UNIT 4 CONTENT ANALYSIS APPLICATIONS (GENERATION OF INFORMATION SERVICES AND PRODUCTS)

Structure

4.0 OBJECTIVES

The Web is full of different types of sources of information, including primary.

secondary and tertiary sources. These documents have made Internet a kind of information system. In this unit we will look at Internet more as a service delivery system. After going through this unit you will be able to:

- understand Internet as an information system.
- comprehend the general services offered on Internet like email, searching, bulletin board service, discussion forums etc.,
- know the information specific services like bibliographic services, alert service, document delivery service etc., and
- understand the Information Retrieval that takes place over Internet, protocols related to it like Z39.50.

4.1 INTRODUCTION

In this unit, our emphasis will be more on service generation. WWW itself has developed as a huge system. It has many sub-sections in the form of different kinds of documents. Basically, Internet has become a potential medium for generating services particularly on demand as well as in anticipation.

One of the most important aspects which is associated with Internet is retrieval of information. There are many approaches for searching like generating the keyword index out of the text of a document or categorization of the documents. Basically the former method is the most used by the search engines for information retrieval over Internet but by experience we know that it is not entirely satisfactory.

4.2 INTERNET AND INFORMATION SYSTEM

A few years ago, the scenario of information storing started changing and electronic media started being used for storage — for example, the microfiche, diskettes and CD-ROM. Later, Online Information Systems emerged. Electronic data formed the core of the online information system, which uses a basic database to generate various information services. These information systems had definite features in terms of the design and the management of information. There are steps involved in the generation of services using Information systems, such as (1)

- Database design
- Data compilation
- Planning of Services
- Planning of Interfaces
- Incorporation of feedback
- Evaluation and adjustment
- System Management

Internet is today, a huge information system. Only the parameters/components of this information system are not in controlled environment and hence monitoring the various parameters is a problem. Though the various components

of information systems are present in Internet as an information system, there are marked differences:

Library / Information System	Internet
Controlled environment	Uncontrolled environment
Selective	Arbitrary
Purpose oriented scope	Disjointed of and varied scope
Limited	Unlimited
Organised	Chaotic
Time tested methods for organisation and retrieval -- includes automatic and manual methods	Lacks organization and retrieval aspects -- only automatic methods
Permanent	Volatile
Emphasis is on ownership	Emphasis is on access

4.2.1 Internet as Information System Model

Libraries or information systems are defined by their function as systems that comprise of a spectrum of activities from generation and location of information to dissemination and use of it. Control and input by users are of course parameters that keep the model dynamic. (1)

Generation	Storing	Organization	Retrieval	Dissemination	Control

The nature of the activities in the Internet as information system conforms to above model from the stage of generation of information to its use.

Generation

From the generators point of view, publishing information on the net, or *electronic publishing* differs to a large extent when compared to traditional publishing. Just any person who has connection to the Internet and has the tools to make information available can make it a part of Internet. The formalities are minimum with the cost of publishing being negligible because of free tools available, thus opening the doors much wider for information generators to make their contributions to literature. Further, the information generated is not restricted to any geographical boundaries and can reach globally, thus widening its scope. Hence, Information explosion with Internet has gone beyond imagination.

Storing and Access

Internet is not one big centrally located system in which all the documents are stored. It comprises of various machines round the world networked together. The information or documents in these machines are interconnected to the Net and access is given through the various communication channels. The shift is from ownership to access. The storage model for Internet collection is disjoint. However access to information is possible with efficient technology, through protocols like HTTP, Navigation tools and Web browsers.

Several online databases are available on the Internet. These databases store large amount of information and can be accessed and used globally. However, different databases have different web based interfaces. Protocols like, Z39.50 define a standard way for searching more than one database for information retrieval. It makes the usage of large information databases easier by standardizing the procedures and features for searching and retrieving information.

Organization

Libraries use specific organizing techniques for their collection using various classification schemes. Traditional classification schemes aim at facilitating browsing the shelves of the library. Internet does not require such linear arrangement of documents. Here the classification schemes, tools and techniques are needed to be applied more for retrieval of documents.

Classification of Internet Documents

One of the attempts towards classifying Internet documents has been made by the Search engines. Generally for a search request, search engines retrieve a set of documents based on pattern matching like technique. Search engines do not make any attempt to provide context to the search. They only provide a vague kind of subject approach. However Search Engines also follow a kind of classification of the universe of knowledge that constitutes the Internet of their reach. They divide the universe of Internet documents under some areas referred to as 'directories'. This is not the conventional classification of basic subjects as in Library Science but provides some classificatory approach by the most popular areas sought by the users, such as Education, Computers, Entertainment, business, politics, etc. one can look for these categories in any search engine. These are rather arbitrary and have not included the nature of universe of knowledge of subjects and their analysis into different basic subjects. As a result, the subject headings may not be exhaustive or exclusive with respect to the concepts on the Internet and hence makes retrieval inefficient.

Subject Gateways

The experience of library science in classification can be applied to organization of Internet resources. Access to information should be through the subject approach. This is possible through subject gateways. These subject gateways are domain specific and include organization and retrieval approaches. There are today several such services on the Internet in fields such as Medicine, Engineering, Geological studies, etc. These subject gateways maintain information on a given subject and provide access through subject/keywords or resource types. They are updated periodically.

Retrieval

Libraries consist of well-organized collection of information resources. There are definite ways to point out particular information items through time-tested methods in information retrieval such as cataloguing and indexing. These methods are based on well-defined rules for document description and identification.

The retrieval aspect depends to a large extent, on the description of the document and their location. In traditional library science, the catalogue performs this function. In the Internet parliance, the information items are basically 'Internet resources' or most commonly 'homepages' or 'WebPages'. The Internet Resources are peculiar by their content and structure. Efforts are currently being made to identify descriptive elements (also called metadata or data about data), which would be used to adequately describe the Internet resources. The Dublin core suggests metadata elements which helps authors themselves (and not the cataloguer) to describe their documents in a formal way so as to facilitate the search engines for efficient retrieval. Besides, there are models where a

third party catalogues a web page and makes a database of access data. For example, the project NetFirst by OCLC which is a database of more than 1,00,000 records is maintained by skilled OCLC staffs. This project is an extension of project InterCat by OCLC.

Dissemination

If one envisages the Internet as a virtual library, then it follows that many of the services of conventional libraries should be designed and delivered in the 'cyberspace' also. Again, instead of the conventional book, the information is in the form of web pages. Once there are ways and means of identifying and locating relevant information, services can be defined accordingly. In addition to the familiar services, a few others could also be offered using the Internet as the base. To name a few information services:

- Reference Service
- Referral Service
- Webliographies
- BBS/ Discussion forums
- Alert services, announcements
- OPACs
- Newspaper clipping services

The above list is only partial. There can be innumerable such services that may be generated as tremendous amounts of information is available on the Internet and communication channels enable the speedy transfer of data across the globe. Internet based information services are becoming a part of the library's routine work. The potential is unlimited. Familiarity with a few techniques has become a necessity for handling the net resources that would help to a large extent to generate and provide these Internet-based services.

Control

Information centres have various functions that are dynamic in nature. There is a need to monitor the activities from beginning to end to bring efficiency in the services. This involves the concept of 'control', with 'planning and management'. In a virtual library such as the Internet, the aspect of control takes on a new dimension. The virtual library has to co-ordinate the information need and the information items, identify, retrieve and deliver the document just-in-time. This involves a great deal of instantaneous decision-making and application to achieve maximum efficiency of the system. Tips can be taken from the traditional management techniques which can be applied with desired modifications and orientation.

Self Check Exercise

1) Describe the Internet as an Information System.

.....

.....

.....

.....

4.2.2 Services

General Internet Service

Generally the Internet offers the following services:

- Email
- Searching
- Remote Access (Telnet)
- File Transfer (File Transfer Protocol)
- Chatting or Conferencing
- Bulletin Board Systems
- Discussion Forums
- FAQs

Email

Email is the most common facility used over Internet. It provides point-to-point delivery of mails. To use the email facility one needs to have an email account on any Mail Server. There are many public mail servers available which gives free accounts to users. A few most popular ones are, Hotmail (www.hotmail.com), Yahoo (www.mail.yahoo.com), Rediff Mail (www.rediffmail.com) etc. Besides there can be private or organizational Mail Servers based on the individual or organizational needs. A typical example of an email address is,

rama@hotmail.com

userid domain name

An email typically has two parts.

User ID:

A user ID is decided by the user for example '*rama*'. It is an individual account identification for the user.

Domain name:

A Domain name is the name of the server on which the account of the user is created. In the above example it is *hotmail.com*. This part is also known as Host name.

Our postal address on a letter is read from bottom to top; email address is read from right to left. A Mail sever *Daemon* reads Host name. A *Daemon* is a program, which performs certain jobs mostly in the background. All the servers on the Internet are known by their numeric IP address, for example the server of National Informatics Centre of India has numeric IP 209.92.33.165. In a mail the originating server converts the host name *hotmail.com* into the IP numbers as 64.4.44.7, 64.4.43.7, 64.4.52.7, 64.4.53.7, 64.4.54.7. A host can be mapped to a single IP number as well as multiple IP numbers. Once the mail

is reached to destination the destination server first checks for the validity of the user for example, *rama*. Unless the validity is ensured it rejects the email back to originating server.

Searching

Search engines offer searches based on key words. A search engine indexes a fixed amount of bytes or lines to generate the index of terms and stores in its database. But there are search engines, which provide full document indexing, search for example, AllTheWeb (www.alltheweb.com).

Remote Access

Another very common service offered by Internet is remote access. One can access any machine in the world using TELNET, a system communication protocol. In a telnet session typically, a connecting machine sends a request to a server over the network for connection. Once the connection is established a session is started between the requester and server. And the requester works virtually on a sever located at a different geographic location. Most of the servers provides user accounts with authentication requirement for remote access but there are a few which provide guest account temporarily also. Example,

grex.org

arbornet.org

Given below is a typical telnet session:

Option 1:

Go to START menu of computer, select RUN, type

telnet arbornet.org

Option 2:

Go to dos prompt

C:\> telnet arbornet.org

It prompts you to type the login and password, once the authentication check is done, access is given to the user.

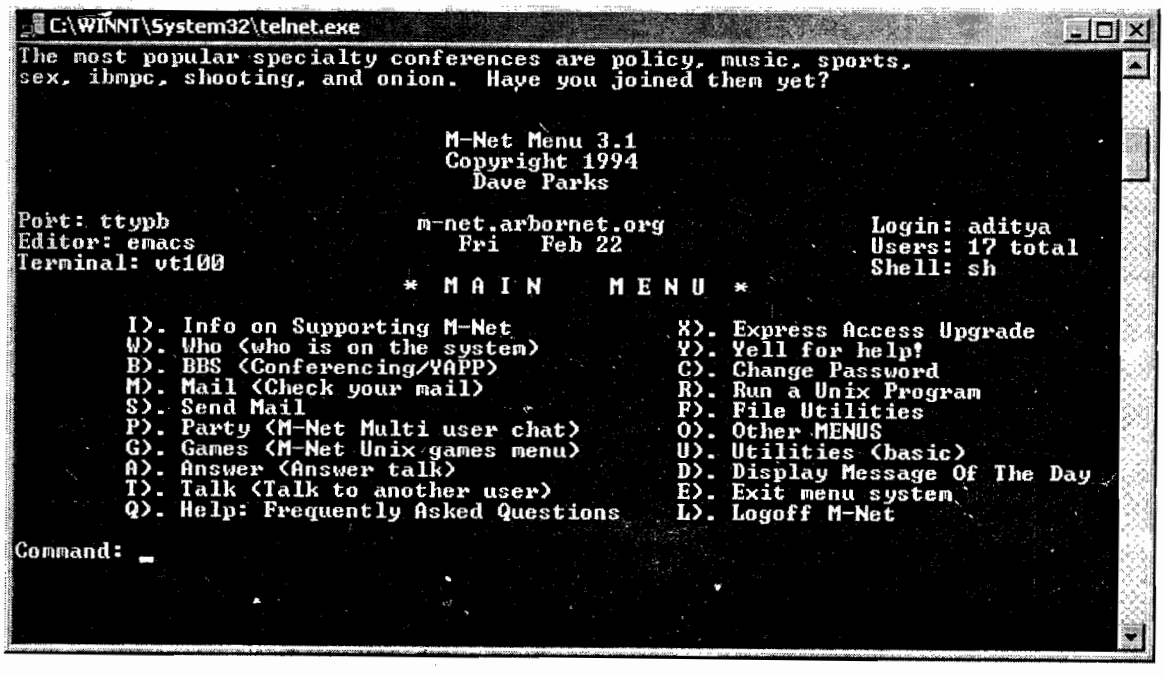

Fig.1: Remote access to arbornet.org

The above mentioned example is of a UNIX based machine which provides a text based remote access. There are services on Internet which provide GUI (Graphical user interface) for remote access. For example, *www.xdrive.com*, *www.tripod.lycos.com*, etc.

File Transfer (File Transfer Protocol)

Internet provides a mechanism for transferring files from one place to another. There are two approaches to it. One approach is; one can send the file as attachment to the mail. The other approach is use of File Transfer Protocol (FTP) to transfer the file. Even when one uses email for sending a file, still he uses FTP. To upload the file on the mail server one uses FTP. Basically what we see on Internet is GUI based program so it is difficult to understand the nature of protocol used. On UNIX machines the same can be done using ftp command, which gives a clear picture on how it works:

Option 1:

Go to Start menu then RUN.

Type ftp sunsite.unc.edu

Option 2:

Type

C:\>ftp Sunsite.unc.edu

It asks for a UserID and password. Once authentication is checked, the user is allowed to upload or download the files. Many servers provide anonymous FTP because these servers keep freely available downloadable software. The common approach is giving *anonymous* as UserID and Emailid (or guest) as password.

For example:

User: anonymous

Password: yourname@domain.com

The mentioned example *sunsite.unc.edu* is an anonymous ftp server.

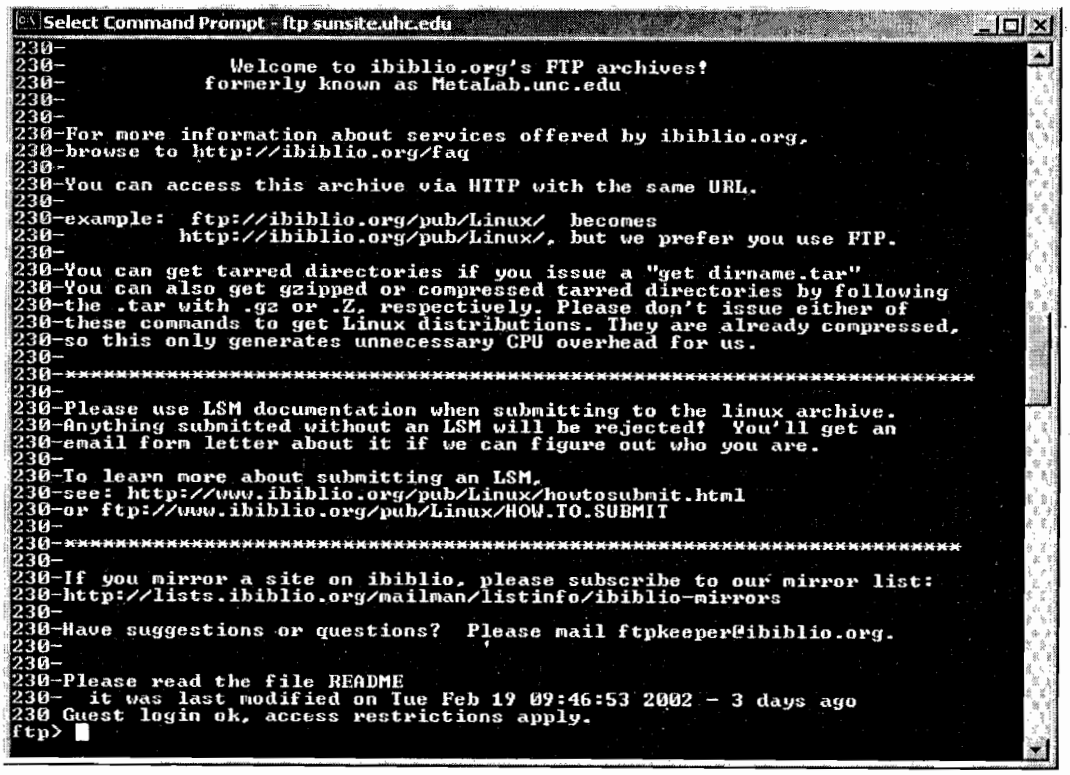

Fig 2.: Using File Transfer Protocol on *ftp://ftp.ibiblio.org*

Chatting and conferencing

One of most important and common services of Internet is interactive chatting. Basically there are two modes of chatting. One is Public chat — for example Yahoo (*www.chat.yahoo.com*). In this mode one enters in the chat room with a chosen ID and chats. Second is a private chat or in other words *off the website* chat where one can download a software (like, Yahoo Messenger or MSN Messenger) and install it on their machine. Then they can add friends to the list and as soon as they login to messenger (a program for interactive chatting) it shows how many friends are currently logged on or if any of the friends of the list logs-in it prompts that so and so has just logged in. One to many chats can be held which is like conferencing.

Bulletin Board System

Bulletin board is a place where information is posted. It is one of the best communication channels over Internet. It is a widely used technique to communicate in an organization. Traditionally there are three types of bulletin boards.

- Chalkboards
- Pinboards
- Magnetic boards

Bulletin boards are a useful way to make information available to a community or a group, to create interest in events, ideas, or products, to create motivation to read, and to display artwork and other paper items of interest. A bulletin board can have information about Advertisements, Artwork, Information about meetings and classes, Maps, Medical or community development information, News items, Photos, Posters, Messages, Stories etc.

Discussion Forum

In a Discussion Forum, people from all over the world participate in *discussions* on any topic in specific areas of interest. You participate in a discussion by reading the messages and responding to them.

If one subscribes to a forum, the header of each new message will be delivered to the subscriber's e-mail box. If he doesn't subscribe, he has to manually check the website as new messages are posted. Another approach is where a mailing list is maintained and as soon as message is created it is delivered to all the participants of the forum.

On Internet these discussion forums appear in the form of threaded messages divided under several subheadings. But these are temporal in nature. Sun Micro-system runs a discussion forum on many of the topics for user-support. One such example is its discussion forum on *StarOffice Suite*.

Frequently Asked Questions (FAQ)

FAQ stands for Frequently Asked Questions. These are the questions about the product in question, answered in anticipation. Usually they are associated with a software or commercial products.

Self Check Exercise

2) What are the general Internet services available?

.....

.....

.....

.....

4.2.3 Generation of Library and Information Services

Libraries offer many services. Some are conventional, some have been very recently developed. Libraries generate different types of information consolidation packages in anticipation as well as on demand. There are steps for developing such products:

- User Study and Criteria Setting
- Evaluation and Selection of Resources and Information
- System design
- Dissemination
- Analysis, Synthesis and Evaluation of Service
- Restructuring of Service
- Packing and Repacking Information in Service
- Feedback

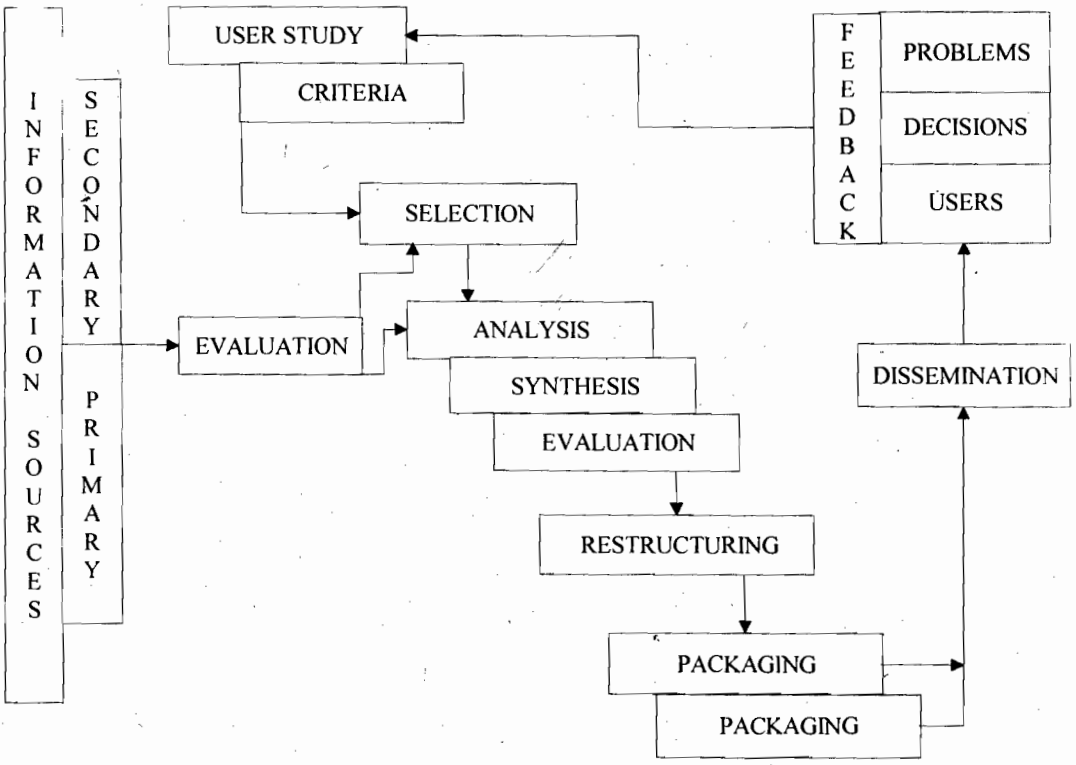

Fig. 3: Generation of Library Services

Library Services

Libraries have a number of bibliographic services as well as document delivery services. Many of the bibliographic services lead to document delivery services. The document delivery involves provision of actual document as well as making a consolidation package. Some consolidation services, which can be generated within the library are as follows:

- Current Awareness Services
  - Current Contents
  - Research in Progress Bulletin
  - News Paper Clipping Service
  - Alert Services
  - Selective Dissemination of Information

Current Awareness Services

This is brought in anticipation or on demand to keep the user abreast of primary information generated in the field. CAS has the following characteristics:

- This service is basically in the form of publications depending on the media. The services can be brought out on Internet or Intranet of organizations.
- Service is generated keeping in the mind a certain domain of users. It does not answer specific query of a particular users.
- The service can deliver bibliographic details. But may lead to actual document delivery.
- This service alerts the users to the recent developments in their field.

Further CAS can be delivered in many forms:

Current contents

A list of articles, which is published in journals acquired by the library is delivered to individual users. A Current Content list contains the name of journals and under each journal the articles appearing in the journal are listed. This service is a periodical exercise by the library. For the sake of convenience the content page of the journal can be photocopied and used. The journals are arranged usually in alphabetical sequence. For example: Uncover service on carl website.

Research-in-Progress Bulletin

This is a kind of alert service, which informs the progress of on going research projects. It is brought out in the form of bulletin. This bulletin contains information about the place at which the project is being done, names of principal and associate investigators and associates, the allocation of funds, duration and so on. It also gives a description about the project and its progress. Research-in-progress services are very common over Internet. These sites give all the above-mentioned information in a nutshell as well as direct links to the website. For example, *www.sourceforge.net*

Newspaper Clipping Service

A variety of information is published in the newspapers. Consolidating the news clippings under the various subheadings is a service known as newspaper clipping service. In a newspaper clipping service a library subscribes to one or more newspapers, carefully chosen for their coverage of areas of interest to the organization of which the library is a part. Each of these newspapers are scanned and any item of news that is considered to be of interest to the user group is clipped and pasted on sheet and in web environment hosted on website. The

clippings can be assigned a subject heading and are arranged accordingly. Many newspapers are today available on Internet. And the dailies maintain archives. Hence an integrated system can be developed to maintain e-clippings in neatly indexed databases with web interfaces.

Selective Dissemination of Information (SDI)

Designed by H. P. Luhn, SDI is one of the most important services a library offers to its users. In an automated SDI service there are several components.

- Document Profile
- User Profile
- A mechanism or program to match the profiles
- User feedback system

Process of SDI:

i) **Selection:**

Expression of user interest is matched with the Documents. Basically the User profile is created with controlled vocabulary. Similarly the document profile is also created with a controlled vocabulary. Both profiles should follow a standard vocabulary. Finally the User profile is matched with document profile. The matched documents are selected to be intimated to the user.

ii) **Notification:**

The selected documents for dissemination have to be notified. The notification list is a form of bibliographic detail about the document with an abstract. This notification list is given to the particular user for further consultation.

iii) **Feedback:**

Once the notification list is delivered to the user, he consults it and finds the relevancy of the documents listed according to his/her needs. If the documents he/she feels are relevant he/she can further consult the document, but if the listed documents are not of his/her interest they send a feedback to the library about their interest.

iv) **Modification:**

The feedback is further used to modify the User's profile. And the same exercise is repeated.

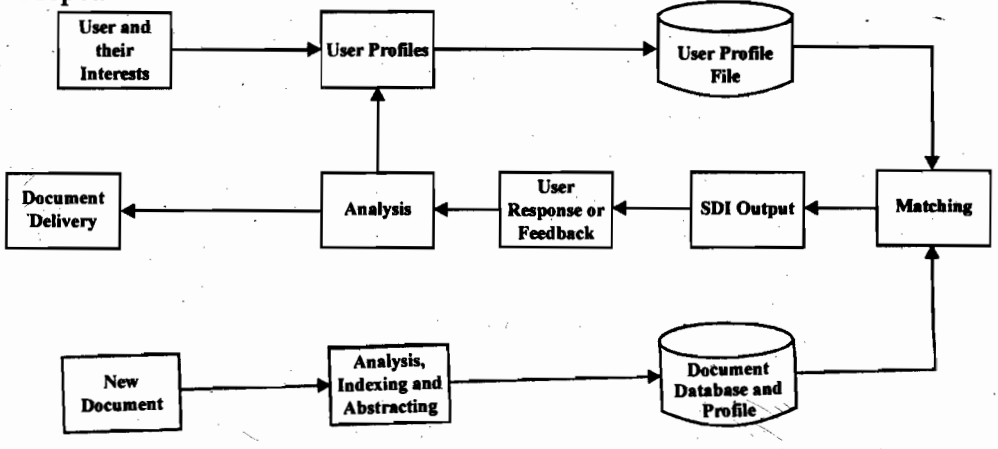

Fig.4: Process of SDI

Self Check Exercise

3) Describe the steps involved in automatic SDI.

.....

.....

.....

.....

4.3 INFORMATION RESOURCE MANAGEMENT

Information Resource Management (IRM) is a very important part of library activity. In fact, the whole library operation is dependent on IRM. It is important to define IRM and examine its scope for the proper understanding of the activities involved. There are various popular definitions, which, put together give a holistic picture. IRM is the management (planning, organization, operations and control) of the resources (human and physical) concerned with the systems support (development, enhancement and maintenance) and the servicing (processing, transformation, distribution, storage and retrieval) of information (data, text, voice, image) for an enterprise. IRM is the recognition by an organization that data and information are valuable resources and the application of the same principles of managing data and information as are used in managing physical resources such as personnel. Information Resources Management (IRM) is the process within the information management arena that serves the corporate interest. IRM seeks to harness information for the benefit of the organization as a whole by exploiting, developing and optimizing information resources. The interests of the organization are usually manifested by its corporate goals and objectives. Thus, IRM is the managerial link that connects corporate information resources with the organization's goals and objectives. IRM is the managing of information resources-a major strategic responsibility of both managerial end users and traditional Information System (IS) management.

There are many definitions in use for IRM, reflecting the various perspectives on the scope and impact of the information resource. Zenona Atkociuniene has derived a definition of IRM by breaking up the term "information resource management" into its component parts.

- Information as something told or items of knowledge
- Resource as a stock or supply that can be drawn on
- Management as the professional administration of business concerns (Hoven 2001)

The information management profession actually consists of a number of largely isolated parts. These include librarians, records managers, archivists, and computer information systems professionals.

4.3.1 Information Resource as an Asset

The area of IRM stresses upon Information as a resource just as any other resource like human resources and financial resources in organizations.

Information is a resource with a final value established according to information quality criteria (novelty, reliability, precision, etc.), potential, and effectiveness of its application. Information Resource Management is the engine that is driving the information economy. It is having and will continue to have a profound impact on business management, competitive advantage, and productivity. Information resource management is an integral part of corporate strategies and can be used by organizations to gain competitive advantages in their markets. IRM and the management of information resources affect all functional areas and all management levels of an organization.

4.3.2 Principles of Information Resource Management

Horton's views expressed at ASLIB meetings for the first time made information and related processes and people prominent. It gave a new professionality to the activities carried out as information resource management. The trend took a definite shape when Nick Willard proposed a model based on traditional resource management principles. The model eventually became known as 'The Willard Model'.

The Willard Model identifies five key elements of IRM:

- **Identification**

The discovery of information resources and the recording of their features in an inventory

- **Ownership**

The establishment of responsibility for the upkeep of an information resource

- **Cost and Value**

Assessment of the cost of an information resource and its value to the organization

- **Development**

The further development of an existing information resource to enhance its value to the organization

- **Exploitation**

The processes which may allow a resource to generate further value through conversion into an asset or a saleable commodity

Based on the theoretical framework of Willard, ASLIB held a series of other workshops to explore each of the above mentioned elements. The IRM network of ASLIB has concluded that the scope and definitions and activities under the denotation of Information Resource Management is totally encompassed by Knowledge Management which is the term used in the latest context in Information arena and a natural extension of the ideas of IRM. Knowledge management will be discussed in UNIT 5 of this Block.

Self Check Exercise

4) What are the principles of Information Resource Management?

.....

.....

.....

.....

4.4 INDEXING

One of the most important tools in information retrieval is indexing. Librarians have been using indexes in their libraries for quite a long time and it is the most reliable source in the library for information retrieval. Besides index there are tools like cataloguing and classification which are heavily used in a traditional library. But with the inception of Internet the role of classification and cataloguing has become limited though they have not lost their value. But indexing has become the chief tool for accessing web resources. Since we are discussing the automated environment we know that we require automatic indexing systems, which can generate indexes without human intervention because of the large number of available web resources.

The objective of Indexing is as follows (Salton and McGill)

- To allow the location of the items dealing with topics of interest of the user
- To relate items to each other, and thus relate the topic areas, in identifying distinct items dealing with similar, or related, topic areas
- To predict the relevance of individual information items to specific information requirements through the use of index terms with well-defined scope and meaning.

If we look at the automatic indexing and manual indexing we find certain differences. We find that manual indexing is more perfect as there are tools for assisting the indexer, like thesaurus in addition to human thinking. Besides there can be scope note of the document that help to identify its sequence and to decide the indexing term. Often the indexer uses some kind of Vocabulary control device (VCD). But when it is an automated environment some of these things are not available. Another problem with automated indexing is that it is very difficult to use multiple terms indexing and the decision of the lead terms. Yet another very serious problem is that of 'anaphora'. Here a noun may be referred successively by pronouns or other equivalent term may be used. In such cases automatic indexing may fail.

Basically, there are two methods of automatic indexing, statistical approach and Artificial Intelligence.

4.4.1 Statistical Technique

It is more of a kind of content analysis where one identifies the terms and finds the number of occurrences of it. The table is prepared with the frequency of

terms and the terms with maximum frequency or minimum are left. One needs to take the decision of what should be the threshold frequency for the cut off point of the terms. The terms having lesser or greater frequency than the threshold should be considered for indexing.

4.4.2 Artificial Intelligence

Another technique is use the of AI. Here rules are developed based on grammar to parse the sentence. The advantage of this technique is, one can define combined indexing terms. A huge dictionary is used to stop the puff words. The progress in the areas of Natural Language Processing (NLP) and Expert Systems have made it possible to design a few automatic indexing system which are capable of syntactic and shallow semantic analysis.

4.5 WEB INDEXING

Web indexing is basically more a secondary information service, where the words are picked up from the first few lines or a few bytes from the file. The terms read are then broken into the single term and stored in a database. Whenever a query is made to the search engine, it searches its' database and fetches the result.

Indexing the Web is not a simple task, and what is evolving to meet the informational needs of Web users are two different kinds of indexing: a back-of-the-book style of hard-coded index links within a Web site and subject trees of reviewed sites. Some organizations are seeing that including indexes on their web sites is just as important as including indexes in books and online manuals.

4.5.1 Back-of-the-Book Style Web Indexing

It is a good idea to give this style of index. Here one can search the alphabetical list of terms which are indexed and then could browse by the term. Many web sites opt to provide a search function for the site. But is this kind of searching, the problem is the same as it used to be with search engines i.e. no relevancy of items found via the search. For example the searching for terms 'Home made' will bring all the pages which contains HOME because it is a word used to direct one to the homepage of site and it will be present in almost all the pages of the site. If there is a site index, you can go directly to the 'H' section, and find the one relevant page, thus saving time. Not only will an index weed out such irrelevant items, but of the many relevant ones. Having sub-headings gives users a clue as to which resources are more likely to answer their questions.

4.5.2 Subject Tree and Reviewed Site Indexes

Some Web search tools review each site and with human intervention decide which categories and keywords fit the site, and then index it accordingly. An example would be Yahoo (*http://www.yahoo.com*), where hordes of people are building indexes to the Web, which is also searchable by a search engine. Another site that reviews the sites listed in their search engine is Magellan (*http://www.maellan.exite.com*). (American Society of Indexer)

Self Check Exercise

5) What is Web Indexing? Describe different kinds of web indexing.

.....

.....

.....

.....

4.6 METADATA AND WEB INDEXING

The META tag in HTML has been used with the goal of giving hints about web page content to search engines. META tag can be misused by putting terms unrelated to the actual content of the page.

In response to it, movements to standardize META tag content have emerged. Corporations and governmental bodies with many web sites often develop a public portal to their web content. They can improve search results for users by the careful use of structured META tags to guide their on-site search engines. Indexers can apply their analysis skills to create these structured tags.

Examples — Digital Object Identifier (DOI), Dublin Core Initiative (DC), Government Information Locator Service (GILS).

Self Check Exercise

6) Write about the roles of metadata schemas.

.....

.....

.....

.....

4.7 SUBJECT HEADING

Traditionally subject headings are given to categorize the documents into different categories. Subject headings are derived for easy access of document. Subject heading is used in cataloguing of document. But in the web parlance manual provision of subject heading is a difficult job. So there must be an automated system which can do this job mechanically. OCLC's Project scorpion aims to derive subject headings using Dewey Decimal classification as a basis.

4.7.1 Project Scorpion

Scorpion is a research project of OCLC. Since subject information is the key to advanced retrieval thus primary focus of Scorpion is building tools for automatic subject recognition based on Dewey Decimal Classification.

In traditional catalog entries the subject portion is most important when it comes to building advanced search and retrieval system. Scorpion assists to automatically assign subject headings or concept domains to electronic items.

Scorpion is a project to combine indexing and cataloging, understanding that these are complementary activities. It cannot replace human cataloging. There

are many aspects of human cataloging that are difficult to automate. However, Scorpion has some idea to produce tools that help to reduce the cost of traditional cataloging by the automating subject assignment when items are available electronically. It assists cataloger to choose the most appropriate subject.

Scorpion overview

The system which maintains the Dewey Decimal Classification electronically isknown as Editorial Support System (ESS). The ESS contains raw information for producing DDC Schedule and Tables. The query is fired against this ESS and ranked results are retrieved. Dewey Decimal Classification is maintained electronically via the Editorial Support System (ESS) at OCLC Forest Press. The corresponding ESS records contain all the raw information used to produce the printed Dewey Decimal Schedules and Tables. ESS records is comprise of a variety of labeled fields. Some or all of these fields can be used to build *ranked retrieval databases* for automatically assigning subject codes to documents. By treating documents as queries against such a database, the result set can be viewed as possible subjects for the document. Figure 5 contains an overview of this process. First, a group of selected ESS records are identified for inclusion in the ranked retrieval database. Then, selected fields from these records are used during the building up process to actually build the database. To automatically assign subjects to an electronic resource, the resource can be turned into a database query with the ranked results being treated as a list of potential subjects for the resource.

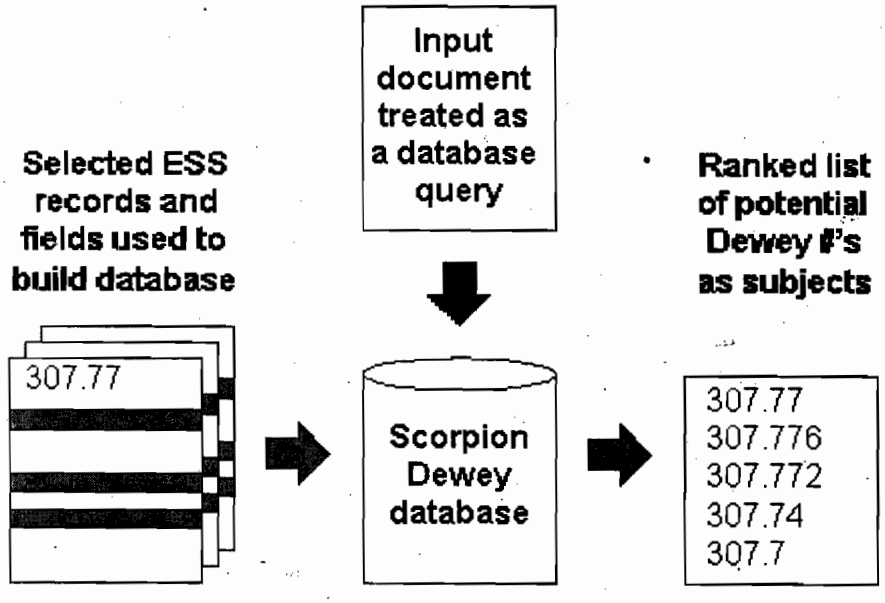

Fig. 5: Simple Scorpion Process Flow (From project Scorpion Website)

4.8 Z39.50

The world wide web has come a long way in providing information, practically rendering many of the Internet services like gopher, wais etc. redundant. Web is user-friendly, allows multimedia components and even interfaces to databases. But http (Hyper Text Transfer Protocol), the protocol of the WWW, is not without drawbacks, when it comes to accessing more than one database using a single interface.

- HTTP does not support the concept of "sessions".
- As it deals with unstructured data, it results in poor indexing and noise in the retrieval.

A "session" is a fixed period when an active connection is restored between client and server.

Most of the bibliographic database managements systems support the notion of a search "session". If it is required, the queries of the earlier search sessions could be reused for any further refined search. It is often the case, that we frame queries depending on the results of the earlier queries. Unfortunately, the HTTP is inherently stateless and presently there is no way of using "session" concept with plain http.

Traditionally database management systems deal with structured data as opposed to editors, word processors, HTML editors that deal with unstructured data. In an ideal situation, the data available on web pages should have been produced by database management system. If the data is structured, practically one can decide which data elements are to be indexed and which data elements need not be indexed, rather than indiscriminately indexing all the content of the web pages (present scenario). With the advent of Common Gateway Interface (CGI) scripting (both server side scripting using PHP, JSP, ASP etc. and client-side scripting using Perl, Python, Tcl etc.) now it is fairly easy to develop web interfaces to a backend database. However, soon the problems become apparent if one wishes to use more than one database in the same site, much worse across internet sites, as the structure of no two databases are expected to be alike. Here comes into the picture the role of Z39.50 standard a protocol for retrieval on the net.

Self Check Exercise

7) Write a short note on Project Scorpion.

.....

.....

.....

.....

4.8.1 History of Z39.50

The Z39.50 (Version 1) standard came into existence along with the OSI (Open Systems Interconnection) model, where 39.50 is an application layer protocol. The Version 1 of this standard was first publised in 1988. It was developed by the National Information Standards Organisation (NISO), an ANSI-accredited standards development body that serves the publishing, library, and information services communities. (8)

In 1992, the second version of Z39.50 came into existence. Around this time a few organizations interested in this standard were involved in a project on Z39.50 Interoperability Testbed project (also called ZIT), with the objective of making many interoperable Z39.50 implementations running over TCP/IP on the Internet. The systems using the second version were primarily host-to-host based systems. The users on one system can search databases on other systems

on the Internet. The main advantage over a simple Telnet session was the ability to use a familiar local system interface.

In 1995, the third version (present version) of Z39.50 appeared. As the World Wide Web was becoming more and more popular, efforts were made to put Web-based gateways onto their Z39.50 host based clients. This version supports the display of holdings information and circulation status. This version also supports access to a variety of data types.

The version 3 is an ANSI/NISO standard and officially referred to as *Information Retrieval (Z39.50): Application Service Definition and Protocol Specification.* Later International Standards organization has adopted the same standard and referred to it as ISO 23950. Presently the standard is maintained by the Z39.50 Maintenance Agency at the Library of Congress.

4.8.2 Basics of Z39.50

A brief introduction of the common terminology used in Z39.50 is presented below(9)

- **Origin:** The source of the Z39.50 request/query. In client-server terminology this would be the client. The origin can be any system on the Internet interested in accessing information on server systems on the Internet.
- **Target:** In client-server terminology this is nothing but a server, which can provide accessibility to a client. Obviously, these servers make their databases accessible to Internet clients.
- **Version:** The information about the Z39.50 version implements in either Z39.50 compliant client or server is important in resolving compatibility issues. The version 1 which was basically WAIS-based, is obsolete. As of today the majority of Z39.50 software conform to Version 2.
- **Gateway:** Gateway means a Web to Z39.50 style implementation. A gateway allows anyone with a Web browser to access Z39.50 compliant databases.
- **Profile:** Profiles provide information on the search attributes like author, title etc. and the types of records that can be returned like USMARC, GRS (Generic Record Syntax), SUTRS (Simple Unstructured Text Record Syntax - text only).

The standard is designed to facilitate interoperability between computer systems whether on Internet or on Intranet. The protocol does not deal with aspects of interaction between user and the origin or target machine. It only deals with interaction between origin and target machine. The essential functions are searching and retrieving information from databases available on multiple hosts. The protocol basically specifies data structures and interchange rules that allow a client machine (origin) to search databases on a server machine (target) and retrieve records.

In a typical implementation of Z39.50, the origin and the target should somehow translate their messages into a common language. Both the origin and target

should be Z39.50 compliant. The origin's search query should the mapped as Z39.50 query. On the target side the Z39.50 query should be mapped as its database query and the results are presented.

The Z39.50 standard does not broadcast searches to multiple servers, but a client can open Z39.50 sessions with multiple servers either sequentially or simultaneously. However, manipulating multiple results from different targets, removing duplicates and presenting retrieved records in a uniform fashion to the end-user are not covered by the protocol.

On the target side, the http/ Z39.50 gateway resides on the web server (like Apache or IIS). In addition, browser-based implementations using either Java or Active X applets reside on the target and are to be downloaded to the user's machine. As no two databases are expected to be alike with regard to the structure (data elements) and searchable fields, it is required to develop a common abstract model of the target databases. The model should contain the abstract data structure (schema) having the data elements like author, title etc. and also the searchable elements as all data elements need not be indexed.

Although Z39.50 is not a database indexing standard, Z39.50 profiles developed for specific communities require a commonly agreed upon database indexing standard. These profiles normally include a minimum set of access points and they should be supported by the database indexes to ensure interoperability between target systems.

The Z39.50 protocol should also help in

- Identifying the characteristics of the server data base
- Locating the databases distributed across the Internet

The protocol provides 'Explain' facility, which provides information in a structured way about the capabilities of the server software, and the characteristics of the information stored in each database on the server. The elements defined by the 'Explain' facility include contact information for the host institution and specifications of the available access points i.e. indexes for searching. This rigid structured information allows the client software to automatically configure itself and adapt to each server system.

To locate the Z39.50 compliant servers, the Z39.50 URLs are useful. These URLs make Z39.50 servers appear like any other document on the Internet. One can collect and classify various Z39.50 servers. The search engines on the Net can be used to identify new Z39.50 servers. However, the standard itself provides 'Locator' facility to do just the same.

Of the many facilities provided by Z39.50, the search and retrieval are the most important. A summary of the major facilities is given in the appendix-1 and 2. A search is performed using one of the pre-defined query-types. For example, Type-1 query can be performed using Boolean operators and Reverse Polish Notation, query Type-101 extends type-1 and allows proximity searches. As the query is performed by the client, the user can choose one or more Z39.50 compliant servers and can express his query in one of the query type. In other words, a typical query can express, "I want to search the OPAC of Library of Congress and I want records where 'Rangnathan' is the author and the title of

the record should contain the word 'classification'". How this query is expressed depends on the user interface i.e. either menu or form based. Many client software allow customization.

The queries are associated with attribute sets. These attribute sets are of two types – domain specific and facility specific. The purpose of attribute sets is to define the abstract database i.e. which include a set of common data elements across databases. As version 3 of the standard is extended to other than bibliographic services these include the following:

Domain Specific:

Bib-1 – Bibliographic

GILS - Government Information Locator Service

STAS - Scientific and Technical

DL - Digital Library Collections

CIMI - Museum Collection Information

GEO - Digital Geospatial Metadata

etc...

Z39.50 facility specific:

CCL-1 - Common Command Language

Exp-1 - for use with an Explain database

Ext-1 - for use with an Extended Service database

The Information Retrieval (IR) protocol is defined within the Open SystemsInterconnection (OSI) framework as an application layer service and protocol. This protocol includes searching and presentation of the retrieved results. The used query presented either in menus or forms is converted into a standardized syntax of this IR protocol.

Another important service provided Z39.50 protocol is the "present" service. This service handles the look and contents of the retrieved records depending of the user specification i.e. what fields are to be displayed or what fields are to be omitted in the output.

In addition to the above facilities, the protocol supports a few important facilities like:

- *Browse*, which allows the client to scan the contents of wordlists or indices on the server. This can be particularly useful in the case of controlled keyword lists or facets.
- *Access control and resource control*, which allow authentication of users, and cost control and online charging for commercial services.
- *Sort*, which allows the client to request different orderings of query results, e.g. relevance ranking, sorting by date or version number, etc.
- *Explain*, which allows the client to interrogate the server about a number of details about its contents and its level of support for the application profile.

- *Item Order*, which allows offline ordering of materials in cases where they cannot be delivered electronically, or where per-unit charging (e.g. online charging) is required. Such services are being supplied in an ad-hoc fashion by online Web-based component repositories such as ASSET. The item order service provides a ready-made, and semantically standardised version of this service.
- *Item Update*, which permits an authorised client to update the contents of the remote database.

4.8.3 Applications

There are a number of potential and existing applications of this standard to libraries:

- Local access to external data sources. The basic search and retrieve functions can be used to extend the number of data sources available for searching at a user workstation. Local and remote databases can be searched using the syntax provided in the local system. This has been the most common implementation of Z39.50 in libraries.
- Creation of virtual or distributed union catalogues. A group of libraries can use the Search and Present services to enable access from a local origin to many targets. In this way, a user on one library can use the syntax and interface of their local system to search catalogues of other systems in the group. With the ILL Protocol, a group of libraries could provide a virtual union catalogue and mechanisms for resource sharing between them. Issues related to this will be discussed later in the paper.
- Copy cataloguing using Z39.50. A local Z39.50 origin can search an external database, specify that the records be presented in MARC syntax, and copy them into their local system for inclusion in a local catalogue. This practice is becoming more widespread.
- Orders for bibliographic outputs. The Extended Services allow a variety of methods to retrieve result sets on a regular basis and have them sent in specified formats. There are a number of possibilities for use of these facilities: SDI services; new and changed records for catalogue purposes; reports for collection development purposes.
- Updating databases. The Update service of the Extended Services can enable simultaneous updating of more than one target by an origin.

4.8.4 Implementations

Europagate

The EUROPAGATE project in the European Libraries Programme is an attempt to build a gateway which will make possible the interoperation of systems based on ANSI Z39.50 (over TCP/IP) and ISO SR (over OSI). In addition to protocol issues, interoperation difficulties arise with respect to data formats, such as national MARC standards and character sets.

GILS

Government Information Locator Service (GILS) is another major project and includes not only the specifications for ANSI/NISO Z39.50, the American

National Standard for Information Retrieval Application Service Definition and Protocol Specification for Open Systems Interconnection (National Information Standards Organization, 1995) in the application but also other aspects of a GILS conformant server that are outside the scope of Z39.50. The GILS Profile provides the specifications for the overall GILS application relating to the GILS Core, which is a subset of all GILS Locator Records, and completely specifies the use of Z39.50 in this application.

4.8.5 Software

Z39.50 Gateway Tools: Libraries and information providers are adopting the Z39.50 information retrieval standard for accessing their on-line catalogues. A Z39.50 to Web Gateway allows users to access these databases using browsers such as Netscape. Alternatively, there are software that work as clients and these can be used instead of web browsers. The search operation usually creates a result set, which is stored on the server and can then be retrieved by the gateway. The features of these gateways include:

1) *Querying* : The ability of the user to specify and submit queries in a search language.
2) *Presentation of results*: The results from the searches are displayed to the user.
3) *Administration*: The setting up and operation of the gateway.
4) *Access and resource control*: Support for authentication and charging for searches.

Following is the list of some gateway tools:

Isite	Unix	http://vinca.cnidr.org/software/Isite/Isite.html
Stanford	Unix	http://lindy.stanford.edu/~harold/z3950/www_gateway.html
WebPAC	IBM AIX	http://www.amlibs.com/product/net/webpac.htm
WebCAT	HP, Solaris OSF-1,AIX	http://www.sirsi.com/webcattoc.html
GeoWeb	AIX, SunOS5.2.x, OSF-1	http://www.geac.com/products/library/geoweb.htm

Z39.50 Client Software: The essential function of any Z39.50 client is to allow the user to search Z39.50 compliant databases. The search operation usually creates a result set, which is stored on the server and can then be retrieved by the client. Some of the client software are:

BookWhere?	Win 3.1, 95	http://www.bookwhere.com/
CanSearch	Win 3.1	http://www.ds.internic.net/z3950/nlc.txt
CIIR's client		ftp://www.usgs.gov/pub/gils/ciir/dtic_a02/
DRAFind	Win 95, NT	http://www.dra.com /products/DRAFIND/DRAFIND.HTM

GeoPac	Win 3.1, 95, NT	http://www.geac.com/products/library/geopac.htm
IrTcl	Unix	http://vinca.cnidr.org/software/Isite/Isite.html
UFO (Fiat lux)	Win 95, NT	http://c134.lib.uci.edu/fiat_lux.htm
Willow	Win 3.1, 95	http://www.washington.edu/willow/
WinPAC	Win 3.1	http://www.als.ameritech.com/winpac.htm
Znavigator	Windows 3.1, 95	http://www.sbu.ac.uk/litc/caselib/software.html

4.8.6 Apprehensions about Z39.50

- It is still under development
- Not widely used
- It is too complex to implement
- It is not required any more as we have web
- It does not work

But

- It is a fairly matured standard
- Fairly widely implemented for LIS work
- Organizations like museums, art galleries, archives have started using it. Latest version supports non-bibliographic information
- It is still useful in web environment. In fact, Web provides access to more than one Z39.50 enabled backend databases
- It promises interoperability across databases
- Supports maintenance of centralized union catalogs.

Self Check Exercise

8) Define the applications of Z39.50 in the library context.

.....

.....

.....

.....

4.9 SUMMARY

Internet conforms to the information systems model. This system offers a number of services for the users, using primary, secondary, tertiary documents and other resources. Over Internet even complex bibliographic services like CAS, Newspaper clippings services, or document delivery services can be offered. But to find out the relevant information one needs to use the searching tools which are highly dependent on the technique of indexing used to index the

system. These indexing techniques basically use automatic methods. For assigning the relevancy to the document many a time statistical approach is used to find the number of occurrence of the particular term.

When it comes to searching by the categorization of the subject, the subject approach is the best method to get the most relevant documents. Scorpion is a project by OCLC which automatically provides the subject heading for a document and categories it for better retrieval.

Z39.50 is a protocol which enables one to access multiple databases in one attempt. It supports sessions for searching.

4.10 ANSWER TO SELF CHECK EXERCISES

1) The basic function of the library and information system is Acquisition, Storage, Processing and Dissemination. Internet also supports all the above mentioned activities which are essential to prove it to be an Information System. Activities supported by Internet are:

   i) Generation
   ii) Storing
   iii) Organisation
   iv) Retrieval
   v) Dissemination
   vi) Control

2) Generally Internet services are:

   - Email
   - Searching
   - Remote Access (Telnet)
   - File Transfer (File Transfer Protocol)
   - Chatting or Conferencing
   - Bulletin Board Systems
   - Discussion Forums
   - FAQs

3) While operating both the manual and automatic SDI have the same steps:

   i) Creation of user profile and document profile. Both profiles are created with a controlled vocabulary,
   ii) Selection: Expression of user interest is matched with the Documents. The matched documents are selected to intimate the user.
   iii) Notification: The selected documents for dissemination have to be notified.

iv) Feedback: Once the notification list is delivered to the user, he sends a feedback to the library about his interest if any modification is needed to be included in his profile.

v) Modification: The feedback is further used to modify the User's profile. And the same exercise is repeated.

4) According to Willard the following are the principles of Information Resource Management:

Identification: The discovery of information resources and the recording of their features in an inventory

Ownership: The establishment of responsibility for the upkeep of an information resource

Cost and Value: Assessment of the cost of an information resource and its value to the organization

Development: The further development of an existing information resource to enhance its value to the organization

Exploitation: The processes which may allow a resource to generate further value through conversion into an asset or a saleable commodity.

5) Web indexing is basically keyword indexing where the terms are picked up from the first few lines or first few bytes from the file. The terms read are then broken into the single terms and stored in a database. Whenever a query is fired to the search engine it searches its database and fetches the result. That is why many a time we end up with missing the links.

Web indexing is of two kinds -back-of-the-book style of hard-coded index links within a Web site and subject trees of reviewed sites.

i) **Back-of-the-Book Style Web Indexing**

Many web sites opt to provide a search function for the site. While this is certainly better than nothing, users encounter the same problems in that scenario as they do in other full-text database searching. The major problem is, of course, relevancy of items found via the search. For example, on a software publisher's site a search for a product called Home Office, ends up retrieving all documents with the word "office" in them, because at the end of every page is the word "home". If there is a site index, you can go directly to the "H" section, and find the one relevant page, thus saving time for other projects. Not only will an index weed out such irrelevant items, but of the many relevant ones, sub-headings well give users a clue as to which are more likely to answer their questions.

ii) **Subject Tree and Reviewed Site Indexes**

Some Web search tools review each site with human eyes and brains to decide which categories and keywords fit the site, and then index it accordingly. An example would be Yahoo (*http://www.yahoo.com*), where hundreds of people are building an index to the Web, which is also searchable by a search engine. Another site that reviews the sites listed in their search engine is Magellan (*http://www.maellan.exite.com*).

6) **Metadata means data about the data**

The META tags are used for better search results. It gives a hint to the search engines about the search terms i.e. to which category it belongs or in other words gives contextual information. Dublin core is one metadata schema that ensures the description of web documents with its 15 elements.

There are various domain specific metadata schemas are available. But use of these metadata schemas is not very common. A typical example of metadata schemas is,

For example, Digital Object Identifier (DOI), Dublin Core Initiative (DC), Government Information Locator Service (GILS).

7) Scorpion is a research project attempting to combine indexing and cataloging, based on the observation that these are complementary activities. Scorpion specifically focuses on building tools for automatic subject recognition by combining library science and information retrieval techniques. For instance, to assign subject codes to a document, the document can be treated as a query against a Dewey Decimal System database using ranked retrieval. The results of the search can then be treated as the subjects of the document. Subject assignment in this manner provides clear differentiation from the traditional computer indexing behind the currently available free search services.

Scorpion cannot replace human cataloging. There are many aspects of human cataloging that are difficult if not impossible to automate. However, Scorpion should produce tools that help reduce the cost of traditional cataloging by automating subject assignment when items are available electronically. For instance, Scorpion could present a list of potential subjects to a human cataloger who could then choose the most appropriate subject.

8) There are a number of potential and existing applications of Z39.50 to libraries:

- Local access to external data sources
- Creation of virtual or distributed union catalogues
- Copy cataloguing using Z39.50
- Orders for bibliographic outputs
- Updating databases.

4.11 KEYWORDS

Boolean query : A query that is a Boolean combination of terms. For example, INFORMATION AND RETRIEVAL, VISION OR SIGHT.

Classification : The process of deciding the appropriate category for a given document.

Collection : A group of documents that a user wishes to get information from.

Document : A piece of information the user may want to retrieve. This could be a text file, a WWW page, a newsgroup posting, a picture, or a sentence from a book.

DOI : Digital Object Identifier, the opaque string used as an identifier by the DOI System.

Entity : Something that is identified.

Indexing : The process of converting a collection into a form suitable for easy search and retrieval.

Information Retrieval : The study of systems for indexing, searching, and recalling data, particularly text or other unstructured forms.

Inverted File : A representation for a collection that is essentially an index. For each word or term that appears in the collection, an inverted file lists each document where it appears.

Metadata : Data that describes something.

Postcoordination of terms : The process of using single terms to describe a document which are then combined (or coordinated) based on a given query. For example, this page may be indexed under the words INFORMATION, RETRIEVAL, and GLOSSARY. We'd then have to combine these terms based on a query like "INFORMATION and RETRIEVAL".

Precision : A standard measure of IR performance, precision is defined as the number of relevant documents retrieved divided by the total number of documents retrieved.

Query : A string of words that characterizes the information that the user seeks. Note that this does not have to be an English language question.

Recall : A standard measure of IR performance, recall is defined as the number of relevant documents retrieved divided by the total number of relevant documents in the collection.

Relevance Feedback : A process of refining the results of a retrieval using a given query. The user indicates which documents from those returned are most relevant to his query.

Relevance : An abstract measure of how well a document satisfies the user's information need.

Robot : Any browser program which follows hypertext links and accesses web pages but is not directly under human control.

Spider : Also called a robot, a program that scans the web looking for URLs.

Stop word : A word such as a preposition or article that has little semantic content. It also refers to words that have a high frequency across a collection.

Term Frequency : Abbreviated as TF, the number of times a particular term occurs in a given document or query.

Term : A single word or concept that occurs in a model for a document or query. It can also refer to words in the original text.

URL (Uniform Resource Locator) : A URL (Uniform Resource Locator) is the address of a file (resource) accessible on the Internet. e.g. *http:/www./ignou.ac.in*

4.12 REFERENCES AND FURTHER READING

Madalli, Devika P. (2002). Tracing the development of Information Resource Management. *Information Retrieval Management*, DRTC, , Paper AA.

What is a bulletin board? *http://www.sil.org/lingualinks/literacy/otherresources/glossaryofliteracyterms/WhatIsABulletinBoard.htm*

Information resources management: topics, concepts, and resources for teaching irm to business students. *http://gise.org/JISE/Vol1-5/INFORM1.htm*

Information resource management: manager of data, information, and knowledge. *http://www.hb.se/bhs/seminar/semdoc/atkociuniene.htm*

IRM Framework. *http://www.irm.org.uk/irmnetpublic/framework.htm*

Salton, Gerard and McGill, Michael J. (1983). Introduction to Modern information retrieval system. McGraw-Hill : New York,.

Indexing the web. *http://www.asindexing.org/site/webndx.shtml*

Z39.50 Maintenance Agency Page. *http://www.loc.gov/z3950/agency/*

Prasad, ARD. A Brief Introduction to Z39.50 Protocol, NACLIN 2002, Hyderabad.

Devika. P. Madalli . Gearing up to the Internet: revamping the LIS curriculum. CALIBER 1999. University of Nagpur, Nagpur

Foskett, AC (1982). Subject Approach to Information, 4th edition, London : Clive Bingley

The scorpion project. *http://orc.rsch.oclc.org:6109/*