KARNATAKA STATE OPEN UNIVERSITY
MUKTHAGANGOTRI, MYSORE –570 006

M.Lib.I.Sc - 3

Master of Library and Information Science

CONTENT ANALYSIS, ORGANIZATION AND DEVELOPMENT

BLOCK – 1

M.Lib.I.Sc – 3

CONTENT ANALYSIS, ORGANIZATION AND DEVELOPMENT

Block 1: Content Organization

Unit -1

Content Organization: Definition and Scope

Unit -2

Significance and Importance of Content Organization in the Printed World and Digital World

Unit -3

Techniques of Content Organization. Classification and Cataloguing of Web Documents: Metadata Schemas, MARC

Unit -4

Document Content Structure and Presentation

INSTRUCTIONAL DESIGN AND EDITORIAL COMMITTEE

COURSE DESIGN

Prof. D. Shivalingaiah — **Chairman**
Vice Chancellor
Karnataka State Open University
Mukthagangotri, Mysuru-570006

Prof. M. Mahadevi — **Convener**
Dean (Academic)
Karnataka State Open University
Mukthagangotri, Mysuru-570006

COURSE COORDINATOR

Shilpa Rani N R
Chairperson
Department of Studies in Library and Information Science
Karnataka State Open University, Mukthagangotri, Mysuru-570006

COURSE EDITORS

Prof. M A Gopinath
Professor (Retd.) in LISc
DRTC, ISI Building, Mysore Road,
Bangalore

Prof. A Y Asuudi
Professor (Retd.) in LISc
Bangalore University
Bangalore

Dr. N. S Harinarayana
Senior Lecturer
Dept. of Library & Information Science
University of Mysore, Mysore -06

Prof. V. G. Talwar
Professor in LISc
Dept. of Library & Information Science
University of Mysore, Mysore -06

COURSE WRITER

Dr. A Neelameghan
Rted. Professor
702, Upstair, III Block, 42 Street
Rajajinagar, Bangalore 560010

BLOCK EDITOR

Dr. N B Pangannaya
Professor of LISc
University of Mysore
Mysore

PUBLISHER

Registrar
Karnataka State Open University
Mukthagangotri, Mysuru-570006

Developed by Academic Section KSOU, Mysore

M.Lib.I.Sc -3 : CONTENT ANALYSIS, ORGANIZATION AND DEVELOPMENT
Block 1: Content Organization

Block Introduction

Content organization is not a new field of research. A lot of studies have concentrated on this area. Why should be worry about content organization? It is basically because the format and presentation will have a direct say in effectiveness of its use. Due to impact of IT and Internet Communication, various kinds of information are now available online. The web designers struggle very hard to find the effective way of organizing their content.

The Block 1 concentrates on 'Content Organization' written by one of the leading personalities in our field. It discusses about the scope of 'Content Development' in the digital environment, with particular reference to developing Internet / Web-enabled information resources and services. A bird's eye view has being given in the area of web information architecture. There are various issues that need to be considered for content organisation and content development. The blocks concentrates on the factors to be taken into account in Content Development, particularly the types Services to be provided, Content Organization and the related formats, and Indexing for fast retrieval. One of the most important concerns for institutions is the effective hosting of information on the net. These issues are also discussed in the block. The relationships between various aspects of information retrieval and its relation to database content and its organization have been dealt in detail. The tools and techniques for accomplishing the different tasks involved in content organization and content development have been discussed in these blocks.

It is hoped that a threadbare analyses of the subject has been provided by one of the excellent exponent in the field. Reading of these blocks would help the students not only for improving their knowledge but also in their professional life as the block provide some very useful inputs that could be directly used in the day-to-day life.

Prof. N.B. Pangannaya

UNIT-1

CONTENT ORGANIZATION

Structure

1.0. OBJECTIVES

After completing this module together with the related study / learning materials and practical exercises a student should able to understand:

- The scope of 'Content Development' in the digital environment,
- Development of Internet / Web-enabled information resources and services;
- Web information architecture;
- Web Servers and Browsers
- Issues in Content Hosting

1.1.CONTENT AND CONTENT DEVELOPMENT

1.1 Introduction to Content : A Basic Component

The generation of various services to users is based on one or more collections of information sources – books, papers, reports, data tables, graphics, etc. on paper, film, or any digital medium.. Within an institution such a collection of information sources may be described in a database – a machine-readable, formally defined, centrally controlled collection of data / information. The data / information records are physically organized and stored so as to facilitate sharing, availability, accessibility, evolvability, and integrity. The content may be textual, graphics, images, sound, or multimedia. It is, therefore, an essential component of all information storage and retrieval system, the generation of information products and services there from, digital library development, library automation systems, etc, In setting up an information service system, the first task is to plan the contents (the database(s)), what it would

consist of, its various attributes, its design, the sources of the data / information, and so on.. Hence the importance of learning and practicing the design and development of databases of various types, processing, organizing and indexing the content for fast and precise retrieval of information; integrating such information resources within the institution through intranets and with those located elsewhere, even at global distances, through extranets, such as, the Internet. The latter aspect raises additional issues in content development and organizing the content for providing information services.

Local or institutional databases may be created and stored in the hard disk of a PC, or ported on to a high density disk (40 MB, 130 MB etc.) or on to CD-ROM (s), on magnetic tape, or a mainframe computer (server) which may be remotely located. The latter may be accessed via a local area network (LAN), wide area network (WAN), and / or the Internet. The design of the databases and the software used may be such that the content can be accessed and / or manipulated by two or more users simultaneously, for example, in a networked environment..

1.2 Content Development

The concepts 'Content Development' and 'Content Organization' are not new in library and information work and service. For information / knowledge management, a variety of tools and techniques have been developed depending upon the availability of content in a physical format at a point of time. However, the subject has acquired new dimensions due to developments in information and communication technology (ICT) – e.g. web resources, multimedia, CD-ROM and convergence of technologies. For example, there have been efforts to develop a range of techniques and tools using automated methods to improve the performance of search engines and

browsers. Such tools have, however, not been able to provide satisfactory indexes to Internet resources to ensure high precision in retrieval. The rapid growth of Internet resources and their increasing use have given a fillip to content development and on the Internet / World Wide Web (WWW or simply the Web).

Internet and web technologies offer several advantages to libraries and information centres to build information resource collections, organize them and deliver information services to their users. The main advantage is the ability to provide integrated access to a variety of information services using web browser as the common interface, from the users' desktop personal computer, laptop, etc. Other advantages include: integration of multi-media content, hardware and software independence, and providing access to information any time as it were, across geographical boundaries. This module will consider: How libraries may organize web content and develop services for delivery over the Internet. In this context we will also examine the factors to be taken into account.

1.2. WEB-BASED INFORMATION

E-mail continues to be important for sending and receiving messages; however, most of the information on the Internet, both static and interactive content, is delivered via the the Web. Therefore, before planning to launch web-based or web-enabled services, it is necessary to understand as to how the Web is organized on the Internet and intranets (networks within an institutional complex). This covers web servers and browsers, web sites and URL (Uniform Resource Locator) and the harware (H/W) and software (S/W) components of a web-site, Web information architecture and organization of information in a website.

1.2.1 Web Information Architecture

1.2.2. Web Servers and Browsers

Web servers store a variety of web compatible documents and provide access to these on the Internet or an intranet

a. **PCs, RISC-based workstations/servers:** These documents are accessed using Web browsers like Netscape and Internet Explorer Records may be stored on Palm tops, Laptops, PCs, workstations, etc.

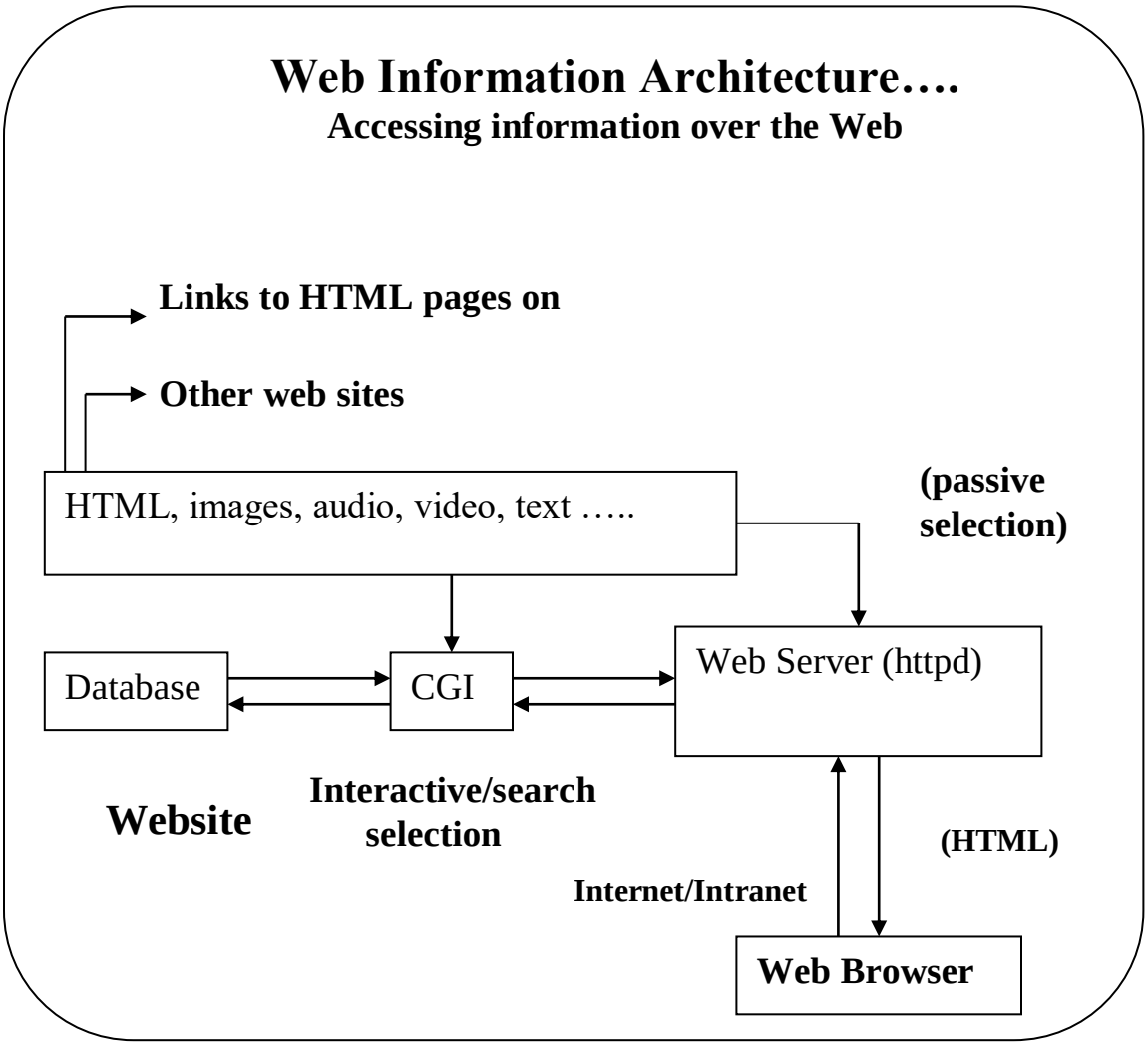
b. **Web sites and URL**

One or more web servers identified with a unique web-site address on the Internet (e.g. www.iisc.ernet.in) Documents available on a Web site are uniquely identified using the URL (UniformResourceLocator)scheme,suchas, ***access protocol://host.domain [:port]/path/file name*** (Ex.: http://www.ncsi.iisc.ernet.in/ncsi/database.html)

c. **Anatomy / Structure of a web-site:**

- HTML pages integrate access to the information
- The pages are organized hierarchically
- Home page (or root page) provides links to second level HTML pages which in turn link to third level HTML pages, and so on. These pages may contain images and provide access to databases through search forms, PDF files, audio, video, etc. or link to documents on other servers.

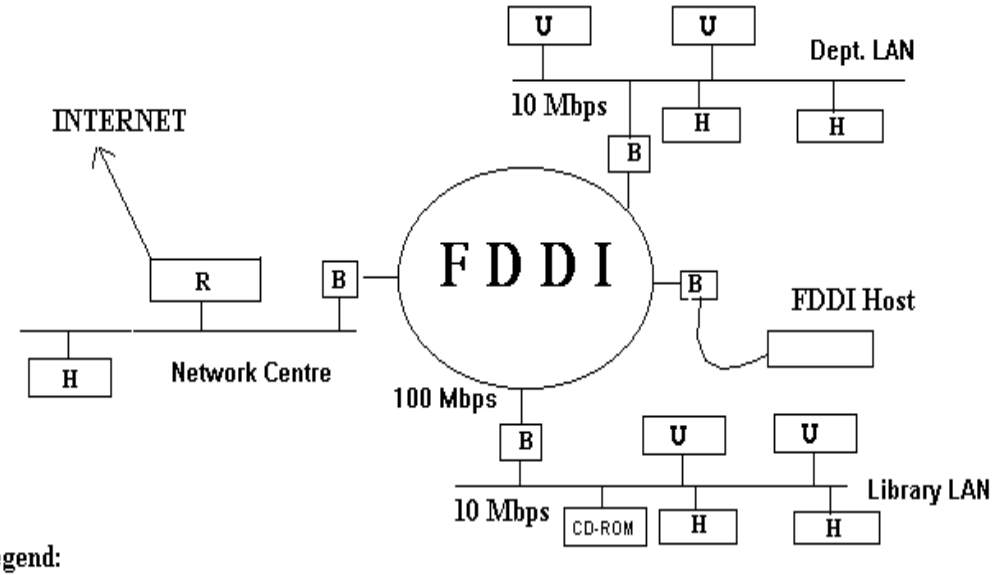
Accessing information on the Web and the links between the components is shown in below diagram



1.3. APPLICATIONS

In planning the content, it's processing and organization, we need to consider the applications (information sources and services) to be delivered to the clientele over the Internet and intranet. The library / information centre needs to have a typical network infrastructure including a LAN

with enough networked computers (e.g. PCs) for users and staff of the centre, and that this LAN is connected to the institutional intranet and to the Internet. The library LAN should also have one or more servers for web, e-mail, CD-ROM, and services.. In such a situation, the library website becomes the integrating factor for tying together all the applications. A library can have applications, such as: information about the library and its services, OPAC (Online Public Access Catalogue), locally owned electronic information sources (e.g. staff publications, reports, manuals), locally hosted, licensed/purchased information sources (e.g. electronic journals, courseware, reference sources), networked CD-ROM databases, remote information sources (e.g. electronic journals, data sets, databases), information sources hosted on web servers within the intranet, profile-based alerting services (push services) (Selective Dissemination of Information / SDI, bulletin boards, listservs, discussion forum, and video conferencing and house keeping operations (e.g. book acquisition, processing, serials management, accounting, etc.).

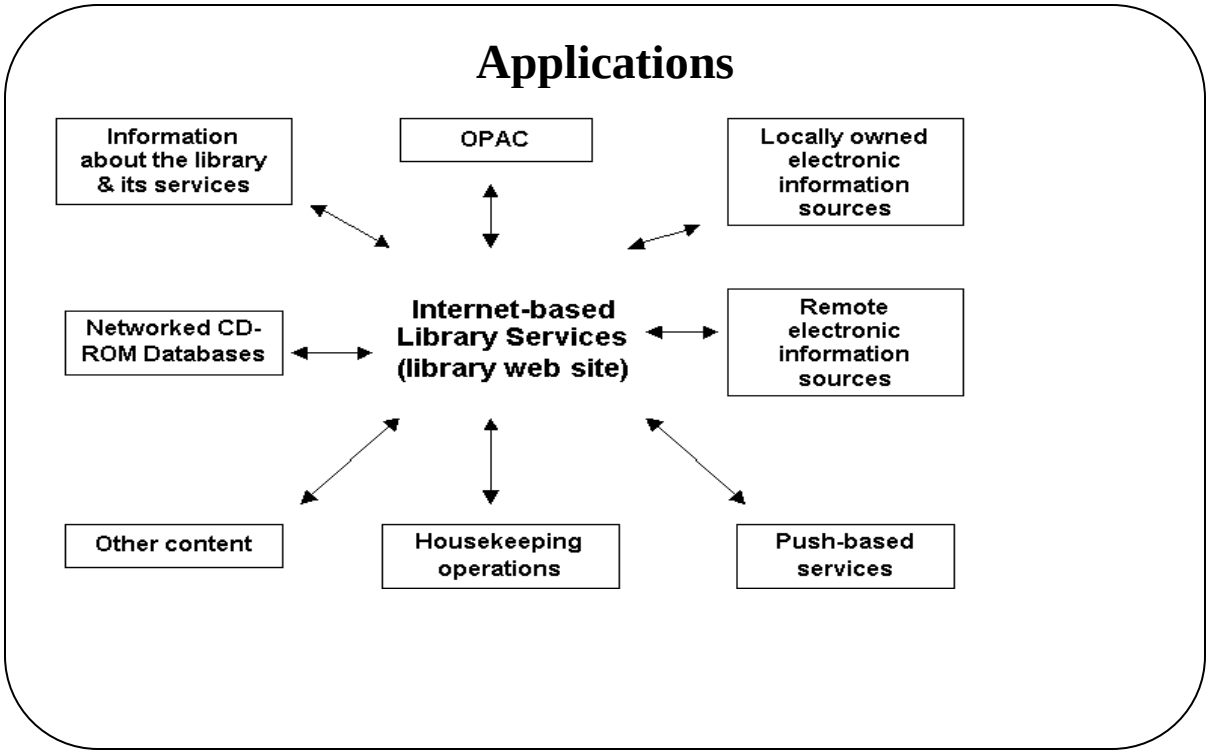


Legend:

R - Router
H - Host/ Server
U - User/ Client
B - Bridge

Servers:
Web, E-Mail, CD-ROM, ERL, etc.

Library LAN Within an institutional Intranet



1.4. ISSUES IN CONTENT HOSTING

The next step is to consider the processes involved in hosting information sources and services (applications) to be provided, the access and delivery desired by and convenient to, the different categories of users. In this connection several key questions need to be answered, including the following:

- How will the content be delivered (ftp, http, telnet, e-mail)?
- If the content is web-enabled, is it browser-aware or does it require plug-ins and helper applications at the user-end?
- Will the content be browsed, searched or both? Is field-based searching required?
- Who is the user (Internet, intranet)?
- What bandwidth is available to the user?

- Granularity of information access – file level, record level?
- How will the content be created (paper to web, electronic to web)?
- What content creation tools are required?
- What security concerns need to be handled for purchased content?

The answers to the issues raised above will be very useful in deciding the content type (running text, images, database, directory, graphics, tables, charts audio, video, and multimedia), storage level (records and files), access modes (browse, search of meta data, abstract, summary, full record, full text), facility for Boolean search (site specific, collection specific); content formats to be handled (HTML, XML, PDF, ASCII, MSWORD (text), GIF, JPEG, BMP (images), WAV, RA (audio), AVL, MOV, MPG (video) etc.); and the type of processing required on these formats for delivering information / record / file to the users' or in print form. Several solutions exist for providing web access to legacy OPACs and CD-ROM databases. Integrating access to remote information sources (free and/or purchased) via the library website also requires careful consideration particularly if there are a large number of remote information sources (e.g. electronic journal). These may require cataloguing using standards like Dublin Core (see below), and providing browse/search access using databases.

There are alternatives for designing email-cum-web solutions for push-based services.

1.5. Check your progress

1. Content analysis is nothing but the

(a) Studying the content of human communication (b) studying the analysis of citations (c) statistical analysis (d) none of the above

2. Information is a trinity consisting of content, conduit, and

(i) value (ii) resource **(iii) context** (iv) power.

3. Web servers store a variety of web compatible documents and provide access to these on the Internet or an intranet (True/ False)

Ans:- True

4. Web servers store a variety of web compatible documents and provide access to these on the Internet or an intranet. (True/ False)

Ans:- True

5. PLONE is a

(a) Content Management System (b) Digital Library Software (c) ILMS (d) Federated Search Engine

1.6. Summary

In designing and developing functional multimedia digital libraries, one faces many barriers. While there are technological and logistical challenges, the major difficulties lie in the serious lack of sufficient large-scale multi-formatted digital contents, and the complex issues related to intellectual property (IP) and copyright. Without a solution to IP and copyright, there will continue to be issues related to the "quality" of contents. This is because of the reluctance of people putting up "quality," "valued," or "treasured" content resources on the web for public access. Thus, even the technology is ready; there are inadequate amount of multimedia digital resources to support the current digital library development. In addition, there is also a serious lack of quality descriptive and annotated information on the available digital resources because of the labor-intensive nature of the work as well as the difficulty in having the involvement of subject specialists.

1.7. Glossary

- OPAC: Online Public Access Catalogue
- URL: Uniform Resource Locator
- LAN: Local Area Network
- WAN : Wide Area Network

1.8.Questions for self study

1. What are the advantages of new artifacts of content?
2. Enumerate the issues in content hosting**.**

1.9 References

1. Martin, James. (1992). *Principles of data-base management.* . 7th Indian reprint. New Delhi: Prentice-Hall of India.

2. Neelameghan, A. (1995). *Online database searching and retrieval: strategies, procedures, commands, and problems*. Bangalore: Sarada Ranganathan Endowment for Library Science' [with tutorial diskette].

3. Neelameghan, A. and Prasad, K.N. *eds*. (2001). *National Seminar on Classification. In*
4. *The Digital Environment. (Papers contributed to the National Seminar on*
5. *Classification in the Digital Environment, Bangalore, 9-11 August 2001*). Bangalore: Sarada Ranganathan Endowment for Library Science; 2001.

6. Raghavan, K.S. and Neelameghan, A. (2002). Composite multimedia works on CD:
    a. catalog entry according to ISBD(ER) and AACR2 revision 1988. *Cataloging &*
7. *Classification* v.33 (3/4); p.193-209.

8. Rajashekar, T.B. (2001). Content organization for Internet-based information services. In:
    a. Content organization in the new millennium (Papers presented to the Seminar
    b. on Content Organization in the New Millennium, Bangalore, 2-4 June 2000 /
    c. edited by A.Neelameghan and K.N. Prasad. p. 51-92.. Bangalore: Sarada
    d. Ranganathan Endowment for Library Science; 2001.

9. Urs, Shalini R. and Raghavan, K.S. (2000). Metadata formats: an overview.
10. *Information Studies*, v.6(2); 2000; p. 111-123.

Unit-2

Significance and Importance of Content Organization in the Printed World and Digital World

---

Structure

2.0.Objectives

2.0.Objectives

After completing this module together with the related study / learning materials and practical exercises a student should able to understand:

- The scope of 'Content organization'
- This unit will explore why attention to the semantic organization of content is particularly important.
- Landmarks and Page Sections
- Paragraphs and Page Elements
- Components of content organization in printed and digital world

2.1Introduction to Content Organization:

A key promise of the web is the removal of barriers to communication and interaction between people, across boundaries, at relatively low cost. The theory is that this promise should be delivered regardless of hardware, software, language, culture or location. However, to truly meet the goal of universal access, the web must be accessible to people with a range of hearing, movement, sight and cognitive abilities.

2.1.1 **Screen Readers**

Screen readers are audio interfaces that—rather than display web content visually for users in a window or screen on the monitor—convert text into audible speech so that users can listen to content. Screen readers present content in a linear manner to users, one word at a time. However,

this is not how sighted users typically view content. Bear in mind that a non-sighted user cannot immediately comprehend the overall layout, style or other meta-level aspects of content. This requires a certain level of mindfulness when organizing content, but also requires some knowledge of how to code for those considerations. There are some types of content that are commonly generated and are a good place to start when considering accessibility, as it relates to semantic web content.

2.1.2 Links

Despite the linear nature of screen readers, there are some ways in which screen-reader users can skim content. One way to do this is to use the TAB key to skip from link to link. This gives the user an idea of where a page links to, and can be a useful way to run through content if the user is looking for a specific link. For this reason, links should make sense when read out of context. Also, any important information pertinent to the link should be at the beginning of the link so that users can hear that information quickly and then decide whether or not to move on.

2.1.3 Headings

Another way for a user to get an overall impression of a page's content is to TAB from heading to heading. This will give a user a general idea of the page's main ideas, and then they can go back to parts of the page that are relevant to their interests. Unfortunately, too many pages lack proper headings that make sense semantically. Content authors should organize content with headings and should not skip from <H1> to <H3>, but instead should be more considerate of the semantic flow and consider the value of going from an <H1> to an <H2> heading.

2.1.4 Landmarks and Page Sections

Screen-reader users can also navigate via ARIA landmarks. ARIA landmarks are attributes that can be added to elements of a page to define areas like the main content or a navigation region. When parts of a page are defined with these attributes, it gives screen-reader users the ability to easily jump from one section to another and know where they are going. For example, we might code our main menu like the following.

```
<div id="navigation">
<ul>
<li>Home</li>
<li>About Us</li>
<li>Contact</li>
</ul>
</div>
```

When screen-reader users encounter something like this, they will hear a collection of links and have to guess, based upon the context, that it is the main menu. However, ARIA landmarks can provide semantic information about the navigation area. By adding the attribute role=”navigation,” and by changing the DIV element to an HTML5 <NAV> element, we are making better sense of this as a navigation region.

```
<nav id="nav" role="navigation">
<ul>
<li>Home</li>
<li>About Us</li>
<li>Contact</li>
</ul>
```

```
</nav>
```

In this case, the screen-reader user will hear something like, “navigation landmark.”

2.1.5 Paragraphs and Page Elements

Visually impaired users can jump from paragraph to paragraph, listening to a sentence or two before moving on to the next paragraph. When possible, place the distinguishing information of a paragraph in the first sentence in the same manner you would with links.

This concludes a high-level overview of some quick wins you can accomplish with accessibility and shows a recommended mindset a developer should possess when generating content. Semantic content matters, not just for search engine optimization, but also to provide universal access to visually impaired users.(1)

2.2Three components of content organization in printed and digital world

As readers and content both go increasingly online, findability becomes an ever greater concern. The organization of content therefore becomes more and more important. Content can’t be effective if it is not found.

There are three components to content organization: **classification, relationships, and stickiness**. Traditionally, we have focused most of our effort on classification as a means of organization, but this needs to change.

a. Stickiness

Content organization is a response to the activity of the information seeker, designed to make that activity more successful.

In other words, we organize content (or anything else) so that **the reader's attempts to find it will be successful**. By this definition, anything we do that makes it more likely that any reader information-seeking behavior will succeed is content organization.

Admittedly, this does turn the usual approach to content organization on its head. Traditionally the approach was: "I have organized this content for you. Now let me teach you how to find stuff according to my organizational scheme." This approach says: "I have been watching how you look for stuff, and I have tried to organize my stuff so that your efforts are more likely to succeed."

If we take this approach — if we **start by looking at how readers seek information** — we will notice that a lot of our reader's information seeking behavior is based on stickiness.

Stickiness is simple. Stick your hand in a jar and the stickiest item in the jar is the one that will most likely come back in your hand. Any way you look for content on the Web, there are usually thousands if not millions of pages that relate to what your are looking for. **The ones you will actually see are the ones that are the stickiest.**

Google essentially works by measuring and recording the stickiness of content. Content gets sticky when people like it and refer to it and read it. Content gets sticky when it is chosen from search results, when it gets liked and tweeted and plused, when it gets referenced or voted up on social networks and Q and A sites. **Readers make content sticky, not writers.** Writers can only make content that is likely to get sticky, and put it where it can begin to pick up stickiness.

Stickiness, then, is a form of organization that you court, rather than create. But stickiness is also, in many ways, **the most sophisticated form of organization**, because it incorporates so many factors, and because it filters out individual judgement in favor of the judgement of the crowd.

b. Relationships

Relationships are all the navigable connections between one piece of content and another. Classification provides one kind of relationship. Even stickiness can provide a kind of relationship, because one piece of sticky content will tend to stick to other pieces of similarly sticky content. But there are relationships between pieces of content that are not a product of either classification or stickiness.

Content, in itself, describes the relationship between things in the world. Content on one subject mentions content on countless other subjects. Every one of those relationships between subjects is a potential relationship between pieces of content. Readers can discover these content relationships for themselves using search, or the author can provide them by creating hypertext links.

It may seem hard to think of hypertext links as a form of organization. For one thing, they tend to be **irregular, capable of pointing off almost anywhere** from almost anywhere in the content. The web of relationships they create often has no obvious order about it. Indeed, a visualization of a web of links often looks completely chaotic.

A top-down visualization of hypertext links looks chaotic, but links can accurately follow the complex and irregular relationships that content actually describes.

But again this is to confuse organization with classification. Classification creates orderly rows and columns. But **the world does not organize itself in rows and columns**. Indeed, the parts of the world that do organize themselves in rows and columns can be described very well by spreadsheets and databases. **Content's domain is precisely those relationships that are not regular**: the bumpy, the lumpy, the unexpected, the counterintuitive, and the downright odd. To confine content to orderly classifications is to rob it of the ability to rule its own domain.

The purpose of organization is to help readers find information, and for the irregular relationships that content describes, and that readers may wish to follow, links fulfill that purpose.

Relationships shine as a means of organization on the Web. Footnotes and cross references provide relationships on paper, but they are so expensive to navigate that their value is minimized. On the Web, on the other hand, relationships are cheap to establish and cheap to navigate. **Relationships therefore play a far greater role on the Web** than they did in the paper world.

c. Classification

That leaves classification. It's what we know best, and often what works best for us to manage the content we create. No wonder we tend to rely on it at the expense of stickiness and relationships. But **genuinely useful classification can be hard to achieve**.

The problem with classification is that its usefulness depends on how well the user understands the classification schema. One of the things our new definition of organization tells us is that if we are going to classify content, we had better observe how the reader classifies things and base

our classification on theirs. Asking them to learn our classification scheme is not going to work unless their motivation is unusually strong.

One of the problems with trying to classify things the way the reader classifies them is that **readers don't usually classify things at all; they usually just name them**. Classification is something you do when you have to manage all the members of a set. If you only deal with a few items from that set, you generally just give them names and ignore the rest.

Another problem with classification is that it gets more and more difficult, both to do and to navigate, as the number and diversity of items increases. The more different things you have to classify, the more arbitrary the classes become. As quantities increase, classes become too large and have to be subdivided, creating classification schemas that are more and more deeply nested and often more and more arbitrary.

In the paper world, classification was by far the most important form of organization, despite its drawbacks, because **stickiness was so hard to measure and relationships were so hard to navigate**.

But **on the Web, stickiness and relationships are far more powerful and far easier to use than classification**. People are navigating much larger volumes of content, so the problems of classification are magnified. And people no longer accept that they have to study the classification schema of a field before they are allowed to look stuff up in that field.

The result is that classification drops to third place for most Web content, behind relationships and stickiness. The problem this poses for writers is that they are more familiar with

classification, and their tools are built more for classification, than they are for stickiness and relationships. This needs to change.

2.3.Check your progress

1.What are Screen Readers?

Ans:- Screen readers are audio interfaces that—rather than display web content visually for users in a window or screen on the monitor—convert text into audible speech so that users can listen to content.

2.TAB key is used to…….

Ans:- skip from link to link

3.Content authors should organize content with headings and should not skip from <H1> to <H3>, but instead should be more considerate of the semantic flow and consider the value of going from an <H1> to an <H2> heading. **(True/ False)**

Ans:- True

4.Three components of content organization are…….

Ans:- Classification, relationships, and stickiness.

5.Objectives of content assessment and organization are to gather a list of the necessary content and to organize that content relative to users needs. **(True/ False)**

Ans:- True

2.4. Summary

The objectives of content assessment and organization are to gather a list of the necessary content and to organize that content relative to audience's needs. This process works "hand in

glove" with the process of defining your Audience. Both these processes require that you have defined the Purpose of your website.

Create a list of all the information sources, services, processes, and other content you offer (or plan to offer) that can be made available through the Web. Eliminate items that don't directly advance the purpose of your site or may not fulfill audience objectives.

At this time it may be a good opportunity to enlist a focus group of your audience to help define and describe your offerings.

1. Assess your service offerings by mapping them to the audience based on their needs.
2. Next, categorize the items in your content inventory according to both user needs and the purpose of your site.
3. For example, if you have content that concerns the graduation process and part of your purpose is to offer that content to your users, then graduation may be a likely category. Continue to group all of your content into their respective categories.
4. After all the content is categorized, organize the content within each category by its relative importance to users. Finally, name each category with a concise and descriptive title. These will become your main "category" links for your Web site.
5. By completing this process you have collected content that satisfies the needs of your target audience, categorized your content into groups that form the foundation for your site structure, and prioritized the relative importance of the content in each category.(3)

2.5. Glossary

TAB key: to skip from link to link

Screen Readers: Screen readers present content in a linear manner to users, one word at a time

ARIA landmarks: ARIA landmarks are attributes that can be added to elements of a page to define areas like the main content or a navigation region.

Relationships: Relationships are all the navigable connections between one piece of content and another.

2.6.Questions for self Study

a. Importance of content organization

b.Three components of content organization

2.7 References

1. "Accessibility-and-the-importance-of-content-organization" accessed from https://ey-intuitive.com/we-see/blog-post/accessibility-and-the-importance-of-content-organization/
2. "Three components of content organization" http://everypageispageone.com/2014/08/11/three-components-of-content-organization/
3. "Content Organization"

   http://warc.calpoly.edu/planning/conceptualization/content_organization.html

Unit-3

TECHNIQUES OF CONTENT ORGANIZATION. CLASSIFICATION AND CATALOGUING OF WEB DOCUMENTS: METADATA SCHEMAS, MARC

Structure

Objectives

3.0. OBJECTIVES

After completing this module together with the related study / learning materials and practical exercises a student should able to understand:

- The factors to be taken into account in Content Development, particularly the types Services to be provided,
- Content Organization and the related formats, and
- Indexing for fast retrieval;
- The major issues in content hosting on the Internet;
- Content / Resource Description

3. INTRODUCTION TO CONTENT ANALYSIS AND DESCRIPTION

Content / Resource Description

S.R. Ranganathan viewed a document as consisting of three aspect, namely, (1) a soul, that is the ideas (thought content) embodied, (2) a subtle body, that is, the mode of exposition (language, form of expression, organization of content, illustrations, style, etc., and (3) a gross body, that is, the medium (recording and carrier) for storing and distribution. This view can apply equally well to digital resources. Most end-users of the resource seek the ideas embodied; however, the mode of exposition and the recording / carrier medium affect the means and method of access and effectiveness of retrieval and use of the information content. These apply even more so to digital contents. According to IFLA's ISBD (ER):

> "Electronic resources consist of materials that are computer-controlled,
>
> including materials that require the use of a peripheral (e.g. s CD-ROM player)

> attached to a computer; the items may or may not be used in the interactive mode. Included are two types of resources: data (information in the form of numbers, letters, graphics, images, and sound, or a combination thereof) and programs (instructions or routines for performing certain tasks including the processing of data). In addition, they may be combined to include electronic data and programs (e.g. online services, interactive multimedia)."

Such digital resources may be in the form of CDs, increasingly being used by libraries in lieu of or in addition to the corresponding print versions(e.g. secondary resources including large databases), or in the form of subscription and/or access to online information sources including Web resources located in remote servers. Some of these resources hold traditional information sources (e.g. text, film, map, sound recordings, etc.) in different formats (e.g. in gif, tif, pdf, doc, txt, wav). Digital resources, however, are not limited to presenting in an alternative format their more conventional counterpart. For example, a record or a file in a CD can be linked to related digital resources on a remote server. The growing abundance of digital resources raises the issue do we need to integrate catalogues of digital resources into the existing information retrieval and access tools, e.g. OPAC in the library? Retrieval tools are becoming more compatible with and capable of handling different metadata. Provision of effective access to digital information sources is complicated by several factors such as the following:

- Digital resources often include multimedia material – optical disc, home page, interactive video disc, Web page, WWW site (IFLA ISBD(ER), 1997). Within each of these categories, there are variations in form, format, and other physical characteristics;

- A digital resource may have several distinct components each one stored separately with 'link' facilities to provide the user a perspective of the whole or of any individual component as may be desired. The chain of links may be long connecting to many subsequent levels. Such an assemblage of resources from several different sources and in different formats is a challenge to the existing techniques of bibliographic control applied to static, print-based collections;
- Media technology is developing rapidly such that by the time appropriate terminology is developed for describing the attributes of a digital resource, the new developments and or changes need to be taken into account;
- Some commonly used digital resources, such as, still and moving images, are difficult to describe for effective organization, indexing and retrieval.

Cataloguing codes and standards have been extended to include a range of newer publication media – microforms, sound and video recordings, films and computer files. Bibliographic standards and the formats associated with them have also been adapted to describe the types of material found in computer networks especially the Internet. However, the objectives of cataloguing formulated by C.A. Cutter over 125 years ago have not continue to be relevant even at present. Nevertheless, it is necessary to extend and rephrase these rules and codes to cover digital resources.

With respect to description or cataloguing of digital resources, the distinction of Work as the ideas embodied (thought content) from the carrier medium is important. The newer carrier media have a significant impact on issues of accessibility, retrieval facility and use of the embodied ideas. For example, two levels of responsibility may be recognized in the development of a CD resource: (1) the responsibility for creating the ideas embodied; and (2) responsibility for

creating the CD product adding value to the accessibility, retrieval, and use of the content of the CD. Similar levels of responsibility can be discerned in the case of Web resources.

The description, analysis, preparation of indexes to and organization of the contents of a collection of information resources / database are crucial to the rapid identification, and efficient retrieval of pertinent information responding to the requirements of the end-user. In manual cataloguing of documents, codes and rules such as the Anglo-American Cataloging Rules (AACR2), the International Standard for Bibliographical description (ISBD) and S.R.Ranganathan's Classified Catallogue Coide are generally used. Recent revisions some of these guides, e.g. AACR2 rev. 1998, ISBD(ER) cover bibliographic description of electronic resources also. Unesco's Common Communication Format (CCF/B), and Library of Congress Machine Readable Catalogue (MARC; MARC21) format are widely used in selecting fields and their tags and for bibliographic description of documents. There are versions of MARC for cataloguing of music sheets and non-bibliographic materials.

A widey accepted bibliographic record structure to facilitate exchange of recrods among databases is !SO2709. The structure of ISO2709 data format is described in the CCF/B manual.

Library and information science professionals are familiar with content analysis of documents for identifying the parameters for cataloguing – e.g. creator(s) of the thought content, title, edition, those responsible for editing, translating etc., publishing, dates and so on related to the documents. Importantly the analysis is needed to determine the subject content, the keywords and descriptors; for preparing abstract, summary and similar records. These are all required in the preparation of index to the sources materials. Subject indexing may be guided by vocabulary management tools, such as, classification schemes, thesauri, subject heading lists, taxonomies, etc

3.1 Metadata

3.1.1Need for Metadata

Search engines are the most widely used devices for retrieving records / information from the Internet. The rapidly growing volume and variety of information sources on the Internet has necessitated the development of automated means for fast search and retrieval of Internet resources thus improving the performance of search engines. These devices are called Web crawlers, Spiders, Robots, Wanderers, Worms, etc. These devices use different methods to navigate and collect information from Web records / documents for indexing. Although these are useful and powerful tools they are not able to provide sophisticated indexes to give the desired level of precision and recall in retrieval. A reason for this inadequacy of the devices is due the fact that the records / information resources on the Internet are not well organized in respect of their structure, content and quality. An index of the available items is an essential requirement to save users' time and network overload. It is in this context that the concept of metadata has emerged, and the associated initiatives and programmes have become significant. Metadata is a critical component in knowledge organization, data mining and information retrieval of Web-based information resources.

3.2.2 Definition

There are several definitions of metadata: "Documentation about documents and objects." (Younger); "Data associated with objects which relieve their potential users of having to have full advance knowledge of the existence of characteristics." (Dempsey and Heery). In its broadest sense, metadata is data about data. In the past library professionals have developed

metadata schema, such as, MARC family of formats, controlled vocabularies and indexing languaes, codes and rules for cataloguing and description of documents. Also database records in a library's OPAC or in an abstracting / indexing service may also be described as metadata. However, the term metadata is more often used to specify records that refer to digital resources on the Internet. This implies that a "metadata record refers to another piece of information capable of existing in a separate physical form from the item it refers to or represents

3.2.3 Uses of Metadata

The main purpose of metadata is to facilitate identification and accessing relevant information resources on the Web in a seamless way and irrespective the type of resource – scientific data, museum object, visual object, text, etc. – and its location.

Additionally, metadata facilitates:

- Documentation of information resource;
- Selection, evaluation and assessment of information resource;
- Improvement of the quality of research results in the sense that it ensures that the resource content is not misrepresented;
- Reuse of content;
- Efficient content development and archiving;
- Protection of intellectual property rights (IPR); and
- Electronic commerce to encode prices, terms of payment etc.

In the ever growing Internet space, effective management of networked information resources will increasingly rely on the creation and effective management of metadata. Metadata may be considered at different levels. "At one end is data used in services, such as, the search engines that support location and discovery of information resources on the Internet/Web. At present metadata may be of limited utility in this area where web-crawlers extract data from information resources. At the other end is metadata that supports much richer functionality. This is associated with research and scholarly activity, requiring special knowledge to create and maintain to meet the requirements of specific domains."

3.2.4 Forms and Characteristics of Metadata

- Metadata is readable by human beings as well as by computers;

- It takes a variety of forms, both general (e.g. Dublin Core / DC) and specialized, and may be part of a larger framework as in Technology Encoding Initiative (TEI). New metadata sets may be developed as the need arises.

3.2.5 Metadata and Traditional Catalogue Data

There are differences between traditional catalogue data and metadata. To understand the nature of metadata records of resources on the Internet we need to examine the characteristics of the Internet's electronic / digital resources which will help us recognize the differences between traditional physical resources held n an institution or electronic / digital resources held on a LAN, CD, etc. Anyone can publish on the Web because there are no controls. Web publishing does not appear to conform to any norms and guidelines. Here are some of the differences:

1. A metadata record usually contains within the record information on the location of the information resource (document), e.g. access information (e.g. access modes – FTP or HTTP), access restrictions (e.g. password) and network address (URL). Such information enables direct delivery of the document using appropriate application software. This is not the same as the location details given in traditional catalogue records. A record in the catalogue of a library (e.g. OPAC) of an institution usually gives location information relating to that institution / library. On the other hand, metadata information may refer to remote locations not related to the institution / library.

2. The Internet/Web environment in which metadata are developed and used change rapidly. The library catalogue environment is much more stable.

3. A Web resource is often accessible in several locations on the Internet, and this may increase rapidly with the growing number of mirror sites. The metadata record thus resembles more like a record in a union catalogue.

4. One and the same document / resource may exist in different format, e.g. ASCII text, PDF, Postscript, etc. Generally metadata formats permit for different versions of the same document to be described in one record, whereas in the hard-print version these may be regarded as different editions / versions.

5. Data on the Internet is relatively transient. Files may be moved around on web servers and the original URL for the resource becomes outdated. Authors may also change and develop documents with an existing URL which means such documents are simply working documents.

6. The level of indexing – granularity – Web resources differs from that of traditional documents catalogued in an OPAC. Questions such as how should web pages be described when a Web document contains separate sections need to be addressed? Should each section be indexed?

For analysis, description and indexing for retrieval of information sources in machine-readable form GSDL, DSpace, TexttAnalyst, WEBSOM and other software packages are available. Some of these (e.g. TextAnalyst) produce summary of one of more text records. The software Picasa is useful for indexing images. The summaries may be used in preparing digests, and executive reports. Quantitative techniques (e.g. statistical analysis) and related software (e.g. SPSS, PolyAnalyst, IDAMS) are helpful in preparing tabulated and graphical presentation of data extracted from texts. Geographical Information System (GIS) is a useful tool to present analyzed data / information in an integrated way text, tabulated data, graphics, and pictures. Large amounts of information can be visualized in a small space.

3.2.6 Metadata Formats

There is no single metadata format that is universally used for describing all types of information sources. There are diffeent types of metadata. A metadata for documentary resources may include the elements found in the catalogue of library materials and of bibliographic databases – e.g. name of author(s), title, edition, imprint, collation, indication of intellectual content, etc. In addition metadata for Internet / web resource description may include information to help client application depending upon the type of users served and conditions of use – e.g. terms of use of the item described, extended documentation about a resource for a researcher. Differences in the metadata may arise from the nature of the Internet resource: resources that are created

temporarily to meet a particular need; these may need only minimal description. Other resources may be valuable either for research and scholarly work or commerically; such resources require detailed description. There may be several information service providers.

Metadata may be embedded within the object / resource it describes, that is, as an intrinsic part of its compostion, or it may exist outside the resource described as a database of metadata records.

Although metadata formats, such as, the MARC family of formats, have been in use much earlier to the advent of the Internet, the latter called for Internet compatible metadata formats. While Dublic Core, a widely used format, was formulated specifically for describing Internet resources, TEI evolved out of the efforts to standardize electronic texts. Efforts are underway to bring together different metadata formats under a common semnatic framework. A mapping between different metadata formats helps in developing programs for inter-format conversion; this would enhance interoperability and searching across databases. Brief descriptions of the formats are given below. For more details on Dublin Core, TEI, and MARC metadata formats please refer to Metadata formats: an Overview / by Shalini R. Urs and K.S. Raghvan. *Information Studies*, v. 6(2); 2000 April; p. 111-123.

Dublin Core (DC):- The Dc metadata standard is being coordinated by OCLC and NCSA. DC consists of a set of fifteeen elements covering the description of a range of network resources. The semantics of the elements have been established by an international inter-disciplinary group of professionals representing library science, computer science, text encoding, museums and other related fields. The decisions are arrived at in the DC workshops. The DC home page is maintained by OCLCat the URL http://purl.oclc.org/metadata/dublincore.

Each DC element is optional and may be repeated. A limited set of qualifiers, that is,

attributes for use to further refine the meaning of an element isavailable. The DC is essentially for application to document-like objects, but may be used to describe similar or cognate entities, such as, profiles of persons, institutions, and research projects. DC is relatively simple, international in scope and may be extended as need arises.

Text Encoding Initiative (TEI):- The objective of the TEI international project is to formulate guidelines for preparation and interchange of electronic texts to support scholarly research and for a broad range of uses by the language industries. The sponsors of TEI are:The Association for Computers and the Humanities (ACH) and the Association of Literary and Linguistic Computing (ALIC), with support from the U.S. National Endowment for the Humanities (NEH), the Commission of the European Communities (CEC-DG-XIII), the Andrew W. Mellon Foundation, and the Social Sciences and Humanities Research Council of Canada.

The TEI covers problems of describing encoded work such that the text itself, its source, encoding, and revisions are properly and completely documented. This is achieved by the Header component of TEI which defines both a core set of elements and additional tag sets that may be used as extensions when necessary. The components of the TEI Header are:

- A file description containing full bibliographic description of the electronic resource including the sources from which the electronic text was derived.
- An encoding description of the relationship between an electronic text and its source(s).
- A text profile, containing classificatory and contextual informqtion about the text, for example, the subject, of the work.
- A history of the revisions.

The TEI Header can become lengthy and complex or quite simple. TEI headsers may also be used as free standing documents.

MARC Format:- The Machine Readable Catalogue (MARC) format, formulated by the Library of U.S. Library of Congress, consists of a family of formats. Started in the 1960s for meeting the opportunities offered by the computerization of library catalogues, the MARC has a longer history than that of the other metadata formats. All odf these have a similar record structure and similar method of tagging data. However, there are differences in their implementation. MARC was opriginally intended mainly to provide a framework for the exchange of catalogue records. The MARC format conforms in structure to the ISO2709 specifications and has four components:

- The Record Label (Leader)
- Directory
- Data Fields
- Record Separator.

The MARC format essentially covers documents, videos, and sound recording at the object level.

Given the fact that almost anyone or any organization can be a publlisher on the Internet, retrieval of information becomes more efficient and convenient if the publishers incorporate metadata into their electronic publications. In this context it is highly desirable to identify and agree upon a minimal set of core metadata elements to be adopted by publishers of electronic documents. Agreement on more elaborate and comprehensive metadata formats is also necessary for detailed descriptions of electronic / digital records on the internet.

5.3 Authority Lists / Files

As already mentioned, cataloguing codes and rules help ensuring consistency of choice and rendering of various elements (e.g. Name of Person. Name of Corporate Body, Name of Meeting, Tirle of Document, Title of Series, Name of Publishers, Name of Country,etc.) in catalogues and databases. Additionally authority lists of such elements, with necessary cross references, can facilitate ensuring consistency. The cataloguer and/or indexerwill be able to consult such lists and use them online. This would minimize the need to type / key-in long and complex names, and hence reduce chances of typing errors. The ISBD manual and CCF manual give a list of codes for languages, and for countries. These can be consulted manually or databases of these can be created and consulted online at the time of data entry in a database.

3.4 Text Mining and Data Mining

3.4.1 Text Mining

It is estimated that over 80 percent of the world's online content is based on text. Text processing and analysis is significantly more difficult than processing and analysis of structured data as in DBMS systems. However, the real challenges and the potential payoffs for an effective universal text solution are perceived as challenging and worth pursuing.

A recent view is that in an organizational context, the knowledge retrieval function can be viewed along two dimensions: a semantic dimension and a collaboration dimension. In the former, linguistic analysis, thesauri, dictionaries, semantic networks, clustering (categorization/table of contents) are used to create an organization's Concept Yellow Pages.

These are used as organizational knowledge maps (conceptual and physical). The techniques consist of both algorithmic and ontology generation and usage. The Collaboration dimension's aims at achieving "value recommendations" identified by experts and advisers to the organization, community building activities, and collaborative filters. Domain experts, who may be accessed across the globe and who hold valuable tacit knowledge can be explicitly identified and consulted for critical decisions.

As in data mining, (see below) text mining adopts analytical methods and their results which are often visual and graphical. Data visualization and information visualization techniques attempt to create an interface that is well suited for human decision-making.

Text mining uses natural language processing (NLP), and deals with diverse and eclectic collections of systems and formats (email, web pages, notes, databases, newsgroups, etc.). It is a cross between information retrieval and artificial intelligence. Besides NLP, text analysis uses indexer or phrase creator, entity extraction, conceptual associations (automatic thesauri), domain-specific knowledge filter (using vocabularies or ontologies), automatic taxonomy creation (clustering), multi-document support and multi-language support. Core text mining analysis can be classified into four main layers: linguistic analysis and NLP, statistical /co-occurrence analysis, statistical and neural networks clustering/categorization, and visualization.

3.4.2 ***Data Mining***

In an institution, the internal data assets and those obtained from external sources and warehoused are processed and analyzed in depth to gain insights on the research object, event or situation. The selected data is then processed, transformed to be ready for the data mining step after which the resulting "knowledge" is interpreted and evaluated. Data mining techniques used have to be specific to the domain and also depend on the area of application. Important requirements are that the data collected should be relevant and of a high-quality.

Some analytical techniques used in data mining include statistical methods, such as, regression analysis, discriminant analysis, factor analysis, principal component analysis, word usage and co-occurrence analysis, and time-series – as well as mathematical modeling. In-depth classification and related indexes are valuable in data mining.

3.5. Check your progress

1.Metadata is divided into

(a) Descriptive, Administrative & Structural

(b) Descriptive,Systematic & Analytical

(c) Logical, Descriptive & Systematic

(d) None of these

2. What is Dublin Core?

(a) Content management tool

(b) E-library software

(c)Metadata standard

(D) Internet Protocol

3. The first set of RDA vocabularies published on the

(i) OAI

(ii) Metadata

(iii) AACR2

(iv) Open Metadata Registry

4. Metadata Dublin core refers to

(A) Data elements in database

(B) Bibliographic elements in Database

(C) Field elements in database

(D) Subject elements in database

5. METS stands for

(A) Machine Encoded Transmission System

(B) Metadata Encoded for Textual Sources

(C) Machine Encoded Textual Standard

(D)Metadata Encoding Transmission Standard

3.6 Summary

Content analysis is a research method for studying documents and communication artifacts, which can be texts of various formats, pictures, audio or video. Social scientists use content analysis to quantify patterns in communication, in a replicable and systematic manner. One of the key advantage of this research method is to analyse social phenomena in a non-invasive way, in contrast to simulating social experiences or collecting survey answers.

Practices and philosophies of content analysis vary between scholarly communities. They all involve systematic reading or observation of texts or artifacts which are assigned labels (sometimes called codes) to indicate the presence of interesting, meaningful patterns. After labeling a large set of media, a researcher is able to statistically estimate the proportions of patterns in the texts, as well as correlations between patterns.

3.7 Glossary

Dublin Core (DC): The Dc metadata standard is being coordinated by OCLC and NCSA

NLP: Natural Language Processing

MARC: Machine Readable Catalogue

TEI: Text Encoding Initiative

ALIC: Association of Literary and Linguistic Computing

3.8.Questions for self study

1. Define the term "Metadata? Mention its uses"
2. Describe Text Mining and Data Mining
3. List out the difference between Metadata and Traditional Catalogue Data

3.9. References

1. Hartley, R.J., Keen, E.M., Large, J.A., and Tedd, L.A. (1990). *Online searching: principles and practice*. London: Bowker-Sauer.

2. Kashyap, M.M. (2005). *Database system: design and development.* Ed. 2. New Delhi: Ess Ess Publications

3. Kini, Srinivas Narasimha and Jacob, K. Paulose. (2000). XML for creating information
4. content on the Internet. *Information Studies*, v.6(4); p. 241-254.

5. Litton, Gerry M.. (1987). *Introduction to database management: a practical approach*. Indian ed. New Delhi: S. Chand & Co.

6. Martin, James. (1992). *Principles of data-base management.* . 7th Indian reprint. New Delhi: Prentice-Hall of India.

7. Neelameghan, A. (1995). *Online database searching and retrieval: strategies, procedures, commands, and problems*. Bangalore: Sarada Ranganathan Endowment for Library Science' [with tutorial diskette].

Appendix 2: Mapping between Metadata Formats

Dublin Core	TEl	MARC
Subject (the topic Addressed by the work)	<encodingDesc> <classDecl> <taxonomy> <ProfileDesc> <textClass> <keywords> <classCode>	560 651 653
Description: A textual description of the content of the resource	<encodingDesc> <projecrDes>	520 505
Title (the name of the object)	<fileDesc> <titleStmt><title>	245
Author/Creator	<fileDesc> <titleS tmt><author>	700 710 711 720
Publisher (Entity responsible for making the Resource available in its present form)	<fileDesc> <publicationStmt> <publisher>	260$b
Other Agent/Contributor (persons such as editors, transcribers, illustrators who have made other	<fileDesc><titleStmt> <sponsor> <funder> <principal>	700 710 711 720

significant intellectual contributions)	<fileDesc><titieStmt> <respStrnt> <resp><name>	
Date	<fileDesc> <publicationStmt> <date>	260$c
Object Type (The genre of the object such as novel, poem or dictionary)	<fileDesc> <notesStrnt><note>	655 516
Form (the physical manifestation of the object such as P9stScript file, PDF	<fileDesc> <notesStmt><note>	856
Identifier (String number used to uniquely identify the object)	<fileDesc> <publicationStmt> <idno>	024 ' 856
Relation (Relationship to other objects)	<fileDesc> <notesS tmt><note>	787 "
Source (Objects - print or Electronic from which this object is derived)	<fileDesc> <sourceDesc> <bibl>	786
Language	<fileDesc> <notesStmt><note>	546 041
Coverage		500$a
Rights Management	<fileDesc> <publicationStmt> <availability>	540 856

UNIT - 4: DOCUMENT CONTENT STRUCTURE AND PRESENTATION

Structure

4.0. OBJECTIVES

After completing this module together with the related study / learning materials and practical exercises a student should able to understand:

a) Various aspects of information retrieval and its relation to database content and its organization; and
b) The use of some of the important tools and techniques for accomplishing the different tasks.
c) To introduce the Principal Stages of the Search in online databases
d) To enumerate the Index-based search strategy
e) To understand the Search Language

4. INFORMATION RETRIEVAL

4.1 Search Strategy

The objective of a search and the associated search strategy is to obtain the best search results. It means that the retrieved data / information / record (s) match with the inquirer's information requirement as closely and completely as possible. All the records and/or information in the information system that meet the user's interests at the moment are retrieved, and no irrelevant information is retrieved. The result should be available when the user needs it, in the form and format he/she is comfortable with, and with minimal effort and cost on his / her part.

A search in a database (information resource collection) is an interactive, iterative and heuristic process. It is often a trial and error exercise in which the searcher in a conversational or dialogue mode with the system and, if necessary and possible, with the participation of the end-user, attempts at successively refining and reformulating the search strategy and the initial search expression on the basis of responses of the system and end-user's reactions to the intermediate search results.

Searching for and retrieval of records / information from a database may be done by the end-user directly, alternatively an information professional – an intermediary, such as, a reference librarian – may do the search and retrieval operations for the end-user. In either

case the success of an online search for and retrieval of, information from a database can be affected by several factors. In a large measure it also depends on the searcher's understanding of the search languages and of the characteristics of the databases accessed and his/her ability to interact with the system, to formulate an appropriate search expression in the search language of the system so as to precisely represent the information need. If an information intermediary is involved in the search, then he/she should be able to effectively interact with the end-user.

4.2 Principal Stages of the Search

The principal stages in online database search may be summarized as follows"

- Recognition of an information need and defining that need by the end-user in terms of a specific subject, type of information desired (e.g. bibliographical references with or without abstract, numerical data, articles in particular language(s) only, whether full-text will be required, translation requirement, if any, to be kept up-dated, etc.), how soon the information is needed and other relevant particulars.
- Communication of the information need to the database service centre / library (in person, by letter, e-mail, telephone, or through another person), especially if the search is to be done by an intermediary.
- Recording the search request in a Search Record form (online or printed form).
- If necessary and if possible, the information intermediary should arrange for a discussion with the end-user for specifying as precisely as possible, the various aspects of the end-user's information need. Aids, such as, a scheme for classification, subject headings list, thesaurus, etc. associated covering the subject of the query, known documents or a specialist in the subject of the query can help in the interview with the user. Display of pertinent parts of these aids can be done online for the end-user and/or searcher to view.

- Selection of appropriate database(s).
- Formulation of the query in the search language of the database and the software used. This implies adoption of search strategies and expressions appropriate to structure, organization, search language and capabilities of the system. Vocabulary management tools, such as, thesaurus, taxonomies classification scheme, subject heading lists, subject map, etc. associated with the database(s0 to be searched can assist in this step. If possible, the end-user should be enabled to participate at this stage to obtain satisfactory results.

4.3 Search Operation

1. Fast access files, such as inverted / index file(s) created by the system should be used. The **Any File** of CDS-ISIS is also a fast access file. The end-user can assist in the evaluation of the initial / intermediate search results and provide feedback.

2. If necessary, modification of search strategy and refinement of the search expression on the basis of the successive end-user feedback on the successive intermediate retrieval results.

3. Selecting the most relevant references / abstracts and arranging for copies of the documents as desired by end-user.

4. Recording / logging the query (e.g. in a Search Request Form), the search procedure adopted and the results obtained for future use and analytical studies.

Guinchat and Menou (1990, p. 224) present a schematic representation of the search procedure.

4.4 Search Language

4.4.1 General Features

The documents represented by surrogates (e.g. catalogue entries, bibliographic citations) in the database are assigned index terms preferably selected from a related vocabulary tool, such as, a thesaurus, classification scheme, subject headings list, Wordnet, subject map, etc. Often these tools indicate the nature of relations (hierarchic and non-hierarchic associative relation) among concept / terms. Such tools can not only aid in the indexing of documents at the cataloguing / data entry stage but also in the correct specification and representation of user's query and in evaluating and refining the search expression to begin with and on the basis of evaluation of intermediate search results as mentioned earlier. The index terms should be selected by a trained indexer and / or subject specialist. The vocabulary tools can also be used online while entering data in a database or formulating search expressions for searching in the database. Terms in the search expression are matched with the index terms assigned to document surrogates to effect retrieval of relevant records. The retrieved references are deemed to represent documents pertinent to the end-user's query and those not retrieved are deemed not pertinent.

Database systems generally permit the following types of search:

1. **Index-based search strategy**
2. **Display the index and select appropriate terms**

   1.2. Enter the search terms (without selecting them from the index) but the system will scan the index for matching terms, and if found, will select them for search in the database.

   In the index-based search, much like looking up the index to a book, the system will directly identify the matching documents, if any (random access)

3. **Free-text search** (not using the index) strategy. Here the search will be sequential, record by record from the first to the last, much like going through page by page of a document to find the match. Hence free text search takes relatively a longer time...

4. **Combination of (1) and (2)** strategy

Generally the terms in the fields frequently searched by users are indexed (Canon of Recall Value), e.g in a bibliographic database, Name of author, Title words, Series title and/or words from the title of the series, Subject descriptors, Keywords, etc. Search by date of publication, name of publisher, place of publication, physical description etc. may be required only very occasionally. Hence contents of such fields may not be indexed. However, should the need arise, search in any one of these fields can be carried out using free text search. For example, all documents published in 2002; all documents in microfilm form, etc.

Combination of index-based search and free-text search is useful in responding to queries such as the following:

Documents by C.R. Rao published during 1999 and 2002

In this case the author field will normally be indexed but not the date of publication field.

Videocassettes on 'Environmental pollution' .

In this case the subject descriptor field and / or keywords fields are usually indexed, but may not be the Type of material' field.

The indexed field search is carried out first and then a free-text search in the output of the first search.

Obviously the Combination search strategy is useful when the number of records retrieved (or number of hits) by the indexed-based search is large and a further selection from the set may be made by free-text search – e,g, year or period of publication.

4.5 Search Operators

Search languages permit the use of various operators in formulating search expressions in order to improve the search results; firstly. retrieve all relevant references and not retrieve non-relevant references; secondly, enable evaluation of the retrieved references and organize or rank them in a sequence of decreasing relevance.

Boolean Operators: The widely used Boolean search operators include::

1 OR, AND, AND NOT

In the search expression for a particular software each of these operators may be represented by a symbol. For example in CDS-ISIS : OR by +; AND by * and NOT or AND NOT by ^

The search expression: aves **OR** birds (aves + birds in CDS-ISIS) implies that the retrieved records may contain either of the terms 'aves' or 'birds' or both the terms.

The search expression: information **AND** management (information * management in CDS-ISIS) implies that the retrieved records must contain both the terms 'information' and 'management'

The search expression: oil NOT mineral oil (oil ^ mineral oil in CDS-ISIS) implies that the retrieved records should contain the term 'oil' but should not contain the term 'mineral oil'

Other operators:

Parentheses are used to shorten the search expression and also ensure that the combination of terms in Boolean expressions are correctly applied. For example, the information need Pollution in Bihar, Orissa, and Tamil Nadu may be expressed as three separate search expressions:

POLLUTION * BIHAR
POLLUTION * ORISSA
POLLUTION * TAMIL NADU

The three separate expressions may be combined (using CDS-ISIS operators) as:

POLLUTION * BIHAR + POLLUTION * ORISSA + POLLUTION * TAMIL NADU

Using parentheses, the search expression may be made compact as:

POLLUTION*(BIHAR + ORISSA + TAMIL NADU)

Truncation: A person interested in information on 'FOREST' may also find useful / related information in documents that are indexed under each or combination of the terms FOREST, FORESTS, FORESTRY, FORESTATION. Using the OR operator the search expression, can be:

FOREST + FORESTS + FORESTRY

However, most retrieval software permit truncation, for instance, by suffixing a character (other than those that may normally occur in the search terms) to a common root of the search terms. CDS-ISIS uses $ (single dollar character) for the purpose. Thus, the above search expression can be represented as:

FOREST$

Care should be taken regarding the position in the string where the $ is placed. For instance, FORE$ will also OR terms such as FOREMAN, FOREIGN, FOREWORD, etc. which will lead to considerable noise in the retrieval. (retrieval of irrelevant documents)

Some software use ? (question mark) as the truncation character and also allow specifying the number of characters that may be present after the truncation character.

Example:

RAMA RAO, K.?1

This means, the matching term in the record may contain just one character after K. For example, RAMA RAO, K.N, RAMA RAO, K.T will be a match, but not RAMA RAO, K.A.N.

Masking.: The role of masking is somewhat similar to that of truncation but normally not suffixed but placed within a term to mask a character. For example:

NORMALI?ATION

This will OR the terms NORMALISATION and NORMALIZATION

Field Specific Search: If a user wishes to retrieve references to documents *on* S.R.Ranganathan but not to works *by* him, then it may be necessary to specify in the search expression that the search term Ranganathan, S.R. should occur in the Subject Keyword / Descriptor field only. The use of an **operand qualifier** in the search expression can be a solution. The search expression can be RANGANATHAN,S.R. /(field tag for subject descriptor) or SU=RANGANATHAN, S.R

Adjacency: In order to improve precision in retrieval with search expression(s) containing two or more terms, it may be necessary to specify that the selected:

two terms should **follow one after the other with no other term separating them**
the two terms may be **separated by one, or two, or three…or n other terms only**
the two terms may be **separated by not more than one, or two, or three or … n other terms.**

CDS-ISIS prescribes **adjacency operators $ (dollar)** and **. (dot)** for the purpose**.**

4.5 Limitations of Boolean Logic

The Boolean retrieval process generally uses the inverted or index file to the database (s) being searched. Some of the limitations of this mechanism are as follows: (Gerrie, 1983)

1. A query must me fully and precisely state as a logical expression compatible with the indexing practrice of the database, irrespective of whether the index is made up of derived or assigned terms. This implies that an incomplete or open-ended search expression is not permitted.

2. The Boolean mechanism tends to have an all-or-nothing quality, that is, if the logical search expression is true for a specific document, the corresponding document / entry is retrieved; otherwise it is not retrieved; there is no partial matching.

3. The logic of sets compels value-judgment regarding a document into a binary scale – that is, either relevant or not relevant.

4. There is no order of preference within sets of retrieved records, each document retrieved is taken as important as any other in the set.

5. Although a query can be precisely expressed using Boolean logic, the set of retrieved documents is not arranged in any sequence in terms of relevance to the query. A single retrieved set can be ranked after the retrieval.

6. With Boolean logic it is difficult to vary the depth of a search in order to vary the number of records or quantity of information required unless terms are either grouped in classes so that they can be automatically substituted for one another or the searcher is willing to study the properties of the index language in order to construct the right kind of query

7. Boolean algebra or calculus of sets is not suited for considering doubtful classes or variable product classes. String ABC may represent a different class (BAC) in a different context. For example, the three terms *Paint, Destruction, Microorganisms* in combination may connote either

Destruction of **Paint** by **Microorganisms**

Or

Destruction of **Microorganisms** by **Paint**

4.6 Mismatch Between Information Need and Search Results

Problems of online searching identified in Fenichel's study (1980-1981) are summarized below:

Approaches to searching vary considerably, from one searcher to another even on the same system and database.

Major problems of most searchers are with the search strategy and not with the mechanics of the retrieval system.

Many searchers perform simple searches and do not browse the retrieved references to check the adequacy of a search formulation or to improve a search, that is, they do not make full use of online interactive facilities.

The online search process tends to be sensitive to and affected by many factors – including economic considerations, administrative policies, the search environment, the charging policy - in addition to the skill of the searcher and the nature of the question.

1.

The major causes for the deviation between the actual information need of end-user and search results occurring at different stages of the search procedure include the following (Guinchat and Menou, 1990, p.314):

Problem recognition and actual information need of end-user

Information need as perceived by end-user

- Definition of problem
- Knowledge of available information
- Knowledge of information sources

Information need as expressed in end-user's query

- Communication means used (letter, phone, e-mail, in person, etc.
- Conscious retention of the exact parameters of the problem
- Communication ability of end-user

Information need as interpreted by Information specialist (IS)

- Subject / domain knowledge of IS
- IS's understanding of end-user's needs
- IS's aptitude for effective dialogue with end-user
- IS's capacity for expression

Query formulation in the language of the system

- Knowledge of search language of system
- Capabilities of the search language

Formulation of search expression and strategy

- Searcher's experience in searching particular database
- Adequacy of the search logic

Search and retrieval operations

- Capacity and flexibility of system
- Structure and organization of database

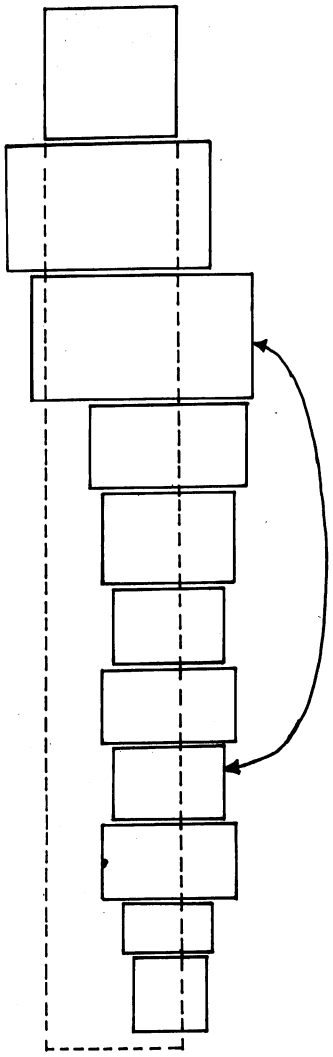
Intermediate results

- Adequacy of database
- Validity of judgement of pertinence

Final result

- Refinement of search expression
- Modification of search strategy

Fig: The Search Process



The unsatisfactory search results can arise from inadequate und-user / information specialists interaction. Selection of inappropriate database(s), inadequacy of and / or inadequate understanding of the search language of the database resulting in less effective searcher-system interface, etc. Therefore, a knowledge about and an understanding of the characteristics of the database system, the search language, formulation of search expression and searching, and specification of user needs can assist in adopting measures and techniques to improve search results vis a vis end-user needs.

Some of the differences between databases and errors within databases may not affect significantly searching in a printed page but they do so when searching online. For a searcher without adequate knowledge of the design and structure of the searchable fields, the capabilities of the query language, of the indexing language, the search engines, etc. obtaining satisfactory search results can be quite difficult. Variations can occur across database files, across the searchable indexes, and across the query languages used. When a number of databases are to be searched in different online services or websites, the difficulties may be compounded.

4.7 Major Approaches in Search Strategy

Major approaches in searching are (Markey and Atherton, 1981; Harter, 1986, chapter 7; Hartley, 1990, p. 170-71):

1. Briefsearch
2. Block-building method – reducing a request to a series of sub-problems

3. Successive fractions method – incolving iterative refinement of a broad starting strategy
4. Citation pearl growing method – a form of prototype matching
5. Starting a search with the most diskriminating facet first;
6. Starting the search with the lowest posted or the most specific term as measured against the file; and
7. A combination of two or more of the above methods – usual in actual practice of searching.

4.7.1 Briefsearch

A Briefsearch is a quick and relatively less expensive way of getting a rough idea about what a database holds about a subject. Usually it involves a single search formulation, a Boolean combination of a few search terms, without spending time to use a vocabulary control tool to select synonyms, related terms etc. There may be minimal interaction between the searcher and the system. Recall ratio may be low.

Example: birds AND India

Schools AND Karnataka

One may not adopt the Briefsearch approach for a full-fledged search. However if appropriate specific terms are used in formulating the search expression, it may result in high precision (relevance ratio) and also reasonably high recall. Example: biodiversity AND antartica. Briefsearch can be useful to examine the terms used (e.g. descriptors) in a database for a particular concept; to select additional terms to build further search steps as in the Building Block Method and Pearl Growing Method. Briefsearch will also be useful to retrieve a record for a document known to be relevant to the query, for example, by words in the title, name of author, etc. Such a record can provide additional terms to build up a more in-depth or complete search strategy.

4.7.2 Building-block Method

In the Building-Block strategy each concept of the query is enlarged or enriched by synonyms and related terms (say, taken from a thesaurus) using Boolean OR to build smaller blocks which are then ANDed together to produce the answer set. The principal steps in the strategy after query formulation, formulation of objectives of the search and selecting appropriate database services and databases accessible include the following:

1. Identification of major concepts and facets and the interrelationship among them;
2. formulating search strings that represent the concepts: words, phrases, descriptors, identifiers, qualifiers, etc.;
3. determining database fields to be searched;
4. for each component of the search expression a set of posting will be created; the sets are then Ored;
5. the sets resulting from step 4 are then combined using BooleanAND and NOT operators;
6. the intermediate results are evaluated and search terms and strategy modified if necessary and the process is iterated to obtained desired results.

Example: Query: Microcomputer software packages for information retrieval

Building blocks: s1 computer OR microcomputer OR PC

s2 software OR packages OR software packages

s3 information retrieval

s4 s1 AND s2 AND s3

This strategy is logical in construction but takes time in keyboarding and requires knowledge of the appropriate vocabulary. Use of a vocabulary control tool is advisable. It is a good way of conducting a comprehensive search requiring high recall.

7.7.3 Successive Fractions Strategy

The Successive Fractions approach is a way of reducing a large set created by using AND or NOT.

Example: Query: Swimming and physical fitness of women

Search sets:	s1	swimming AND (wom?n OR female?)	620 postings
	s2	s1 AND fit?	42 postings
	s3	s2 AND yr >= 1995	7 postings

Terms specifying date of publication and / or language(s) are useful to reduce the number of documents retrieved.

If the search retrieves too few or no records at all or the terms used in the search expression appear to be ambiguous, then the strategy may have to be changed. Two successive facet methods are available for the purpose – most specific concept first and fewest postings first. These are closely related to the Successive Fractions Strategy. All these methods start with a search expression that results in high recall and proceed to reduce the set to a smaller number of postings. They differ slightly in the method for selecting the initial set of postings.

4.7.4 Citation Pearl Growing Method

The Citation Pearl Growing method begins with a very small initial set, even a single item known to be relevant to the query. The initial record(s) set may be retrieved by Name of Author, a known specialist in the subject. The record(s) retrieved can provide additional relevant terms from the title, descriptors, abstract, etc. These can be used to carry out further searches on the subject. The Citation Pearl Growing method is helpful in searches on subjects that are not familiar to the searcher and those in which vocabulary control tools are not available.

Prototype matching technique is similar to Citation Pearl Growing strategy.

4.8 Formulation of Search Expression

Formulation of a search expression consists of translating the concepts of the search request into the language of the database or information retrieval service selected. If the database has an associated vocabulary control tool, such as, list of subject headings, a thesaurus, and/or classification scheme, then it should be referred to, if not already done so while preparing the search request, in order to select appropriate near-synonymous or equivalent terms, and broader, narrower, and related terms.

Example: Query: References on Hyperactivity in Children

Key concepts selected: Children, Hyperactivity
Reference to a related thesaurus gives:

Children	Hyperactivity
Young children	Hyperkinesis
Infants	
Neonates	
Adolescents	

Boolean Expression:
(hyperactivity OR hyperkinesis) AND (children OR young children OR infants OR neonates OR adolescents)

The search expression implies that the retrieved records / documents should contain at least one of the terms from each set of ORed terms (terms placed in parentheses)
When a search in a broad subject area is desired, consultation of the documentation on the database will indicate whether such a feature for search is provided for. For example:

-- by DDC number in MARC records
-- by Biosystematic Codes in BIOSIS
-- by Section Code in CAS Online, Sociological Abstracts etc.

-- by Subfiles in CAB International and TOXLINE

Example with ***Sociological Abstracts***

Section Code 1900 'The Family and Socialization' a search expression could be:

? ESH=1900

Ref	**Items**	**Index-term**	
E1	225	SH=1844	[SH = Section Heading Code]
E2	3456	SH=19	
E3	8615	*SH=1900	
E4	2356	SH=1900/38	
E5	2345	SH=1900/39	
E6	1654	SH=1900/40	

….. more**….**

The searcher may select one or more of the Section Heading Codes, for example, **E5 Adolescence and Youth** for use in a search expression

The EXPLODE feature in relation to the terms in the hierarchical structure of MeSH terms in the *Medical Subject Headings – Tree Structure*, the EXPAND feature in DIALOG, and ZOOM on ESA-IRS can be used for a similar purpose, that is, for searching a broad subject area.

Example: SEARCH (EXPAND DE/CHILDREN)

CHILDREN	785
ADOPTED CHILDREN	73
MIGRANT CHILDREN	107
YOUNG CHILDREN	128
INFANTS	3679
PRESCHOOLD CHILDREN	115

NEONATES	235
PREMATURE INFANTS	189
RESULT: 3175	

If the term(s) of the vocabulary control tool does not adequately represent a concept, the database may use natural language terms as index terms – free terms, text words, identifiers, etc, that may be applied in addition to or instead of, the controlled vocabulary terms. The following example illustrates such use.

Query : Effect of meditation on hypertension
Database searched : MEDLINE
Vocabulary tool : Medical Subject Headings (MeSH)

Let us say the term 'Meditation' does not occur in MeSH. Therefore, the search expression may be:

HYPERTENSION AND MEDITATION(TW) [TW denotes Textword]

At this stage, the retrieval may be limited or narrowed down by language and/or date of publication of the papers and other parameters. Alternatively, these can be introduced at a later stage if the initial search expression retrieves a large number of records.

4.9 Improving Search Results: Recall and Precision Devices

Precision value is the ratio of the number of records retrieved to the number of relevant records present in the database. Precision devices are used to restrict the number of records retrieved. Such devices enable the coordination of linking information at various stages of document indexing, search, and retrieval processes. These devices include logic, syntactic, weighting devices, and bibliographic coupling.

4.9.1 Recall Devices

Recall devices are used to increase the number of relevant references retrieved by a search expression. The more important of such devices are synonym control, hierarchical term linkage, associative term linkage, document clustering, control of word forms, summation of document sets, and bibliographic coupling so as to be able to retrieve related materials relevant to the subject of the query which may otherwise be missed.
Each of these devices can be implemented in different ways in online search. A summary of the implementation of these devices is given in Annex 1:

4.9.1.1 Synonym control

The searcher has no control over the indexing language, method or depth of indexing used by the producers of databases or the online services on which the databases may be accessed. However, if the software used has a facility for mapping of terms, then the searcher may use a non-preferred term from the vocabulary control tool associated with the database but will be guided or directed to the preferred term. In some systems, such as, MEDLINE, it is not advisable to use concurrently two search aids, e.g. expanding (exploding) an access term and the mapping facility. It may result in a 'NO POSTINGS' message.
Terms that are not truly synonymous are treated as such (e.g. FLATS and HOUSES), it may result in loss of term discrimination in the search language.

In some systems synonym control is implemented by confounding synonymous and near-synonymous terms, using different terms in indexing but they may be taken as synonyms in searching. The SAVE Search command in most information retrieval systems can emulate this feature. Judicious use of truncation can help, for example: LAW? can take care of LAW, LAWS, LAWYERS, LAWMEN, LAW SUIT, etc.

4.9.1.2 Hierarchical term linkage

Hierarchical Term linkage refers to the linking of the selected descriptor to its broader terms (BT) and narrower terms (NT) (super-ordinate and sub-ordinate terms). For

example ***see also*** references in subject heading lists and BT and NT in thesauri. Establishing such relationships should preferably be done while formulating the search expression, say, with the help of appropriate vocabulary control tools. The EXPAND / EXPLODE facility of the search software automatically matches generically related search terms taken from the controlled vocabulary. At time this can lead to the inclusion of noise (irrelevant) terms in the search expression.

4.9.1.3 Associative term linkage

Associative Term linkage represent non-hierarchic associative term relations for a given descriptor, for example, the RT terms in a thesaurus. About thirty such relationships have been identified (Neelameghan, 2001; Neelameghan and Ravichandra Rao, 1974).

4.9.1.4 Document clustering

Associative and probabilistic indexing apply statistical techniques to produce clusters of keywords that form the basis for automatic derivative classification for retrieval. The associative groups are used in cluster-based retrieval (van Rijsbergen, 1979). The searcher controls the search and decides on the stopping point, to produce a list of retrieved documents in a sequence ranked according to their closeness to the query.

4.9.1.5 Control of word forms

Such control is essentially based on the prescriptive use of vocabulary in indexing. Judicious use of truncation enables matching of words with a common root. For example:

NETWORK? will match with NETWORK, NETWORKING and NETWORKS

LEGISLAT? will match with LEGISLATION, LEGISLATIONS, LEGISLATIVE, LEGISLATOR. LEGISLATORS, and wrongly spelt words, such as, LEGISLATOIN.

Control of word forms at search point provides different search options. However, the flexibility provided by truncation will be lost if confounding of word forms is done automatically at the indexing stage.

4.9 Summation of document sets

Most online retrieval services allow the use of Boolean operators. Different word forms can be ORed together. For example

NETWORK OR NETWORKING OR NETWORKS

With CDS-ISIS , the ANY FILE and ANY FILE SEARCH provide for this facility extensively.

Gerrie's remarks are noteworthy in this connection:

"Under certain circumstances, the Logical expression A AND B may be too constraining, in that not all wanted documents explicitly contain B. Recall can be improved by identifying alternatives for B or by simply omitting B from the search and identifying some undesirable concept C. (A AND B) or (A NOT C), which will retrieve those documents that contain A and explicitly mentioning B while rejecting those explicitly mentioning C.

Quorum logic, in which retrieval is based on "n out of x terms" is also equivalent to Boolean logic and is useful when retrieval of a polythetic group is required but the actual term combinations are unimportant. Both weighted retrieval and quorum logic have the advantage of being able to express a retrieval function in a concise way when the retrieval objective is clear"

4.10 Precision Devices

Precision ratio is the number of records / documents retrieved to the number of records / documents judged to be relevant to the user's query. Precision devices are used for limiting the number of records retrieved and involve the coordination or linking of information at various stages of the indexing of documents, searching and retrieval processes. Online retrieval provides a variety of linkages and pre-coordinate index terms

may be broken up for post-coordinate retrieval, the distinction between pre-coordinate indexing and post-coordinate indexing becomes hazy in online retrieval. The major precision devices may be grouped into logic devices, syntactic devices, weighting devices, and bibliographic cycling. A summary of the implementation of precision devices is given in Appendices 4A and 4B

4.10.1 Logic Devices

Logic devices include Boolean operators (AND, and AND NOT as distinct from OR NOT), and appropriately applied quorum logic (a retrieval specification "7 out of 10"terms will retrieve fewer records with a higher probability of relevance than a specification "5 out of 10"). Logic devices are based on post-coordinate contextual analysis, an extension of Boolean logic, and takes into account positional information in the retrieval process for enhancing the associative strength between terms so as to improve the precision.

4.10.2 Syntactic Devices

Links and roles are used with syntactic devices. Links show which concepts are related by assigning a common letter or a number code to each of them or by using fixed tags and subfield code combinations so as to avoid false coordination or terms. As there are other methods of avoiding false coordination there is relatively less interest in the use of links and roles. PRECIS as a pre-coordinate indexing system has been using role indicators.

4.10.3Weighting devices

Term weighting technique is used for presenting search results in a ranked sequence. Weighting applied at the stage of indexing a record is based either on the intuitive judgment of a human indexer identifying concepts as major or minor or based on a statistical assessment of a term's weight depending on a calculated frequency of occurrence of terms. Retrieval based on such weights regard weighted occurrences of terms as more specific subsets of the whole set, both weighted and un-weighted.

Weighting at the stage of searching permits the searcher to express an interest in a term over another for reasons unrelated to the actual use of the terms in the set of documents. A user interested in the subject "Comparative export potential of software from Hong Kong and Taiwan" may give greater weightage to a document that presents export figures for the two countries than to separate documents that deal with software industry in Hong Kong and software industry in Taiwan. There are several approaches to ranking of the retrieved sets (Hartley and others, 1990, p. 350-355; Gerrie, 1983, p. 154-155).

4.10.4 Bibliographic Cycling

Bibliographic cycling (or bibliographic coupling) is a measure of concept / subject association determined by the number of cited works that documents have in common. The measure remains over time once the documents are published. Bibliographic association may be used to increase precision in retrieval in a scoring search.

The search begins with retrieval of a document known to be relevant to the query. Documents in a file or database in which the cited documents do not match in any way those in the document chosen are rejected. Documents with matching cited items are presented in a ranked sequence of their decreasing commonness of items with those in the bibliography of the chosen document.

4.11 Sensitivity to Errors

Most database systems require almost perfect input of data in cataloguing / indexing, search statements, command, etc. regarding spacing, punctuation, and spelling. Errors in typing and those relating to understanding of procedures and protocols specific to the particular system / software / service being used are common. Some types of typing errors are:

Character insertions	infformation
Character omission	infomation
Character substitution	imformation
Character reversal	informatoin

A system's lack of sensitivity to such simple human errors can lead to mismatch of search terms with the terms in the record or the index. This can result in recall or precision failures. The chances of retrieval of no records due to misspelling, incorrect usage of punctuation, spacing etc., are quite high. Such failures can be due to the searcher not using all possible approaches to retrieval or the search language used by the searcher not being fully compatible with that of the indexer of the system (Lancaster and Fayen, 1973)

Some systems (cf MEDLINE) ignore superfluous spacing and punctuation in the input. Right truncation can help in some cases as indicated earlier. The use of an abbreviated entry form for lengthy word form can offset the chances of errors in typing longer words

Some systems do not report error in the input search expression. As mentioned earlier, Boolean combinations of terms can be interpreted in more than one way, unless the components are properly enclosed in parentheses.

4.12 Check your progress

1. Who coined the phrase 'Information Retrieval'?

(A) Calvin Mooers

(B) S.R.Ranganthan

(C) J.D.Brown

(D) H.P.Luhn

2. Queries of the users are translated into the indexing system and matching is done with the vocabulary of the system

(a) Query formulation

(b) Query assimilation

(c) Query matching

(d) **Search strategy**

3. The merit of what type of searching is to offer the using of Boolean logic which allows limiting or expanding the search, as required?

(i) manual searching

(ii) online searching

(iii) literature searching

(iv) reference searching.

4. Blair and Maron evaluation study on retrieval effectiveness of full text search is called

(i) SMART retrieval system

(ii) MEDLARS evaluation study

(iii) STAIRS project

(iv) Cranefield –II project

5.Recall devices are used to increase the number of relevant references retrieved by a search expression. (True/False)

Ans:- True

4.13.Summary

A search in a database (information resource collection) is an interactive, iterative and heuristic process. It is often a trial and error exercise in which the searcher in a conversational or dialogue mode with the system and, if necessary and possible, with the participation of the end-user, attempts at successively refining and reformulating the search strategy and the initial search expression on the basis of responses of the system and end-user's reactions to the intermediate search results.

4.14. Glossary

Briefsearch: is a quick and relatively less expensive way of getting a rough idea about what a database holds about a subject.

Adjacency: In order to improve precision in retrieval with search expression(s) containing two or more terms, it may be necessary to specify that the selected:

Masking.: The role of masking is somewhat similar to that of truncation but normally not suffixed but placed within a term to mask a character

Parentheses: are used to shorten the search expression and also ensure that the combination of terms in Boolean expressions are correctly applied.

4.15. Questions for self study

1. What is Search Language? Explain its General Features
2. Describe the Index-based search strategy
3. Explain the role of Search Strategy in Information Retrieval

4.16.References

4.12. REFERENCES

1. Gerrie, Brenda. (1983). Online information systems: use and operating characteristics, limitations and design alternatives. Arlington, VA: Information Resources Press.

2. Guinchat, P. and Menou, M. (1990). Sciences et techniques de l'information et de la documentation: introduction generale. Paris: Unesco.

3. Harter, Stephen P. (1988). Online information retrieval: concepts, principles and techniques. London: Academic Press.

4. Hartley, R.J., Keen, E.M., Large, J.A., and Tedd, L.A. (1990). Online searching: principles and practice. London: Bowker-Sauer.

5. Kashyap, M.M. (2005). Database system: design and development. Ed. 2. New Delhi: Ess Ess Publications

6. Kini, Srinivas Narasimha and Jacob, K. Paulose. (2000). XML for creating information
7. content on the Internet. Information Studies, v.6(4); p. 241-254.

8. Litton, Gerry M.. (1987). Introduction to database management: a practical approach. Indian ed. New Delhi: S. Chand & Co.

9. Martin, James. (1992). Principles of data-base management. . 7th Indian reprint. New Delhi: Prentice-Hall of India.

10. Neelameghan, A. (1995). Online database searching and retrieval: strategies, procedures, commands, and problems. Bangalore: Sarada Ranganathan Endowment for Library Science' [with tutorial diskette].

11. Neelameghan, A. and Prasad, K.N. eds. (2001). National Seminar on Classification. In
12. The Digital Environment. (Papers contributed to the National Seminar on
13. Classification in the Digital Environment, Bangalore, 9-11 August 2001). Bangalore: Sarada Ranganathan Endowment for Library Science; 2001.

14. Raghavan, K.S. and Neelameghan, A. (2002). Composite multimedia works on CD:
a. catalog entry according to ISBD(ER) and AACR2 revision 1988. Cataloging &
15. Classification v.33 (3/4); p.193-209.

16. Rajashekar, T.B. (2001). Content organization for Internet-based information services. In:
a. Content organization in the new millennium (Papers presented to the Seminar

b. on Content Organization in the New Millennium, Bangalore, 2-4 June 2000 /
c. edited by A.Neelameghan and K.N. Prasad. p. 51-92.. Bangalore: Sarada
d. Ranganathan Endowment for Library Science; 2001.

17. Urs, Shalini R. and Raghavan, K.S. (2000). Metadata formats: an overview.
18. Information Studies, v.6(2); 2000; p. 111-123.

19. Urs, Shalini R. and Raghavan, K.S. (2001). Organizing WEB resources: XML for enhancing retrieval effectiveness. In: Content organization in the new millennium (Papers presented to the Seminar on Content Organization in the New Millennium, Bangalore, 2-4 June 2000 / edited by A. Neelameghan and K.N. Prasad. p. 93-06.. Bangalore: Sarada Ranganathan Endowment for Library Science; 2001.

Appendix 3: Example of Authority Lists

Fields of the database (as in IDRC's CDS-ISIS based MIBIS database)

Tag	Name	Repeatable	Subfields
901	Corporate body		abcd
902	See reference(s	R	
903	Other language version(s)	R	
904	Former name(s)	R	
905	Later name(s)	R	
911	Serial title		
912	ISSN		
913	See reference(s)	R	
914	See also other language edn(s)	R	
915	Former title(s)	R	
916	Later title(s)	R	

921	Supplier authority code	
922	Supplier name and adder	abcd
997	Authority record notes	
998	Authority date	

CORPORATE BODY

International Development Research Centre, Ottawa, ON, CA.

SEE REFERENCE(S): IDRC.
OTHER LANGUAGE VERSION(S): Centre de Recherches pour le Developpement International.
Centro Internacional de Investigaciones para el Desarrollo.
Authority record notes: Source: its annual report.
Authority record date: 1988-01-14.

Centre de Recherches pour le Developpement International, Ottawa, ON, CA.

SEE REFERENCE(S): CRDI.
OTHER LANGUAGE VERSION(S): International Development Research Centre.
Centro Internacional de Investigaciones para el Desarrollo.
Authority record notes: Source: its annual report.
Authority record date: 1988-01-14.

Centro Internacional de Investigaciones para el Desarrollo, Ottawa, ON, CA.

SEE REFERENCE(S): CIID.
OTHER LANGUAGE VERSION(S): International Development Research Centre.
Centre de Recherches pour le Developpement International.
Authority record notes: Source: its annual report.
Authority record date: 1988-01-14.

OECD, Development Centre, Paris, FR.

OTHER LANGUAGE VERSION(S): OCDE. Centre de Developpement.
Authority record date: 1988-08-10.

OCDE, Centre de Developpement, Paris, FR.

SEE REFERENCE(S): Centre de D,veloppement de l'OCDE.
OTHER LANGUAGE VERSION(S): OECD. Development Centre.
Authority record date: 1989-02-23.

OECD, Paris, FR.

SEE REFERENCE(S): Organisation for Economic Co-operation and Development.
OTHER LANGUAGE VERSION(S): OCDE.
Authority record date: 1989-02-23.

OCDE, Paris, FR.

SEE REFERENCE(S): Organisation de Coop,ration et de Developpement Economiques.
OTHER LANGUAGE VERSION(S): OECD.
Authority record date: 1989-02-23.

SERIAL

WHO chronicle

ISSN: 0042-9694
OTHER LANGUAGE EDITION(S): Cr¢nica de la OMS
Chronique OMS
Authority record date: 1988-08-04

Cr¢nica de la OMS

ISSN: 0250-8591
OTHER LANGUAGE EDITION(S): WHO chronicle
Chronique OMS
Authority record date: 1988-08-04

Chronique OMS

ISSN: 0373-3556
OTHER LANGUAGE EDITION(S): WHO chronicle
Cr¢nica de la OMS
Authority record date: 1988-08-04

Canadian woman studies

ISSN:	0706-8204
SEE REFERENCE(S):	Cahiers de la femme
EARLIER TITLE(S):	Canadian women's studies
Authority record notes:	Name changed with v. 3, no. 2, 1981.
Authority record date:	1987-04-01

Appendix 4A: Recall Devices

Device	Implementation	Remarks
Synonym control	a) Prescriptive use of an index language while indexing	Possible loss of term diskrimination
	b) Prescriptive use of an index language while indexing with term mapping while retrieving	
	c) Confounding synonyms	Advantageous to have relationships defined by searcher
Hierarchical linkage of terms	a) Indexing documents both specifically and generically	Difficult to control search level
	b) Explosion by automatically linking generically related search terms at point of search	Spurious/undesirable generic-specific relationships in search expression
Associative linkage of terms	a) Automatic keyword classification	
	b) Associative probabilistic	

	indexing	
Document clustering	a) Cluster-based retrieval	
Control of word forms	b) Truncation and suffixing c) Prescriptive use of an index language search options d) Confounding word forms e) String searching	Control of word forms at point of search provides for a variety of search options
Summation of document sets	a) Boolean OR b) Weighted retrieval c) Quorum logic	Qualitative retrieval and brevity of search expression
Bibliographic cycling	a) Forward cycling b) Backward cycling	Links between documents may be spurious; on the other hand associations that would otherwise remain hidden are revealed

---

Appendix 4B: **Precision Devices**

Device	Implementation	Remarks
Logic devices	a) Boolean AND, AND NOT b) Quorum logic x out of n c) Contextual logic	Reformulation of a search to increase precision is easier with quorum logic
Syntactic devices	a) Links b) Roles	
Weighting	a) At the point of indexing b) At the point of search c) At the point of retrieval d) At the point when document enter the system	Easier to vary the specificity of a search by assigning theshold and cut-off values. Weighting at the point of search provides more flexibility including the retrieval of a conti uous set of documents in a sequence of their probable relevance
Bibliographic coupling	Measures association among documents by their co-citation frequency	