain movie may contain 'sunrise' and 'sunset' both, so that both of them may be used as keywords. If the need is to locate one of these scenes, either the offset in the movie has to be stored, else do the image processing of neighboring frames to find out the motion of sun (relative to horizon), to locate the frame automatically. Some demo on searching video and images library can be seen at `www.altavista.com` and `www.ctr.columbia.edu/VideoQ/`

If the data is static in nature it can be sorted, once for all, and searched fast. If the data is of dynamic nature, that is constantly changing, it needs to be indexed. Metadata is useful for both the types of data. Various sorting and searching techniques are described in [2].

2.2 Browsers for searching

Universal availability of browsers is of utmost importance to the user, as it provides a consistent user interface, irrespective of the familiarity of the user with the data access equipment (e.g. a computer node). Netscape Navigator and Microsoft's Internet Explorer are two such browsers, but their size may be too bulky, for mobile Internet devices of today.

Browsers mentioned above are bulky because, one thing they support is interpretation of scripting languages. The *scripting language* used decides the outcome of a query. A simple query on 'Bill' can result in answers related to 'Bill Gates', 'Parliament bill', 'Bill/cash memos', 'Bill of material in CAD/CAM' and so on. To make the results more precise, new tags will be required, say 'Bill' as 'Name', which XML allows but not HTML. HTML allows using <H2>Apple</H2> without reference to Apple as a company or as a fruit. XML differs from HTML in three areas:

(i) Information providers can define new tags in extended markup language (XML)

(ii) Document structures can be nested to any levels of complexity.

(iii) XML document can contain an optional description of its grammar, for use by applications that need to perform structural validation.

The next important issue is *authenticity* of data. To decide on the information quality obtained through search engine, someone has to review it and rate it either with a digital score or with stars (so probably **** is better than ***). Some criteria like accuracy, appropriateness, organization of information, bibliography, completeness, contents etc. are discussed in (www.onlinemc.com/onlinemag/SeptOL/rettrg9.html).Search engine DirectHit (www.directhit.com) and Intel's (www.intel.com) search engine uses 'stars' to indicate the quality, while MetaCrawler (www.metacrawler.com) ranks the results by the number of previous hits to the web page [7].

Browsers should support *multiple languages*. Unicode is a 16 bit code which replaces American standard code for information interchange (ASCII) and EBCDIC (8 bit) codes. It provides support for 65535 characters and hence character sets like Kanji, Devanagari and so on, being 16 bit. This helps in achieving portability of digital contents across the globe, and increases the user base who can access the digital library.

Security and authorization can be provided through a hardware lock (practically difficult) or through encryption. When the issues like Intellectual Property Rights (IPR) generate hot debate, the governments are fighting with the hackers world-over, the processor power is increasing, one needs industrial strength encryption algorithms. Plans are afoot to launch Advanced Encryption Standard (AES) from Apr 2001. It uses an encryption key of the size of 128/192/256 bits and data block size of 128 bits. Earlier DES standard used 56bit key and 64bit block.

2.3 Hardware for access

Telecom networks of the future (year 2001-2003) are expected to be the third generation (3G) networks. These networks will be broadband networks, which will provide Internet access on mobile phones, and promises a download speed of 2.4Mbps. The cost of the 3G phone will be a crucial factor in making it popular, and hence the Internet access.

Physically handicapped people – especially blind - can now access the computer through special terminals, enabled with Braille (www.ala.org/editions/openstacks/insidethecovers/ mates/mates_toc.html)

3. Challenges at a glance

3.1 Video information systems

These systems will be available as soon as high bandwidth and powerful servers are in place.

Contents	Video for Video-on-Demand.
User interface	GUI software with functions like Fast Forward, Reverse, Pause, View etc.

Cataloguing	Manual methods are accurate but time consuming. Automatic methods are yet evolving.
Searching	As MPEG-7 is deployed as a compression standard, its content descriptors will provide a natural way to mix descriptors and pixel searches of digital video.
H/W for retrieval	Network needs to give real time response. Has to handle variable/constant bit rates, support multicasting.
S/W for retrieval	'Streaming' helps instant viewing during download. Special software like Windows Media Player and RealPlayer are used.
Data volume	Very high. Generally compressed with MPEG scheme
Copyrights	Through watermarking/embedding a symbol on each frame
Problem area	There are many file formats, and migration from a lower version to higher, or from one format to another, but without loss of information, is a challenge, as the formats are proprietary and not publicly documented.

Some good references on the technicalities of a video library are by Australian digital video library (`www.cmis.csiro.au/DMIS/VideoTalk/VideoTalk.html`) and by [6]

Proceedings of the 3rd conference on Digital Asset Management (6-7 March 2000) in audio/video format are available at `www.ec2.edu/DAM/2000/DAM2K_archive.html` . Dynamic search facility is neither provided nor needed here.

3.1 Image libraries

Contents	Photo, terrain maps, animated image sequences, medical images
User interface	GUI software.
Cataloguing	Manual methods are accurate but time consuming. Automatic methods have shown encouraging results..
Searching	Based on metadata generated in cataloguing. Needs moderate computing power.
H/W for retrieval	Network need not give real time response.
S/W for retrieval	'Streaming' helps instant viewing during download. Software like browsers are sufficient.
Data volume	Very high. Generally compressed with JPEG scheme
Copyrights	Through watermarking on images
Problem area	(1) There are over 35+ graphic file formats. They also undergo upgradation, as also the software that access them. Hence the images need to be converted from lower version (say TIFF-4) to higher version (say TIFF-6), with no guarantee for loss of information. (2) Zooming the images beyond the resolution at which it was created may lead to loss of data/distortion.

Some experiences in creating and managing image libraries of cultural contents and bio-medical contents are given in [3][8].

3.2 Textual contents

Contents	Text – Multilingual or otherwise
Interface	GUI/text based interface software, with multilingual support.
Cataloguing	Manual methods are accurate but time consuming. Automatic methods are yet evolving.
Searching	Boolean searches. Needs low computing power.
H/W for retrieval	Network do not give real time response.
S/W for retrieval	Search engines are available on network. Databases have structured query language (SQL). Efforts are on to interface artificial intelligence with SQL, to optimize the query.
Data volume	Low. Generally compressed with Lempel-Ziv or GunZip etc.
Copyrights	Through special cyber-laws on copyrights.
Problem areas	1) Migration from one format (say *.html) to another (say *.txt) may lead to the loss of structural information. 2) Text files stored under Unix and DOS have different formats.

3.4 How big is "big" ?

How big really are the existing libraries? Elsevier science holds over 700000 digitized articles from Elsevier journals, and expands at a rate of 25000 articles per month (www.elsevier.com). IEEE has around 120+ journals, 600+ conference proceedings and 875+ technical standards digitized. The web site: (www.ala.org/editions/openstacks/beyondthebook/wholelib.html) provides link to 20 largest public libraries of the world, of which the first one viz. The New York Public Library (www.nypc.org) has 105 lakh volumes.

4. Conclusions

Into the digital future, digital libraries will provide compact scalable storage, exhaustive search in shortest possible time, fast retrieval, fast dissemination of information, citation indexing, and distribution of video and audio, in addition to text. One need not wait till digital contents are created. Howsoever a lack of foresight, not anticipating the challenges, and not evolving a strategy, can cripple the digital library on its explosive growth.

5. References

[1] Anthony M. Rutkowski, "Understanding Next Generation Internet", IEEE Communications Magazine, Sep 99, p99-102.

[2] E. Horowitz and S. Sahni, "Fundamentals of Data Structures", Computer Science Press, 1983.

[3] Fred Mintzer, "Developing Digital Libraries of Cultural Contents", IEEE communications Magazine, Jan 1999, p72-78

[4] Gerard J Rego, "Working together in the digitl World", Economic Times, 9/11/2000, p9

[5] Harold S. Stone, “Image Libraries and the Internet” , IEEE communications Magazine, Jan 1999, p99-106

[6] John R Smith, “Digital Video Libraries and The Internet”, IEEE communications Magazine, Jan 1999, p92-97

[7] Steve Lawrence and C. Lee Giles, “Searching the Web: General and Scientific Information Access” ,IEEE communications Magazine, Jan 1999, p116-122

[8] Stephen T. C. Wong and Donny A. Tjandra, “A Digital Library for Biomedical Imaging On The Interne t”, IEEE communications Magazine, Jan 1999, p84-91

[9] SS Hardas and Narendra Gupta, “Into the Digital Future”, CSI Communications, March 2000.