Towards an Information Utility in the New Millineum

V.Rajaraman[*♣]

Summary

An information utility is a distributed repository of a variety of materials such as books, periodicals, news, airline/train schedules, music, video, experimental data, commodity prices etc., of interest to the general public. It is ideally accessible to anyone , any time, anywhere. Such an information utility is now emerging using the internet. It brings with it many new problems of intellectual property rights, security , accessibility, cost and ethics. In this talk we will highlight these issues.

1. Introduction

Traditionally information has been accessed from a variety of libraries. Every University maintains a large library with a diverse collection of books and other materials such as audio tapes, video tapes, microfilms, microfische, etc. Besides a central library, departments maintain their own special libraries of interest to a small group of researchers. The library system is well developed – companies maintain libraries of interest to them, individuals have their own libraries, most cities have public libraries. It has been recognized that access to information is essential in modern civilized society and investment on libraries have grown over the years.

Recently Google (famous for its search engine) has initiated a project to scan and place books of several libraries on their web site which will be accessible to all. There are copyright issues which have been resolved. Out of print books in the library will be put in Google site with arrangements to pay a fee to copyright holders.

Scholarly and other information available in libraries is not the only information people are concerned about. There are a variety of other information people need in their day-to-day life in a complex society. These include government rules and regulations, daily news, up to date information on prices of commodities, shares etc., schedules of public transport, to cite a few.

Hon.Professor, Supercomputer Education & Research Centre
Indian Institute of Science, Bangalore 560 012

The advent of computers half a century ago set in motion a new paradigm of information storage and retrieval. Early researchers worked on methods of classifying information for ease of retrieval in a computer based system. Research was impeded due to non availability of large machine readable corpus of information as disks were of small capacity and manual transcription of information was slow and expensive.

This situation has changed now. There has been a convergence of a number of developments in computer technology in the last five years which has significantly affected the way computers can be used to access information. These developments are:

- Emergence of CDROMs (Compact Disk Read Only Memories) and now DVDROMs (Digital Versatile Disk Read Only Memories) with very high information storage capability. One DVDROM can store upto 7.5 Giga bytes (7.5×10^9 bytes) (To store a typical 500 page book 0.25 Mbytes are needed). The cost of these storage devices is very low, around ten paise per Megabyte.

- Continuous increase in capacity of magnetic disks which can be used for on-line access. Today (2008) desk top PCs have 160GB disks. Storage capacity of disks is doubling every twelve months, at constant price

- Development in computer network technology which has facilitated interconnecting computers not only within the country but also across countries leading to a world wide computer network. Network bandwidths are also doubling almost every 9 months at constant price.

- Wireless technology also rapidly developed. This allows anywhere – any time access to information even when a person is mobile.

- Method of digitizing, compressing and storing text, audio, graphics and video data have continuously improved. Standards have emerged for audio compression, e.g.MP3

♣ [This is a revised (24.11.08) and updated version of an article which appeared in Technorama (Institution of Engineers, India) Vol.49[T], No.3, Dec.1999, pp.4-10]

format, graphics (JPEG) and video data compression(MPEG4). Standards allow easy interchange of these data

- Advent of very powerful processors which can process multimedia information very fast. Processing speeds have been doubling every 18 months at constant price.

- Availability of high resolution video terminals which can display information on multiple windows. Revolutionary developments in display technologies to facilitate easy reading has matured leading to devices such as Kindle of Amazon and Sony book reader which use e-ink technology and are battery driven.

When all the above developments are combined we have a powerful technology to efficiently store multimedia information available in geographically dispersed locations, index them for easy retrieval and access the information from anywhere in the world using a Personal or laptop computer connected to the internet. Even mobile phones can access and display information. These technologies have led to the concept of an *information utility*. In this talk we will answer the following questions:

- What is an information utility?
- What are the unique advantages of a computer based information utility?
- How will such a utility affect our day-to-day work ?
- What is the relevance of these developments to India?

2. What is an information utility?

We attempt to define our concept of an information utility using the analogy of an electrical power utility. In early days of power generation, each city or community had a local generating station which supplied power to the consumers in its immediate vicinity. There was hardly any standardisation. Direct current (DC) was supplied in some cities and alternating current (AC) to others in their neighbourhood. Electrical gadgets could not be used when one moved to a city with a different power supply. Excess generation by a city could not be used by its neighbours. Engineers realised the need for standardization of supply voltages and frequency, need to interconnect generating stations and agreeing on distribution networks and strategies. This led to modern power systems with its attendant advantages of optimization of power generation, fault tolerance, development of a large consumer market for electrical gadgets, cost reduction due to

economy of scale and availability of power to geographically remote areas. Thus a power utility is characterised by

1. Distributed generating stations.
2. Interconnection of generating stations and creation of a distribution network
3. Standardisation of supply to ease access and enable wide use.
4. Regulation of power generation, tariffs, and adherance to standards.

Electrical power system has now become an essential infrastructure for all civilized societies. Using this analogy we can attempt to define the attributes of an information utility as:

1. A variety of information sources are geographically distributed and interconnected by high speed digital links.
2. Access and storage methods are standardised to enable any user connected to the network to access information regardless of its physical location.
3. Regulations are formulated to control storage and access of information and policies on charges for usage.

Currently the internet and the world wide web to some extent satisfy conditions 1 and 2. However, the last attribute is still being debated and there is as yet no consensus.

The main components of an information utility are:

INFORMATION RESOURCE

- **Textual data -** This consists of books and journals and other useful information such as patents, international standards, specifications, etc., stored in a digital form in a computer's disk store. There are two ways of storing this information. One way is to photograph a page and scan the image with a scanner. The scanner digitzes the image storing a 0 for white and 1 for a dark spot. For good resolution one page will be represented by (800×1000) bits (or 100 Kbytes). This form of storage is called a **bit mapped form**. Bit patterns do not carry information for indexing. This is, however, the only practical way of storing old manuscripts, texts and journals. The image of a page may be retrieved and displayed on the video screen of a computer.

The other way of storing a text is to represent each character by its ASCII code. Texts generated using a word processor are already in this form. Most books and journals produced in the past few years will already be in this form. If a page has 6000 characters it will need 6000 bytes of storage. Further, it will be easy to index the document using arbitrary words in the text. If a table has numeric information, the numeric data would be stored in coded form which allows it to be processed. Photographs or other complex figures in the text, however, will have to be scanned and stored as bit maps.

As it requires less storage to store text in ASCII coded form, software is becoming available to scan printed texts using a scanner and convert them to coded form. Conversion by such software is , however, not 100% accurate and manual correction is required before the text is stored. Good conversion software for standard fonts are currently able to give 95 to 98% accuracy. For old texts using non standard or mixed fonts and for hand-written manuscripts such conversion software is not available.

- **Numeric data** consist of tables of various types such as physical property data of various materials, data from experiments, astronomical tables, stock prices etc. Such numeric data stored digitally may be used (if required) by curve fitting programs, spread sheet programs etc.

- **Graphics data** may be photographs, maps, drawings, land records etc. The simplest way of storing such data is to scan the image and store it as a bit pattern. There are better ways of coding and storing maps, drawings etc., which abstract the information contained in them. For example, maps may be stored using longtitude/latitude as coordinates of cities, a linked list depicting road network etc. Data stored in this form eases retrieval.

- **Photographs** (both colour and monochrome) are stored in bit mapped form using compression algorithms to reduce storage space. Formats known as bmp, tif, gif and jpeg are now commonly used.

- **Audio data** is digitized, compressed using a commonly accepted standard compression algorithm (called MP3 format) and stored. Musical scores may also be coded and stored along with the audio data (if required).

- **Video data** requires enormous storage space due to the need for repeating frames atleast 30 times per second. Thus the data is compressed in such a way that when decompressed the original data is recovered. Common standards for compression have been evolved. The current standard is called MPEG-4 (Motion Picture Experts Group - Version 4) and compresses one 90 minute video movie to occupy 7Gbytes.

For details of methods of acquiring, compressing, storing, processing and disseminationof multi-media data one may refer to the book by V.Rajaraman[8].

INDEXING

- **Indexing and interlinking multimedia data** is extremely important for ease of retrieval. Key words in textual documents are selected and linked to related words with logical links by appropriate software. This is called a *hypertext*. For material in other media (audio, video) also, related elements are selected and linked in what is known as *hypermedia*. Such links would allow an user to navigate through multimedia material. For example, from a multimedia encyclopaedia stored in a CDROM one may request information on the Taj Mahal. The computer would search the data and retrieve a page giving textual information about Taj Mahal which would be displayed on the video screen. If there is a reference to music in the text it may link to an audio clip giving a recording of classical music of that time. Links may also be present to video clips on Taj Mahal and related subjects.

LINKING

- The information collection of the utility will normally not be stored in one computer. It will be distributed in many computers known as *servers*. All these servers will be linked by high speed communication links. The fact that the information is distributed need not be known to a user as it is not relevant from his/her point-of-view. A user gets a “seamless” access to the information based on his/her request regardless of its geographical location.

USER

- A user may access information from anywhere using a terminal or a computer, called a client, connected to the network to which the information servers are connected. New types of services which are now popular are music downloads provided by Apple Computers in a hand held device known as Apple iPOD. A large library is available and one may download individual tracks of an album on payment of a fee. Another emerging facility is YouTube (recently acquired by Google) which provides video clips stored by numerous amateurs and professionals for free download.

Amazon has recently introduced a service using an e-book reader called Kindle. Kindle is battery operated, portable and uses e-ink technology which is easy to read. It uses a mobile network to enable users to download books from Amazon's book list and store them locally. The cost of books is a third of print version. One has to buy Kindle which costs around $ 250.

To summarise, the key components of an information utility are:

* A large collection of digitized and compressed multimedia data.
* All data logically linked together and indexed with key words (or elements) to enable easy search and retrieval.
* The data collection is geographically distributed on a computer network.
* Users are geographically distributed and connected to the network.
* Seamless access is available to all "consumers" to the data stored on servers connected to the network.
* Availability of search programs for accessing desired information.

3. Technologies which enabled creation of information utility

Last few years has seen the phenomenon of internet – an interconnected world-wide network of computers. All computers connected to the internet follow a standardized common protocol (a set of rules) called TCP/IP to communicate with one another.

The internet provides facilities to send and receive electronic mail (called e-mail) which is widely used. Internet also supports a file transfer protocol (abbreviated ftp). Directory of files (which may be text, audio, graphics or video) resident in any computer in the network may be searched and a desired file may be selected and transferred to another computer by the ftp program.

Directory of files and their locations (i.e. address of the computer where they are stored) are available in the internet itself.

To allow browsing of information easily on the internet graphical user interfaces (abbreviated GUI and pronounced gooyee) have been developed. For textual information the idea of hypertext is used. In a hypertext key words in each document are highlighted and linked to other documents where the same keywords or related words occur. By moving a mouse to point to a word and clicking it, the GUI allows a user to navigate from one document to another. The documents may reside in any computer on the network. This idea can be extended to graphics, video and audio information also.

A hypertext system used to link information stored on many computers is called the World Wide Web (abbreviated WWW). One can access information in WWW with a program called a *browser* which assists in displaying hypertext documents, identifying hypertext links and retrieving linked files (multimedia). Two of the popular browsers are Fire Fox and Internet Explorer. Thousands of commercial enterprises, newspapers (e.g. The Hindu) magazines (e.g., India Today), organizations and individuals maintain a location with an address (called a home page) on the web. Each page has its own unique web address called URL (Universal Resource Locator). All the web pages are written using a special language (or a notation) known as Hypertext Markup Language (HTML). HTML allows hypermedia links using URLs. The web page of the Indian Institute of Science, for example, has a URL (or web address):

http://www.iisc.ernet.in

In this address http stands for hypertext transfer protocol (as the file transfer on the web follows this protocols (namely, commonly agreed set of rules for data transfer among computers).

As the amount of information on the world wide web is huge (may be several million files) it is essential to have some method of locating the desired page and searching it by using content descriptors. Tools known as *search engines* have been developed and are easily available from the internet itself. A currently popular search engine is called Google.

HTML is a specific implementation of an international standard for defining device-independent, system-independent method for representing texts in electronic form using descriptive markups known as SGML (Standard Generalised Markup Language). Many other subject specific markup languages for a variety of document types such as manuals, books, chemistry and mathematics,

journals etc., are emerging based on SGML. Currently XML is a popular Markup Language. XML allows users to define their own meaningful markups and publicise them separately.

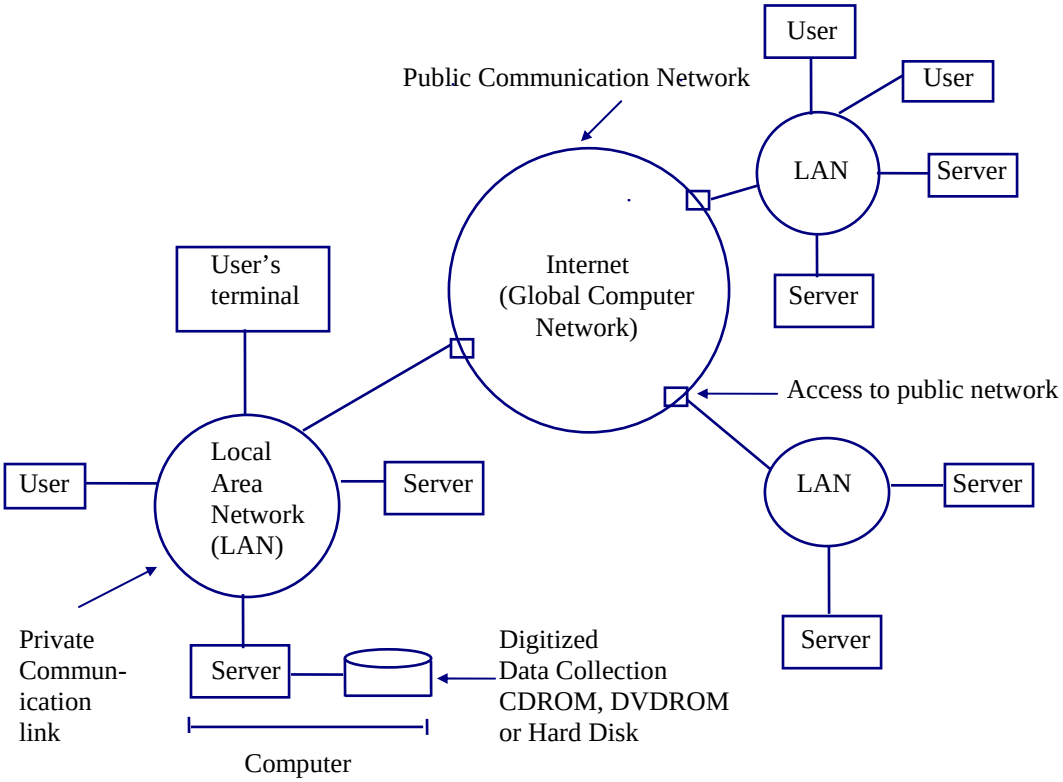

Figure 1. An Information Utility

A schematic diagram of an information utility is shown in Figure 1. We see from the diagram that the internet is an important part of an information utility. The other parts of an information utility are:

Servers which are primarily computers connected to local networks within organizations which have large storage devices. The storage devices are normally hard disks with capacities of 100s of GBs. As reading information from these devices is slow, faster memories are used as buffers. Related information which may be needed by a user are retrieved from disks and stored in semi conductor memory buffer called webcache so that the time to access information is, on the average reduced. The primary purpose of a server is to store indexed, interlinked multimedia data for easy retrieval by search engines.

A variety of servers would be connected to an information utility and maintained by many organization. For instance a tourist office may store local maps, information on hotels, weekly entertainment etc. A Government office may maintain a server with land records, rules and regulations etc. An income tax office may maintain a server with tax rules, necessary forms to file income tax returns etc.

Data Conversion and Compression Hardware particularly for storing audio and video information.

High resolution video displays with multiple windows and good graphical user interface.

The hardware described above is essential but the richness of data stored and indexed along with good search engines are vital for an information utility.

4. Unique advantages of an Information Utility

The fact that all types of data are in a digital form and data is distributed and accessible from anywhere endows some unique advantages to an information utility. We discuss some of them in this section.

Unlike traditional libraries documents are not physically handled by a user. A user views the necessary document and prints the portions of interest in his/her location. Further, many users can simultaneously view a document. Thus documents are not "lost" due to theft or misplacing. Documents do not tear due to wear. Library need not have multiple copies of a popular journal. Rare manuscripts may be allowed to be viewed by many as only images are accessed by users.

Unlike printed text where tables of numbers can be studied but not easily processed, if they are in digital form the numbers in a table can be processed. In print material numbers are "passive" whereas in digital form the same numbers are "alive" as they can be processed and transformed. For instance, a table may be used in a spread sheet program and we can find out how altering some entries change other entries. A user may also try to fit appropriate curves to a set of numbers in a table and give his/her own interpretation of the data.

An information utility can store unconventional information such as readings obtained from some scientific instruments such as spectrometers. All modern instruments incorporate microprocessors

and experimental data is already in digital form. A repository of experimental data can be stored in a server and made available to others to assist in their work.

Searching for information is much easier due to global indexing and use of hypertext and search engines. Fast communication networks allow widespread search for required material.

The ability to digitally store and retrieve multimedia data allows one to effectively provide access to audio and video information. Providing such access is particularly difficult in current libraries due to difficulty in handling such material and wear and tear if many persons access the material. Information Utility would also provide an effective method to preserve rare manuscripts, vintage movies, old music, etc., without denying access to users.

An Information Utility would provide users access to current information which normally takes a long time to reach a traditional library. For example, many authors store their latest research reports in their web page and permit free access. Many conference papers may be found in World Wide Web. An information utility provider can collect articles of interest to a research group and provide it in a local server thereby saving users' time.

The internet allows easy access to discussion groups in specified subjects. Users of an information utility can find researchers working in similar areas and attempt collaborative work. An information scientist can enable this collaboration.

We are also seeing the emergence of journals published only in "electronic form" with no print version. This reduces cost and allows prompt, wide dissemination of the journal. Some journals also provide means for readers to criticise an article and link such review with the article. Such informal review is unorthodox but is very useful for a prospective reader.

Many other types of information not found in a traditional library may be provided by an information utility such as computer aided lessons, lecture demonstrations by musicians, dancers and artists, lecture video of important speeches (such as Nobel lectures) etc.

To summarise the unique features of emerging information utility are :

- Safe storage and multiple access of multimedia data
- Ability to process numerical data published in the literature

- Ability to store variety of data such as audio, video, graphics, data output from experiments, computer aided lessons, lecture demonstrations by artists, famous lectures etc.
- Ability to access information “hot off the press”.
- Accessing information available anywhere in the world any time from anywhere else in the world. This alleviates to some extent the disadvantage felt by persons living in remote areas.
- Ease of search and retrieval.

5. Scientists and Information Utility

The major impact of an information utility on a scientist will be the need to have ready access to such a system if one has to keep up-to-date. As many scientists will place their most recent work on a web page, easy access to it is essential. This implies that a scientist should be connected to the internet and have rapid access to web pages. This necessitates good communication infrastructure in a country. The current situation in India in this respect needs improvement. Recently broadband access is available but the coverage is not universal.

In the last three decades we have seen an acceleration of scientific development and proliferation of publications. The rapidity of change and proliferation of material will be accentuated. A scientist will be deluged with information and has to find a technique of handling this “information overload”. We will see the emergence of “software agents” which can be tailored to satisfy the needs of individual scientists. Such an agent will filter out irrelevant information and seek out relevant information from the system. We will see the widening of gap in the working conditions of scientists with regard to availability of information. The “haves” will have excellent communication and computer facilities with a variety of software agents while the “have nots” will be isolated due to poor communication and consequently non-availability of published material.

Currently a large number of journals insist that articles be submitted in machine readable form to expedite publication of journals. Many journals now have a hard copy form and an electronic form, that is, articles stored digitally in a server and accessible via the internet. Such an access is particularly good for scientists in India as it reduces postal delay. Print form will remain, as it is easy to read. We are also seeing the emergence of purely electronic journals (without a print

version). Pure electronic publications will increase in number as it is cheaper and convenient to publish and scientists will find it increasingly important to have a good internet access to refer to these journals.

Another important issue which will appear will be the charges for use of information. Currently most of the systems are funded by Governments and no charging policy exists. In the long run, however, one may have to pay for accessing and downloading articles. A scientist browsing through many articles may have to pay a substantial amount and this may inhibit browsing. This will particularly affect scientists in India who have limited budgets.

Information providers will have directories of scientists working in similar area all over the world. Access to such directories would promote collaborative work. With improvement in increasing bandwidth of international networks, video conferencing over the network will also be possible. Thus a scientist in India will find it easier to collaborate with scientists elsewhere in the world. Access to electronic discussion groups will also enable scientists in India to know about current trends (topics of research etc.).

One of most interesting things that may happen is the availability in digital form of a variety of experimental results. All sophisticated modern instruments used in experiments employ micro processors and have digital readouts. Thus it is easy to store results of experiments in a digital "server" and provide access to this data to other scientists. Groups of scientists may collaborate to store the data in a repository which may then be used in various research projects. Apart from such experimental data there are other areas where digital access to a large group of scientists would expedite progress in an area. An example is the human genome project in which gene maps are being provided by international groups. Another important development is grid computing in which cooperating organizations (typically Universities) interconnect their computers permitting access from any of the organization to computing power and information resources. Sharing books and journal as well as experimental data and scientific instruments enriches the "virtual organization" created by the grid.

We hasten to add that printed journals and libraries as we know today will remain for quite some time to come. The ease of reading printed material, serendipitous search in traditional libraries and human contacts in libraries cannot be replaced by digital library systems.

6. General Public and Information Utility

So far we have mainly concentrated on information available from library systems which is mostly scholarly information. An information utility will be of great relevance to public at large if it is able to effectively provide information of day-to-day interest to the entire population. Information of this type are: arrival-departure times of trains/planes which are up-to-date, tourist information, information on rules and regulations formulated by state and local Governments on taxation, licensing, copies of judgements, rulings, etc. Providing access to such information is well within the technological capability of our telecommunication and computer systems provided, the required information is selected, codified, authenticated and stored by organizations which own them. The information also needs constant updating. In as much as telephone is not available in all individual's homes such information may be made accessible from public telephone booths which may be expanded as "information kiosks". Such systems called Rural Data Services upto village level is currently developing in many states. Karanatka and Andhra Pradesh have developed these as a public-private partnership. Other states will soon follow. Government of India is providing funds to all the states to set up broad band wide area networks which will reach talukas.

An important problem to be addressed in India is provision of information in local languages. There are very few web sites now which use local script. Technology exists to use local scripts – both hardware solution (GIST card for example) and software solution are available. It is imperative to attempt to organize information provision in at least two languages – English and local language.

Another challenge is the provision of access in remote areas. As of today STD connections exist to almost every town with over 1 lakh population. It is only areas which are remote, in bad terrain, that we have to be concerned about. To connect such places technology exists – wireless transmission of digital information has been perfected. This will allow connecting remote places. For widespread use of an information utility there is need for a "culture change". People must realise the value of timely information and be willing to pay for it. Information providers most appreciate the importance of authenticated, timely information and assume responsibility to provide it.

7. Conclusions

We have seen that the emergence of many technologies - both software and hardware - has led to a revolution in the way information is created, stored and disseminated. Rapid international access to the variety of information distributed across the world will to some extent alleviate problems faced by scientists in the developing world with regard to information availability. There are, however, many problems which need to be resolved before a world wide information utility develops. Some of these are:

- Copyright problem. It is easy to copy digital information. Methods have to be found to prevent illegal copying without inhibiting legal users.

- Material being digitized are in both page image form and ASCII coded form. ASCII form is preferable for wide use of textual matter but technology is not yet available for digitizing archival material to ASCII.

- A disturbing aspect, particularly for developing countries is the rapid obsolescence of both hardware and software. This would put higher financial burden in changing equipment.

- Special precautions need to be taken to prevent corruption of data by vandals bent on mischief.

- Authentication and quality control of information available in the utility is now almost non-existent. This task is a difficult, time consuming and expensive.

- Lastly, we are not used to the idea of paying for information. The question of who pays when information is accessed and how much payment is considered reasonable has to evolve.

References

1. Communications of ACM, U.S.A., Vol. 38, No.4, April 1995 (special issue on Digital Libraries).
2. IEEE Computer, Vol.29, No.5, May 1996 (Special issue on U.S.Digital Library initiative).
3. Fox, Edward, ***Digital Library Source Book***, 1993 (available at web address: http://fox.cs.vt.edu//DLSB.html)
4. Berkeley Digital Library Sunsite (Information on digital libraries) (Web address: http//sunsite.berkeley.edu).
5. Rajasekhar T.B., ***Digital Libraries***, Resonance (Indian Academy of Science Journal on Science Education), Vol.2, No. 4, April 1997
6. Fox, Robert, ***Tommorrow's Library Today***, Communications of ACM, Vol.40, No.1, Jan.1997, pp.20-21
7. "**Towards a World Wide Digital Library**", Communications of ACM, U.S.A., Vol.41, No.4, April 1998,.
8. V.Rajaraman,"**Introduction to Information Technology**" Prentice Hall of India, New Delhi, 2003.