

CIT 741: ADVANCED INFORMATION STORAGE AND RETRIEVAL



NATIONAL OPEN UNIVERSITY OF NIGERIA

CIT 741: ADVANCED INFORMATION STORAGE AND RETRIEVAL

Course Code	CIT 741
Course Title	Advanced Information and Retrieval
Course Developer/Writer	Dr Kenneth Ikechukwu Nkuma-Udah Federal University of Technology, Owerri
Programme Leader	Prof. Afolabi Adebajo National Open University of Nigeria
Course Coordinator	



NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters
14/16 Ahmadu Bello Way
Victoria Island
Lagos

Abuja Office
No. 5 Dar es Salaam Street
Off Aminu Kano Crescent
Wuse II, Abuja
Nigeria

e-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng

Published by
National Open University of Nigeria

Printed 2009

ISBN:

All Rights Reserved

CONTENTS	PAGE
Module 1 Overview of Information Storage and Retrieval (ISR)	1
Unit 1 Concept of Information	1
Unit 2 Information Life Cycle	5
Module 2 Information Representation	10
Unit 1 Representation of Bibliographic Information	10
Unit 2 Representation of Non-Bibliographic Information	19
Unit 3 Standard Metadata and their Description	22
Unit 4 Issues in Information Representation	30
Module 3 Information Organisation and Storage	38
Unit 1 Organisation of Bibliographic Information	38
Unit 2 Organisation of Digital Information	43
Unit 3 Data Representation and Organisation on Information System	49
Module 4 Information Retrieval Models	54
Unit 1 Retrieval of Bibliographic and Digital Information	54
Unit 2 Information Retrieval Models	60
Unit 3 Query Structure	71
Unit 4 User Profiles	77
Module 4 Information Retrieval Systems	83
Unit 1 Information Retrieval System	83
Unit 2 Web Information Retrieval	89
Unit 3 Bibliographic Information Retrieval System and Evaluation	93

Module 1 Overview of Information Storage and Retrieval (ISR)

Unit 1 Concept of Information

Unit 2 Information Life Cycle

UNIT 1 CONCEPT OF INFORMATION**CONTENTS**

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 What is Information?
 - 3.2 Types of Information
 - 3.3 Sources of Information
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider what information is. You will also learn the types and sources of information.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- explain what information is
- describe the types of information
- discuss where information can be sourced.

3.0 MAIN CONTENT**3.1 What is Information?**

Information can be defined as a message received and understood. However, in terms of data, it is defined as a collection of facts from which conclusions may be drawn. Generally, information is the result of processing, manipulating and organizing data in a way that adds to the knowledge of the person receiving it.

3.2 Types of Information

With respect to its content, information can be divided into:

- fact or opinion/analytic information
- objective or subjective information
- primary or secondary information

And with respect to both content and audience, information can be divided into:

- popular, scholarly or trade information

Note that these information types are not exclusive, because information can be a combination of two or more types. For instance information can be both primary and opinion/analytic; subjective and popular; or secondary and scholarly.

Let us discuss these information types in more detail:

1. Fact

A fact is the statement of a thing done or existing. Example, the Nigerian Golden Eaglets won the 2007 FIFA Under - 17 Football World Cup Tournament.

2. Opinion/Analytic information

An opinion/analytic information is a personal view or judgment based on what seems to be true, or an interpretation of fact. Example, the 1994 Super Eagles teams was as good as the 1998 Super Eagles.

3. Objective Information

Objective information presents all sides of a topic. Example, HIV is transmitted in several ways: Heterosexual intercourse accounts for 70% of transmission. Homosexual intercourse and intravenous drug users sharing contaminated needles also spread the HIV virus. The virus also can spread from mother to child by transfer across the placenta or through breast milk. Some infections were recorded from contaminated blood after blood transfusion before screening procedures were introduced in the late 1980s.

4. Subjective Information

Subjective information provides opinions/evaluative information on a topic. It commonly does not provide all sides of a topic. Example, HIV is transmitted in several ways, including via homosexual activity. Thus HIV infection can be seen as a punishment from God resulting from homosexuals' ungodly activities.

Subjective information sometimes sounds like Opinion. The basic difference is that an opinion is often a brief statement based on what seems true to a person (e.g., Blue is the best color for cars), while subjective information often involves a lengthier presentation of information which combines opinion with an incomplete (not multi-sided) presentation of a topic.

5. Primary Information

Primary information is information in its original form. Its information which has not been published elsewhere, put into context, interpreted, or translated. What exactly constitutes primary information can vary by discipline. In the sciences, primary information means the original study and its data. Example, counting the number of dogs in Aba and detailing their age & health conditions. In the languages, primary information means information in the original spoken language before its translated into another language. Example, *the original Hebrew Bible*.

Primary information is also information from "the source," or which has not been commented upon. Example, a person's diary (unedited) is a primary source. The lyrics and notes to a song

are primary information.

6. Secondary Information

Secondary information is information removed in some way from its original form. It may include restatements, examinations, interpretations, or translations. Secondary information is basically primary information that has been put into context, interpreted, or translated. Example, a book which discusses how highways impact populations of domestic carnivores and uses the data from the Aba dog study in justifying its conclusions is a secondary source. A web page commenting on a the text of a person's diary or a DJs comments on air about the meaning and significance of a particular song is a secondary information.

7. Popular, Trade and Scholarly Information

A better appreciation of the popular, trade and scholarly Information can be grasped if the features of the three types are compared and contrasted as shown in table 1 below.

Table 1: Features of Popular, Trade and Scholarly Information

	<i>Popular</i>	<i>Trade</i>	<i>Scholarly</i>
<i>Appearance</i>	Glossy paper Lots of adverts Adverts for everyday products (make up)	May be glossy paper May be color adverts Adverts for items used in field	Plain paper Black & White illustrations Few adverts (books)
<i>Authorship</i>	Professional writers Paid for work	Professional writers or academics Often unpaid	Academics/People with advanced degrees NOT paid
<i>Editors</i>	Review for style, fact checking Not expert in field	Review for style, fact checking Often not expert in field	Review for content (peer review) Expert in field
<i>Format/ Structure</i>	Real people with "issues" on topic plus two "expert" views	"How I Did It Good"	Abstract, Literature Review, Hypothesis, Methodology, Findings, Sources
<i>Sources</i>	Often people Texts not quoted or cited	Mix of people and texts Texts not fully cited	Lots of sources (texts) Complete citations
<i>Audience/ Language</i>	"Anyone" can understand (8th grade reading level) – "General audience"	Written for specialists, but practical not theoretical	Written for specialists Lots of jargon
<i>Purpose</i>	Inform the general public Make money on publication	Help peers Make money on publication	"Advance scholarship /understanding"
<i>Frequency (if periodical)</i>	Frequently – maybe even weekly	Monthly	2-4 times a year
<i>Length</i>	Often short	Often short	Usually long

3.3 Sources of Information

Knowing what type of information one needs, can help the individual find information more effectively because information sources depend to a large extent on types of information under

consideration.

In looking for factual information, the sources include print or electronic reference sources such as: dictionaries, atlases, handbooks, directories, books, articles, and web sites. However, books, articles, and web sites are not efficient ways to find out, e.g., the story of Nigeria civil war.

Sources of opinions and other subjective information include books, articles, and web pages. are all likely sources of opinions & subjective information. Review articles and editorials in newspapers and other publications are especially good sources of opinions and other subjective information. Opinion-only sources, like reviews and editorials should be avoided and just taken for what they are – subject or opinion information.

For objective information, sources may include print or electronic reference sources such as encyclopedias or handbooks. Books, articles and web pages can form sources of objective information, but the source should be considered carefully.

Primary information may be found in an article or report, an un-translated book, and/or any resource from the time/place studied. Secondary information sources may include articles, books, or web pages.

4.0 Conclusion

In this unit you have been introduced to information. You have also been introduced to the the types and sources of information.

5.0 Summary

In this unit, information is the result of processing, manipulating and organizing data in a way that adds to the knowledge of the person receiving it. Information type can be factual or opinion/analytic information; objective or subjective information; primary or secondary information; or popular, scholarly or trade information. Sources of information depend on the type of information under consideration.

6.0 Tutor-Marked Assignment

1. Define Information
2. Describe the types of information
3. Discuss where opinion, subjective and secondary information can be sourced

7.0 References/Further Readings

1. Manning C.D., P. Raghavan, H. Schütze. (2008). *Introduction to Information Retrieval*, Cambridge UP, London.
2. Baeza-Yates R., B. Ribeiro-Neto.(1999). *Modern Information Retrieval*, Addison-Wesley.

UNIT 2 INFORMATION LIFE CYCLE

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Information Life Cycle
 - 3.1 Information Identification
 - 3.2 Information Capture
 - 3.3 Information Organization
 - 3.4 Information Management
 - 3.5 Information Utilization
 - 3.6 Information Archiving
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider what information life cycle is. You will also learn the principle steps in information life cycle.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- explain what information cycle is
- describe various ways of identifying, capturing and organizing information
- discuss how information can be managed, utilized and archived.

3.0 INFORMATION LIFE CYCLE

Information doesn't exist in a vacuum; it's used by man in the context of its surroundings. Without the correct identification, capture, organisation, management, utilization and archiving or storage of information, one would be hard pressed to make any decisions - business, personal, or governmental - let alone good ones. To properly manage information, it's essential to understand the way in which information will be used, by whom, and for what purpose.

The value of any information is determined by the tripartite factors of context (the purpose of the information), content (the information itself) and user (the audience) as shown in figure 1.

Not only does the context of how the information will be used matter in identifying data, but it also guides the capture and presentation of that information. Context is all about asking the right questions to understand the eventual use of managed information.

Content is not always readily accessible and may not be in a manageable form. The form of the information will directly influence the management mechanisms, perhaps limiting the possible solutions to non-computer-based management.

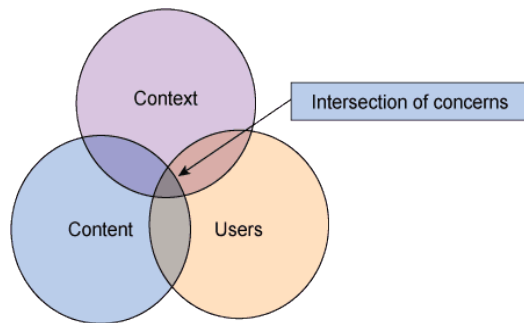


Figure 1: Intersection of the tripartite factors of context, content and users in assessing the value of information for information management

Content is not always readily accessible and may not be in a manageable form. The form of the information will directly influence the management mechanisms, perhaps limiting the possible solutions to non-computer-based management.

Information users are a mixed bag. While some are able to find the exact right set of keywords, others struggle to find useful content. And not all users have same needs; some prefer a complex, detailed display, and others prefer a simplified presentation that allows a free-form ability to browse. An effective information-management policy must support most kinds of users.

In information management, it is important to note that the value of information is variable. Some information are always valuable, such as investment account balances; other information has a defined period of time when it's valuable, such as plane departure and arrival information; and still other information (data) has value only periodically, such as business intelligence. Nevertheless, all information has a life cycle during which it's identified, captured, organized, controlled, utilized, and eventually archived or stored. Figure 2 illustrates these six principle steps in the information life cycle.

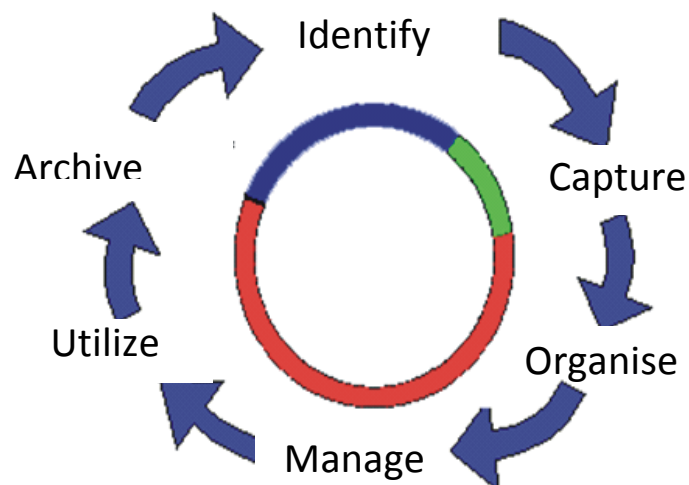


Figure 2: Information life cycle

3.1 Information Identification

The first step in information management is identifying content to be managed. For example, if you're creating a repository of requirements for a development team, the items of value can be initially identified as business requirements, system requirements, and testing requirements. Most if not all information to be managed fall into one of those categories. This also provides you with an understanding of the data source, which may be easy or difficult to manage, depending on the form and period of time between updates.

A frequently updated information source that has an inaccessible format requires a much more sophisticated scheme than one that is periodically updated in a readily accessible form. This approach also provides scope to the effort, which prevents trying to manage everything having to do with developing a new system or modifying an existing system.

3.2 Information Capture

Information capture is the process of collecting data in form of paper documents, forms and e-documents, transforming them into accurate, retrievable information, and delivering the information into business applications and databases for immediate action.

With the information identified, the next step is to capture that information into a manageable repository, where the content format dramatically affects the storage needs. Assuming all the information of interest is binary, then the primary questions of storage are size and bandwidth. The size of the files determines the principle storage needs (including backup) and the level of bandwidth required for capture and eventual display. Large files, like video or music, require a much larger storage space and delivery capacity.

Storage needs are similar regardless of the mechanism (database, network device [tape], or file system). Remember that you must provide sufficient scaling for future needs and sufficient bandwidth to accommodate user downloads of the content. As for processor power, if the metadata associated with a file is properly indexed to the searches (hitting only the indexes), then processor needs tend to scale linearly with the user load.

3.3 Information Organization

Information organization refers to methods of rendering large amounts of information into a form that can be stored, retrieved and manipulated by users or computer system. An example of information organization is Digital Information Organization (DIO).

Organization of content means that all information must be tagged in some fashion so that users can readily locate it later. This tagging may be as simple as document title or as sophisticated as the Library of Congress metacategory method. In either case, it's a good idea to develop a controlled vocabulary in a formal metadata definition document to guide both the initial repository development and the acquisition of new materials.

A *controlled vocabulary* is a hierarchy of categorization labels that are applied to all the information in the repository. For most purposes, a single hierarchy is sufficient, such as for

simple document retrieval; but there may be need to organize materials in a cross-referenced secondary hierarchy if multiple content forms (eg. comedy/video or documentary/audio books) are stored.

With any controlled vocabulary, choosing the granularity for each level of the tagging hierarchy is a critical decision for both maintenance and information navigation. This is the hardest part of organizing information and the one most likely to cause long-term difficulties in adding new materials.

3.4 Information Management

Information management depicts a comprehensive approach to managing the flow of an information system's data from creation and initial storage to the time when it becomes obsolete and is deleted. Unlike earlier approaches to data storage management, information management involves all aspects of dealing with data, starting with user practices, rather than just automating storage procedures.

Managing the repository involves updating materials periodically as older materials are archived and newer ones are added. Depending on the technical storage of the information (database, content management system, or file system), the configuration change-control mechanism either is directly provided by the storage software (such as for content-management systems) or must be layered over the information storage (such as for a file system).

3.5 Information Utilization

As noted earlier, if end users can't effectively find the information they're looking for, the repository won't be effective and will likely fall into disuse. Proper utilization involves two interrelated functions: search and navigation.

Searching is based on the metadata associated with the repository materials; index design based on the expected search categories dramatically speeds discovery of properly labeled materials. *Navigation* is the ability to rapidly move around the information space to locate related information.

Information presentation is also a key factor in utilization. For information-management purposes, presentation is involved in ensuring the accuracy of data. *Accuracy* means ensuring that the tagged information belongs with the assigned category, much like putting a book on the correct shelf. Presentation tools that let the maintainer see and browse the content for a particular category are valuable, especially where content is automatically captured from the information source.

3.6 Information Archiving

The goal of archiving is preservation rather than ready access. Information reaches the end of its life cycle when it begins to lose direct value to the user community. At this point, it's no longer cost-effective to have the data take up space in the primary information store; you should move the data to an archival location where the long-term maintenance cost is reduced. Currently, that

means moving the content to either tape or disk-archive arrays.

Moving content means repeating the identification step, only in reverse; now you're looking for information that isn't frequently accessed by the user community and migrating that information to the archive, freeing up space for new acquisitions.

4.0 Conclusion

In this unit you have been introduced to information life cycle. You have also been introduced to the six principle steps in the information life cycle.

5.0 Summary

In this unit, every information consists of a life cycle, which consists in its identification, capture, organization, control, utilization, and its eventual archiving.

6.0 Tutor-Marked Assignment

1. With the aid of a diagram, explain what is meant by Information life Cycle
2. List and discuss the six principle steps in information life cycle.

7.0 References/Further Readings

1. Baeza-Yates R., B. Ribeiro-Neto.(1999). *Modern Information Retrieval*, Addison-Wesley.
2. Grossman D.A., O. Frieder. (2004). *Information Retrieval: Algorithms and Heuristics*, Springer.

Module 2 Information Representation

Unit 1 Representation of Bibliographic Information

Unit 2 Representation of Non-Bibliographic Information

Unit 3 Standard Metadata and their Description

Unit 4 Issues in Information Representation

UNIT 1 REPRESENTATION OF BIBLIOGRAPHIC INFORMATION**CONTENTS**

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Bibliographic Information
 - 3.2 Bibliographic Information Cataloging
 - 3.3 MARC
 - 3.4 MARC 21
 - 3.3.1 MARC 21 Format
 - 3.3.2 MARC 21 Record
 - 3.3.3 Organization of MARC 21 Record
 - 3.5 UNIMARC
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider what bibliographic information is. You will also learn about bibliographic information cataloging, especially machine-readable cataloging.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- explain what bibliographic information is
- describe how to catalog bibliographic information
- discuss the various types of machine-readable cataloging.

3.0 MAIN CONTENT**3.1 Bibliographic Information**

A bibliography is a list of resources used or referred to by an author. It is also used to mean a list of books and other resources thought to be useful on a particular subject. Bibliographies range from cited works' lists at the end of books and articles to complete, independent publications. As separate works, they may be in bound volumes, or computerised bibliographic databases. Bibliographies used to be lists of written resources. Today, however, they often include information on other resources such as interviews, video and audio tapes, computer resources and speeches.

Bibliographic information is the information about a resource consulted during the process of a work and written in a standard bibliographic format. Information will include, the author, title, publisher, date, place, volume and page (for a printed source); and author, editor, title, organisation, web address or url and last date (for a web source).

Bibliographic works differ in the amount of detail depending on the purpose, and can be generally divided into two categories: enumerative bibliography (also called compilative, reference or systematic), which results in an overview of publications in a particular category; and analytical, or critical, bibliography, which studies the production of books. In earlier times, bibliographic information mostly focused on books. Now, both categories of bibliography cover works in other formats including recordings, motion pictures and videos, graphic objects, databases, CD-ROMs and websites.

A database of bibliographic information is called a *bibliographic database*. It may be a database containing information about books and other materials held in a library (e.g. an online library catalog) or, as the term is more often used, an electronic index to journal or magazine articles, containing citations, abstracts and often either the full text of the articles indexed, or links to the full text.

An example of a bibliography is:

Dechant, Emerald. 1991. *Understanding and teaching reading: An interactive model*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. 522.

3.2 Bibliographic Information Cataloging

For two centuries, bibliographic cataloging has had quite a tremendous development. It was originally the *book style* catalogs such as those of the British Library. Next came the *card* catalogs that comprised thousands of drawers at large institutions. Then the *microfilms (and microfiche)*, which are microreproductions of documents in form of reels (and flat sheets) used for transmission, storage, reading, and printing of information. Thereafter, came the *Online Public Access Catalog (OPAC)* is an online database of materials held by a library or group of libraries and used for searching a library catalog to locate books, periodicals, audio/visual materials or other items under control of a library. Lastly and presently we have the *World Wide Web* or The Web.

Right from the time computers came into being, precisely in the 1960s, bibliographic data format started being developed to catalog bibliographic information online. Consequently, this format was made machine-readable so that compatible computers can read and interpret the information. Therefore, the method of cataloging became known as Machine-Readable Cataloging (MARC).

3.4 MARC

The *MARC* standards consist of the MARC formats, which are standards for the representation and communication of bibliographic and related information in machine-readable form. It defines a bibliographic data format that was developed at the Library of Congress in the 1960s and provides the protocol by which computers exchange, use, and interpret bibliographic information. Its data elements make up the foundation of most library catalogs used today.

"Machine-readable" in MARC means that one particular type of machine, a computer, can read and interpret the data in the cataloging record, while "Cataloging" means a bibliographic recording, or recording information in a catalog card. MARC records contain many information, but usually include:

- a description of the item
- main entry and added entries
- subject headings
- the classification or call number.

1. Description:

The *description* of the item in MARC record is shown in the paragraph sections of a catalog card. It includes the title, statement of responsibility, edition, material specific details, publication information, physical description, series, notes, and standard numbers.

2. Main Entry and Added entries:

The *main entry and other added entries* are the access points, which are the retrieval points in the library catalog where patrons should be able to look up the item.

3. Subject Headings (or subject added entries):

The librarian uses some list of standard subject headings to select the subjects under which the item will be listed. Use of an approved list is important for consistency, to ensure that all items on a particular subject are found under the same heading and therefore in the same place in the catalog.

4. Call Number:

The Dewey Decimal or Library of Congress classification schedule is used to select the call number for an item. The purpose of the call number is to place items on the same subject together on the same shelf in the library. Most items are sub-arranged alphabetically by author. The second part of the call number usually represents the author's name, facilitating this subarrangement.

There are many national and international variants of MARC. The most popular variants are:

- MARC 21: this is the "harmonization" of United States MARC and Canadian MARC; it is maintained by the Network Development and MARC Standards Office of the Library of Congress.
- UNIMARC: this was created by International Federation of Library Associations (IFLA) in 1977; it is the official MARC in France, Italy, Russia, Portugal, Greece and other countries.

3.5 MARC 21

As mentioned above, *MARC 21* is a result of the combination of the United States and Canadian MARC formats (USMARC and CAN/MARC). *MARC 21* formats are standards for the representation and communication of bibliographic and related information in machine-readable

form and is based on the American National Standard Institute (ANSI) standard Z39.2, which allows users of different software products to communicate with each other and to exchange data. MARC 21 was designed to redefine the original MARC record format for the 21st century and to make it more accessible to the international community.

Currently MARC 21 has been implemented successfully by The British Library, the European Institutions and the major library institutions in the United States, and Canada. The MARC 21 formats are maintained by the Library of Congress in consultation with various user communities. The British Library adopted MARC 21 as its cataloguing format in 2004 as part of the implementation of an integrated library system.

MARC 21 Format for Bibliographic Data is designed to be a carrier for bibliographic information about printed and manuscript textual materials, computer files, maps, music, continuing resources, visual materials, and mixed materials. Bibliographic data commonly includes titles, names, subjects, notes, publication data, and information about the physical description of an item.

The bibliographic format contains data elements for the following types of material, with their associated designator codes in parenthesis:

- *Books (BK)* - used for printed, electronic, manuscript, and microform textual material that is monographic in nature.
- *Continuing resources (CR)* – formerly referred to as Serials (SE); used for printed, electronic, manuscript, and microform textual material that is issued in parts with a recurring pattern of publication (e.g., periodicals, newspapers, yearbooks).
- *Computer files (CF)* - used for computer software, numeric data, computer-oriented multimedia, online systems or services. Other classes of electronic resources are coded for their most significant aspect. Material may be monographic or serial in nature.
- *Maps (MP)* - used for all types of printed, electronic, manuscript, and microform cartographic materials, including atlases, sheet maps, and globes. Material may be monographic or serial in nature.
- *Music (MU)* - used for printed, electronic, manuscript, and microform music, as well as musical sound recordings, and non-musical sound recordings. Material may be monographic or serial in nature.
- *Visual materials (VM)* - used for projected media, non-projected media, two-dimensional graphics, three-dimensional artifacts or naturally occurring objects, and kits. Material may be monographic or serial in nature.
- *Mixed materials (MX)* - formerly referred to as Archival and manuscript material (AM); used primarily for archival and manuscript collections of a mixture of forms of material. Material may be monographic or serial in nature.

3.5.1 MARC 21 Format

A MARC 21 format is a set of codes and content designators defined for encoding machine-readable records. Formats are defined for five types of data: bibliographic, holdings, authority,

classification, and community information. These formats are:

a) *MARC 21 Format for Bibliographic Data*: This contains format specifications for encoding data elements needed to describe, retrieve, and control various forms of bibliographic material. The MARC 21 Format for Bibliographic Data is an integrated format defined for the identification and description of different forms of bibliographic material. With the full integration of the previously discrete bibliographic formats, consistent definition and usage are maintained for different forms of material.

b) *MARC 21 Format for Holdings Data*: This contains format specifications for encoding data elements pertinent to holdings and location data for all forms of material.

c) *MARC 21 Format for Authority Data*: This contains format specifications for encoding data elements that identify or control the content and content designation of those portions of a bibliographic record that may be subject to authority control.

d) *MARC 21 Format for Classification Data*: This contains format specifications for encoding data elements related to classification numbers and the captions associated with them. Classification records are used for the maintenance and development of classification schemes.

e) *MARC 21 Format for Community Information*: This provides format specifications for records containing information about events, programs, services, etc. so that this information can be integrated into the same public access catalogs as data in other record types.

3.5.2 MARC 21 Record

A MARC record involves three elements: the record *structure*, the *content designation*, and the data *content* of the record. These elements are:

a) *The structure of MARC records*: is an implementation of national and international standards, e.g., Information Interchange Format (ANSI Z39.2) and Format for Information Exchange (ISO 2709).

b) *The Content designation*: is the codes and conventions established to identify explicitly and characterize further the data elements within a record and to support the manipulation of those data; this is defined in the MARC 21 formats.

c) *The content of most data elements*: is defined by standards outside the formats, such as cataloging rules, classification schemes, subject thesauri, code lists, or other conventions used by the organization that creates a record. e.g., Anglo-American Cataloguing Rules, Library of Congress Subject Headings, National Library of Medicine Classification. The content of other data elements, e.g., coded data, is defined in the MARC 21 formats.

3.5.3 Organisation of MARC 21 Record

A MARC 21 record consists of three main sections: the *leader*, the *directory*, and the *variable*

fields. These sections are:

a) *The leader*: consists of data elements that contain coded values and are identified by relative character position. Data elements in the leader define parameters for processing the record. The leader is fixed in length (24 characters) and occurs at the beginning of each MARC record.

b) *The directory*: contains the tag, starting location, and length of each field within the record. Directory entries for variable control fields appear first, in ascending tag order. Entries for variable data fields follow, arranged in ascending order according to the first character of the tag. The order of the fields in the record does not necessarily correspond to the order of directory entries. Duplicate tags are distinguished only by location of the respective fields within the record. The length of the directory entry is defined in the entry map elements in Leader/20-23. In the MARC 21 formats, the length of a directory entry is 12 characters. The directory ends with a field terminator character.

c) *The data content of a record*: is divided into *variable fields*. The MARC 21 formats distinguish two types of variable fields: *variable control fields* and *variable data fields*. Control and data fields are distinguished only by structure. The term **fixed fields** is occasionally used in MARC 21 documentation, referring either to control fields generally or to specific coded-data fields, e.g., 007 (Physical Description Fixed Field) or 008 (Fixed-Length Data Elements).

3.6 UNIMARC

The primary purpose of *UNIMARC* (Universal Machine-Readable Cataloging) is to facilitate the international exchange of bibliographic data in machine-readable form between national bibliographic agencies. UNIMARC may also be used as a model for the development of new machine-readable bibliographic formats. UNIMARC is maintained by an International Federation of Library Associations (IFLA) committee, the Permanent UNIMARC Committee (PUC).

The scope of UNIMARC is to specify the content designators (tags, indicators and subfield codes) to be assigned to bibliographic records in machine-readable form and to specify the logical and physical format of the records. It covers monographs, serials, cartographic materials, music, sound recordings, graphics, projected and video materials, rare books and electronic resources.

UNIMARC is intended to be a carrier format for exchange purposes. It does not stipulate the form, content, or record structure of the data within individual systems. UNIMARC does provide recommendations on the form and content of data when it is to be exchanged. Records are usually structured in exchange tape format as the last stage in any conversion process, after form, content, and content designation have been converted to the UNIMARC standard. Those organizations intending to use UNIMARC for data interchange will find it useful to co-ordinate their internal format content designators and field and subfield definitions with those in UNIMARC to reduce the complexity of data conversion when the records are converted into the UNIMARC exchange tape structure.

3.5.1 UNIMARC Structure

UNIMARC is a specific implementation of International Organization for Standardization (ISO 2709), an international standard that specifies the structure of records containing bibliographic data. It specifies that every bibliographic record prepared for exchange conforming to the standard must consist of: a *record label* (consisting of 24 characters), a *directory* (consisting of a 3-digit tag of each data field, along with its length and its starting character position relative to the first data field) and *data fields* (of variable length, each separated by a field separator).

The specific layout of a bibliographic record in UNIMARC

RECORD LABEL	DIRECTORY	DATA FIELDS	R/T = Record Terminator
--------------	-----------	-------------	-------------------------

Figure 1: General layout of a bibliographic record in UNIMARC

Record Label:

ISO 2709 prescribes that each record start with a 24-character Record Label. This contains data relating to the structure of the record, which are defined within the standard ISO 2709, and several data elements that are defined for this particular implementation of ISO 2709. These implementation-defined data elements relate to the type of record, its bibliographic level and position in a hierarchy of levels, the degree of completeness of the record and the use or otherwise of International Standard Bibliographic Description (ISBD) or ISBD-based rules in the preparation of the record. The data elements in the Record Label are required primarily to process the record and are intended only indirectly for use in identifying the bibliographic item itself.

Directory:

Following the Record Label is the Directory. Each entry in the Directory consists of three parts: a 3-digit numeric tag, a 4-digit number indicating the length of the data field and a 5-digit number indicating the starting character position. No further characters are permitted in a Directory entry. The Directory layout is as follows:

Directory entry 1			Directory entry 2		Other directory entries
Tag	Length of Field	Starting Position	F/T = Field Terminator	

Figure 2: Directory layout of a bibliographic record in UNIMARC

The second segment of the Directory entry gives the number of characters in that field. This includes all characters: indicators, subfield identifiers, textual or coded data and the end of field marker. The length of field is followed by the starting character position of the field relative to the first character position of the variable field portion of the record. The first character of the first variable field is character position 0. The position of character position 0 within the whole record is given in character positions 12-16 of the Record Label.

The tag is 3 characters long, the 'length of the data' fills 4 characters and the 'starting character position' fills 5 characters. After all of the 12-character directory entries corresponding to each data field in the record, the directory is terminated by the end of field marker IS2 of ISO 646.

Data Fields:

The variable length data fields follow the directory and generally contain bibliographic as opposed to processing data.

Data (Control) Field (00-) layout:

Data	F/T
------	-----

Figure 3: Data field (00-) layout of a bibliographic record in UNIMARC

Data Field (01- to 999) layout:

Indicators	Subfield Identifier	Other Subfields				
Ind 1	Ind 2	\$a (etc.)	Data	Data	F/T

Figure 4: Data field (01- to 999) layout of a bibliographic record in UNIMARC

Tags are not carried in the data fields but appear only in the directory, except for tags in embedded fields. Fields with the tag value 00- (e.g. 001) consist only of the data and an end of field character. Other data fields consist of two indicators followed by any number of subfields. Each subfield begins with a subfield identifier that is composed of a subfield delimiter, IS1 (1/15 of ISO 646), and a subfield code (one alphabetic or numeric character) to identify the subfield.

The subfield identifiers are followed by coded or textual data of any length unless stated otherwise in the description of the field. The final subfield in the field is terminated by the end of field character IS2 (1/14 of ISO 646). The last character of data in the record is followed as usual by the end of field character IS2 which in this instance is followed by the end of record character IS3 (1/13 of ISO 646).

3.5.2 Mandatory Fields in UNIMARC

The following is a list of fields that must be present in the UNIMARC record:

- 001* RECORD IDENTIFIER
- 100* GENERAL PROCESSING DATA
- 101 LANGUAGE OF THE WORK (when applicable)
- 120 CODED DATA FIELD: CARTOGRAPHIC MATERIALS GENERAL (cartographic items only)
- 123 CODED DATA FIELD: CARTOGRAPHIC MATERIALS SCALE AND COORDINATES (cartographic items only)
- 200* TITLE AND STATEMENT OF RESPONSIBILITY (\$a title proper is the only mandatory subfield)
- 206 MATERIAL SPECIFIC AREA: CARTOGRAPHIC MATERIALS MATHEMATICAL DATA (cartographic items only)
- 801* ORIGINATING SOURCE FIELD

The fields marked by an asterisk (*) must be present in every record, without exception. However, when records are converted into UNIMARC from other MARC versions, the remaining fields in the list above are not regarded as mandatory if meaningful fields cannot be produced directly or by computer algorithm. For example, 101 should be omitted if the record would otherwise contain nothing more than 101 |#\$a|||.

3.5.3 Record Linking in UNIMARC

In practice there are situations when it may be desirable to make a link from one bibliographic entity to another. Example, when a record describes a translation, a link may be made to the record that describes the original. Another example is when a link may be made between records relating to different serial titles following a change of name. A technique is provided in UNIMARC for making these links. A block of fields (the 4-- block) is reserved for this purpose.

A linking field will include descriptive information concerning the other item with or without information pointing to a separate record that describes the item. A linking field is composed of subfields, each of which contains a UNIMARC field made up of tag, indicators, and field content including subfield markers. Note that these embedded fields are not accessible through the Directory, since only the entire linking field has a directory entry. The tag of the linking field denotes the relationship of the item identified within it to the item for which the record is being made.

4.0 Conclusion

In this unit you have been introduced to representation of biographic information. You also learnt the MACHINE-Readable Cataloging (MARC) as well as the two main versions of MARC.

5.0 Summary

In this unit, Bibliographic information is the information about a resource consulted during the process of a work and written in a standard bibliographic format. It is mainly represented in a machine-readable form for easy storage, access and utilization. This can be done using any of the Machine-Readable Cataloging (MARC) formats in use.

6.0 Tutor-Marked Assignment

1. Define Bibliographic Information
2. List and briefly discuss the usual information contained in a MARC record
3. What are the types of material with their associated designator codes found in MARC 21
4. Briefly describe the specific layout (contents) of a bibliographic record in UNIMARC

7.0 References/Further Readings

1. van Risjbergen C.J. (2004). *The Geometry of Information Retrieval*, Cambridge UP.
2. Chowdhury G.G. (2003). *Introduction to Modern Information Retrieval*, Neal-Schuman
3. Meadow C.T., B.R. Boyce, D.H. Kraft, C.L. Barry. (2007). *Text Information Retrieval Systems*. Academic Press.

UNIT 2 REPRESENTATION OF NON BIBLIOGRAPHIC INFORMATION

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Non Bibliographic Information
 - 3.2 Differences between Bibliographic and Non Bibliographic Information
 - 3.3 Searching for Non-bibliographic Sources
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider what non bibliographic information is. You will also learn the differences between bibliographic and non bibliographic information and how to search for non bibliographic information.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- explain what non bibliographic information is
- discuss the differences between bibliographic and non bibliographic information
- describe how to search for non bibliographic information

3.0 MAIN CONTENT

3.1 Non Bibliographic Information

Non bibliographic information is any information which is not officially cited in a recognized database or "reported in the literature" and has records which are non structured and non semantic. Non Bibliographic information include information such as listings of audiovisuals, federal legislation, computer simulation models for teaching, class lecture notes and schedules, computer documentation, library online public access catalogs (OPACs), phone directories, public announcements, software etc which are usually generated from many sources.

Non-bibliographic information sometimes called factual information has databases containing data, numbers, and perhaps even some knowledge. While there may be over seven million information reported in the scientific literature, there are data on only a portion of these. As an example, the Beilstein database (a non bibliographic database) dating back about 150 years which does contain data on organic chemicals has about half the number of compounds which are in the Chemical Abstracts Registry database (a bibliographic database).

3.2 Differences between Bibliographic and Non Bibliographic Information

The main differences between bibliographic and non-bibliographic information are outlined in the table below:

Table 1: Differences between Bibliographic and Non Bibliographic Information

Bibliographic Information	Non Bibliographic Information
Cited in recognized database	Not cited in recognized database
Records are highly structured and semantic	Records are non structured and non semantic
Database store references to documents available elsewhere	Database store facts, figures, text or graphics
Information accessed from their databases is usually available in print	Information accessed from their databases in many cases is not available in print

3.3 Searching for Non-bibliographic Sources

The reference sources for non bibliographic information can be sought by determining the type of information needed according to categories outlined below. Non-bibliographic information sources are identified through directories, guides and bibliographies to them. The types of non bibliographic information corresponding to the type of materials being sought include the following categories:

- **People** [*who*]
 - biographic dictionaries
 - encyclopedias
 - pseudonym dictionaries
 - address books
 - genealogy
 - heraldry
 - directories
- **Subjects** [*what*]
 - Encyclopedias
 - Dictionaries
- **Institutions** [*what*]
 - directories (including telephone directories)
 - address books
 - calendars
- **Places** [*where*]
 - geographic dictionaries
 - geographic encyclopedias
 - gazetteers
 - atlases, maps

- **Times, events** [*when*]
 - chronologies
 - calendars

- **Numerical** [*how much/many*]
 - statistics
 - censuses
 - metrology

- **Linguistic** (*dictionaries*)

- **Bibliographies of data sources:**
 - general
 - special (subject)

- **Guides to data sources**
 - printed
 - electronic

4.0 Conclusion

In this unit you have been introduced to representation of non biographic information. You also learnt about the differences between bibliographic and non bibliographic information as well as how to search for non bibliographic information.

5.0 Summary

In this unit, Non bibliographic information is any information which is not officially cited in a recognized database or "reported in the literature" and has records which are non structured and non semantic. This can be differentiated from bibliographic information.

6.0 Tutor-Marked Assignment

1. Define Non Bibliographic Information
2. List all the categories of Non Bibliographic Information you know
3. Differentiate between Bibliographic and Non Bibliographic information

7.0 References/Further Readings

1. Sparck K., Jones, P. Willett. (1997). *Readings in Information Retrieval*, Morgan Kaufmann.,
2. Kowalski G., M.T. Maybury. (2005). *Information Storage and Retrieval Systems*, Springer.
3. van Risjbergen C.J. (2004). *The Geometry of Information Retrieval*, Cambridge UP.

UNIT 3 STANDARD METADATA AND THEIR DESCRIPTION

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 What is Metadata?
 - 3.2 Functions of Metadata
 - 3.2.1 Resource Discovery
 - 3.2.2 Organizing Electronic Resources
 - 3.2.3 Interoperability
 - 3.2.4 Digital Identification
 - 3.2.5 Archiving and Preservation
 - 3.3 Creating Metadata
 - 3.3.1 Creation Tools
 - 3.3.2 Metadata Quality Control
 - 3.4 Common Metadata Standards
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider what metadata is. You will also learn the functions of metadata as well as how to create metadata. Lastly, you will study common metadata standards.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- explain what metadata is and how to create it
- describe the functions of metadata
- discuss common metadata standards.

3.0 MAIN CONTENT

3.1 What Is Metadata?

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. It is the key to ensuring that resources will survive and continue to be accessible into the future. Metadata is often called data about data or information about information. A good example of metadata is the cataloguing system found in libraries, which records for example the author, title, subject, and location on the shelf of a resource.

The term metadata is used differently in different communities. Some use it to refer to machine understandable information, while others use it only for records that describe electronic

resources. In the library environment, metadata is commonly used for any formal scheme of resource description, applying to any type of object, digital or non-digital. Traditional library cataloging is a form of metadata; MARC 21 and the rule sets used with it, such as AACR2, are metadata standards.

Other metadata schemes have been developed to describe various types of textual and non-textual objects including published books, electronic documents, archival finding aids, art objects, educational and training materials, and scientific datasets.

There are three main types of metadata:

- *Descriptive metadata* - describes an information resource for purposes such as discovery, identification and retrieval. It can include elements such as title, abstract, author, and keywords.
- *Structural metadata* - indicates how compound objects are put together or related to one another, for example, how pages are ordered to form chapters.
- *Administrative metadata* - provides information to help manage an information resource, such as when and how it was created, file type and other technical information, and who can access it.

There are several subsets of administrative data. Two of them are sometimes listed as separate metadata types. These are:

- *Rights management metadata*, which deals with intellectual property rights, and
- *Preservation metadata*, which contains information needed to archive and preserve a resource.

3.2 Functions of Metadata

The main reasons for creating metadata, especially descriptive metadata, include to facilitate discovery of relevant information, to aid resource discovery, help organize electronic resources, facilitate interoperability and legacy resource integration, provide digital identification, and support archiving and preservation.

3.2.1 Resource Discovery

Metadata serves the same functions in resource discovery as good cataloging does by:

- allowing resources to be found by relevant criteria;
- identifying resources;
- bringing similar resources together;
- distinguishing dissimilar resources; and
- giving location information.

3.2.2 Organizing Electronic Resources

With the growth of web-based resources, aggregate websites become increasingly useful in organizing links to resources based on audience or topic. Such lists can be built as static webpages, with the names and locations of the resources “hardcoded” in the HTML. The use of metadata stored in databases can allow these webpages to be built dynamically and more

efficiently. Various software tools can be used to automatically extract and reformat the metadata information for web applications.

3.2.3 Interoperability

Describing a resource with metadata allows it to be understood by both humans and machines in ways that promote interoperability. Interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality. Using defined metadata schemes, shared transfer protocols, and crosswalks between schemes, resources across the network can be searched more seamlessly.

Two approaches to interoperability are cross-system search and metadata harvesting. The American National Standards Institute (ANSI) Z39.50 (a client-server protocol for searching and retrieving information from remote computer databases) protocol is commonly used for cross-system search. Z39.50 implementers do not share metadata but map their own search capabilities to a common set of search attributes. A contrasting approach taken by the Open Archives Initiative is for all data providers to translate their native metadata to a common core set of elements and expose this for harvesting. A search service provider then gathers the metadata into a consistent central index to allow cross-repository searching regardless of the metadata formats used by participating repositories.

3.2.4 Digital Identification

Most metadata schemes include elements such as standard numbers to uniquely identify the work or object to which the metadata refers. The location of a digital object may also be given using a file name, URL (Uniform Resource Locator), or some more persistent identifier such as a PURL (Persistent URL) or DOI (Digital Object Identifier). Persistent identifiers are preferred because object locations often change, making the standard URL (and therefore the metadata record) invalid. In addition to the actual elements that point to the object, the metadata can be combined to act as a set of identifying data, differentiating one object from another for validation purposes.

3.2.5 Archiving and Preservation

Most current metadata efforts center around the discovery of recently created resources. However, there is a growing concern that digital resources will not survive in usable form into the future. Digital information is fragile; it can be corrupted or altered, intentionally or unintentionally. It may become unusable as storage media and hardware and software technologies change. Format migration and perhaps emulation of current hardware and software behavior in future hardware and software platforms are strategies for overcoming these challenges.

As mentioned above, *metadata is the key to ensuring that resources will survive and continue to be accessible into the future*. Archiving and preservation require special elements to track the lineage of a digital object (where it came from and how it has changed over time), to detail its physical characteristics, and to document its behavior in order to emulate it on future technologies.

3.3 Creating Metadata

Who creates metadata depends on discipline, the information resource being described, the tools available, and the expected outcome. Though, it is almost always a cooperative effort. Much basic structural and administrative metadata is supplied by the technical staff, who initially digitize or otherwise create the digital object, or is generated through an automated process.

For descriptive metadata, it is best in some situations if the originator of the resource provides the information. This is particularly true in the documentation of scientific datasets where the originator has significant understanding of the rationale for the dataset and the uses to which it could be put, and for which there is little if any textual information from which an indexer could work.

However, many projects have found that it is more efficient to have indexers or other information professionals create the descriptive metadata, because the authors or creators of the data do not have the time or the skills. In other cases, a combination of researcher and information professional is used. The researcher may create a skeleton, completing the elements that can be supplied most readily. Then results may be supplemented or reviewed by the information specialist for consistency and compliance with the schema syntax and local guidelines.

3.3.1 Creation Tools

A growing number of commercial software tools are becoming available. Again, many metadata project initiatives can develop tools and make them available to others, sometimes for free as is currently the situation. Creation tools fall into several categories:

- *Templates*: allow a user to enter the metadata values into pre-set fields that match the element set being used. The template will then generate a formatted set of the element attributes and their corresponding values.
- *Mark-up tools*: will structure the metadata attributes and values into the specified schema language. Most of these tools generate XML or SGML Document Type Definitions (DTD). Some templates include such a mark-up as part of their final translation of the metadata.
- *Extraction tools*: will automatically create metadata from an analysis of the digital information resource. These tools are generally limited to textual resources. The quality of the metadata extracted can vary significantly based on the tool's algorithms as well as the content and structure of the source text. These tools should be considered as an aid to creating metadata. The resulting metadata should always be manually reviewed and edited.
- *Conversion tools* will translate one metadata format to another. The similarity of elements in the source and target formats will affect how much additional editing and manual input of metadata may be required. Metadata tools are generally developed to support specific metadata schemas or element sets. The websites for the particular schema will frequently have links to relevant toolsets.

3.3.2 Metadata Quality Control

The creation of metadata automatically or by information originators who are not familiar with cataloging, indexing, or vocabulary control can create quality problems. Mandatory elements may be missing or used incorrectly. Schema syntax may have errors that prevent the metadata from being processed correctly. Metadata content terminology may be inconsistent, making it difficult to locate relevant information.

The *Framework of Guidance for Building Good Digital Collections*, available on the National Information Standards Organization (NISO) website, articulates six principles applying to good metadata:

1. Good metadata should be appropriate to the materials in the collection, users of the collection, and intended, current and likely use of the digital object.
2. Good metadata supports interoperability.
3. Good metadata uses standard controlled vocabularies to reflect the what, where, when and who of the content.
4. Good metadata includes a clear statement on the conditions and terms of use for the digital object.
5. Good metadata records are objects themselves and therefore should have the qualities of archivability, persistence, unique identification, etc. Good metadata should be authoritative and verifiable.
6. Good metadata supports the long-term management of objects in collections.

There are a number of ongoing efforts for dealing with the metadata quality challenge. These include:

- Metadata creation tools are being improved with such features as templates, pick lists that limit the selection in a particular field, and improved validation rules.
- Software interoperability programs that can automate the “crosswalk” between different schemas are continuously being developed and refined.
- Content originators are being formally trained in understanding metadata and controlled vocabulary concepts and in the use of metadata-related software tools.
- Existing controlled vocabularies that may have initially been designed for a specific use or a narrow audience are getting broader use and awareness. For example, the *Content Types* and *Subtypes* originally defined for MIME email exchange are commonly used as the controlled list for the Dublin Core *Format* element.
- Communities of users are developing and refining audience-specific metadata schemas, application profiles, controlled vocabularies, and user guidelines. The *MODS User Guidelines* are a good example of the latter.

3.4 Common Metadata Standards

Metadata Standards are developed from schemes, which metadata elements grouped into sets designed for a specific purpose, example, for a specific domain or a particular type of information resource. For every element the name and the semantics (the meaning of the element) are specified. Content rules (how content must be formulated), representation rules (e.g., capitalization rules), and allowed element values (e.g., from a controlled vocabulary) can be specified optionally.

Some schemes also specify in which syntax the elements must be encoded, in contrast to syntax independent schemes. Many current schemes use Standards Generalized Mark-up Language (SGML) or XML to specify their syntax. Metadata schemes that are developed and maintained by standard organizations (such as ISO) or organizations that have taken on such responsibility (such as the Dublin Core Metadata Initiative) are called metadata standards.

Many different metadata schemes are being developed as standards across disciplines, such as library science, education, archiving, e-commerce, and arts. In the table below, an overview of some of the available metadata standards is given.

Table 1. Metadata Standards. Source: *Metadata Standards, Wikipedia, the free encyclopedia*, at http://en.wikipedia.org/wiki/Metadata_standards

Name	Focus	Description
	Archiving	Encoded Archival Description - a standard for encoding archival finding aids using XML in archival and manuscript repositories.
	Arts	Categories for the Description of a Work of Art is a conceptual framework for describing and accessing information about works of art, architecture, and other material culture.
<u>Core</u>	Arts	Visual Resource Association – the standard provides a categorical organization for the description of works of visual culture as well as the images that document them.
<u>Core</u>	Biology	The Darwin Core is a metadata specification for information about the geographic occurrence of species and the existence of specimens in collections.
	Book industry	Online Information Exchange - international standard for representing and communicating book industry product information in electronic form.
	Data warehousing	The main purpose of the Common Warehouse Metamodel is to enable easy interchange of warehouse and business intelligence metadata in distributed heterogeneous environments.
<u>LOM</u>	Education	Learning Objects Metadata - specifies the syntax and semantics of Learning Object Metadata.
	Geographic data	Content Standard for Digital Geospatial Metadata maintained by the Federal Geographic Data Committee (FDGC).
<u>-GMS</u>	Government	The e-Government Metadata Standard defines the metadata elements for information resources to ensure maximum consistency of metadata across public sector organizations in the UK.

	Government/ Organizations	The Global Information Locator Service defines an open, low-cost, and scalable standard so that governments, companies, or other organizations can help searchers find information.
	Humanities, social sciences and linguistics	Text Encoding Initiative - a standard for the representation of texts in digital form, chiefly in the humanities, social sciences and linguistics. 27
<u>MIX</u>	Images	NISO Metadata for Images in XML is an XML schema for a set of technical data elements required to manage digital image collections.
<u><indecs></u>	Intellectual property	Interoperability of Data in E-Commerce Systems addresses the need to put different creation identifiers and metadata into a framework to support the management of intellectual property rights.
	Librarianship	MAchine Readable Cataloging - standards for the representation and communication of bibliographic and related information in machine-readable form.
	Librarianship	Metadata Encoding and Transmission Standard - standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library.
	Librarianship	Metadata Object Description Standard - is a schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications.
	Librarianship	XML Organic Bibliographic Information Schema - a XML schema for modeling MARC data.
	Media	PBCore is a Metadata & Cataloging Resource for Public Broadcasters & Associated Communities
<u>-7</u>	Multimedia	The Multimedia Content Description Interface MPEG-7 is a ISO/IEC standard and specifies a set of descriptors to describe various types of multimedia information and is developed by the Moving Picture Experts Group.
	Networked resources	Dublin Core - interoperable online metadata standard focused on networked resources.
	Networked resources	Digital Object Identifier - provides a system for the identification and hence management of information ("content") on digital networks, providing persistence and semantic interoperability.
ISO/IEC	Organizations	Standard that describes the metadata and activities needed to manage data

11179 ^[1]		elements in a registry to create a common understanding of data across organizational elements and between organizations.
ISO 15489 ^[2]	Records management	Standard for records management policies and procedures.
	Records management	A specification describing the MOdel REquirements for the management of electronic records.
	Scientific data sets	Directory Interchange Format - a descriptive and standardized format for exchanging information about scientific data sets.
	Web	Simple Knowledge Organization System - development by W3C of a simple yet powerful framework for expressing knowledge structures in a machine-understandable way, for use on the semantic web.

4.0 Conclusion

In this unit you have studied standard metadata and their description. You also learnt the functions of metadata, the tools for creating it and common metadata standards.

5.0 Summary

In this unit, Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. It is the key to ensuring that resources will survive and continue to be accessible into the future.

6.0 Tutor-Marked Assignment

1. What is Metadata?
2. List and briefly discuss the functions of metadata
3. Enumerate briefly on the metadata creation tools
4. Briefly articulate six principles applying to good metadata

7.0 References/Further Readings

1. Baeza-Yates R., B. Ribeiro-Neto.(1999). *Modern Information Retrieval*, Addison-Wesley.
2. Grossman D.A, O. Frieder. (2004). *Information Retrieval: Algorithms and Heuristics*, Springer.
3. Voorhess E.M., D.K. Harman (2005). *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press.
- 4 Sparck K. Jones, P. Willett. (1997). *Readings in Information Retrieval*, Morgan Kaufmann.,

UNIT 4 ISSUES IN INFORMATION REPRESENTATION

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Issues in Information Representation
 - 3.1 Information Creation
 - 3.2 Acquisition and Collection of Information
 - 3.2.1 Collection Policies
 - 3.2.2 Gathering Procedures
 - 3.2.3 Intellectual Property Concerns
 - 3.3 Identification and Cataloging
 - 3.3.1 Metadata
 - 3.3.2 Persistent Identification
 - 3.4 Storage
 - 3.5 Preservation
 - 3.5.1 Hardware and Software Migration
 - 3.5.2 Preservation of the Look and Feel
 - 3.5.3 Preservation Formats
 - 3.5.4 Standards and Interoperability
 - 3.6 Access
 - 3.6.1 Access Mechanisms
 - 3.6.2 Rights Management and Security Requirements
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider the issues involved in information representation, including the stages in the life cycle of information representation management.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- discuss the issues involved in information representation
- describe the stages in the life cycle of information representation management
- discuss major information archiving consideration.

3.0 ISSUES IN INFORMATION REPRESENTATION

The rapid growth in the, representation of information by authors, publishers, corporations, governments, and even librarians, archivists and museum curators, has emphasized the speed and ease of short-term dissemination with little regard for the long-term preservation of the information.

Generally, information is fragile. The situation is worse when it is considered that we are in the era of digital information, which is fragile in ways that differ from traditional technologies, such as paper or microfilm. It is more easily corrupted or altered without recognition. Digital storage media have shorter life spans, and digital information requires access technologies that are changing at an ever-increasing pace.

Some types of information, such as multimedia, are so closely linked to the software and hardware technologies that they cannot be used outside these proprietary environments. Because of the speed of technological advances, the time frame in which we must consider adequate representation, storage and archiving becomes much shorter. That is, the time between manufacture and preservation is shrinking.

3.1 Information Creation

Information Creation is the act of producing the information product. The producer may be a human author or originator, or a piece of equipment such as a sensing device, satellite or laboratory instrument. Many information managers acknowledged that creation is where information representation starts and for information durability and security, long-term storage, archiving and preservation must start here.

The best practice would be to create the metadata at the object creation stage, or to create the metadata in stages, with the metadata provided at creation augmented by additional elements during the cataloging/identification stage. However, only in the case of data objects is the metadata routinely collected at the point of creation. Many of the datasets are created by measurement or monitoring instruments, and the metadata is supplied along with the data stream.

For smaller datasets and other objects such as documents and images, much of the metadata continues to be created "by hand" and after information creation. Metadata creation is not sufficiently incorporated into the tools for the creation of these objects to rely solely on the creation process. As standards groups and vendors move to incorporate XML (eXtensible Markup Language) and RDF (Resource Description Framework) architectures in their word processing and database products, the creation of metadata as part of the origination of the object will be easier.

3.2 Acquisition and Collection of Information

The acquisition and collection of information is the stage of information representation in which the created information is "incorporated" physically or virtually into the archive. The information object must be known to the archive administration. There are two main aspects to the acquisition of information objects -- collection policies and gathering procedures. The other point is the issue of intellectual property.

3.2.1. Collection Policies

In most countries, the major difference in collection policies between formal print and electronic publications is the question of whether digital materials are included under current deposit legislation. Guidelines help to establish the boundaries in such an unregulated situation. In

Nigeria, however, such policies are lacking, so we tend to depend on external collection policies, when we find ourselves in need of such applications.

It is also the case that there is just too much material that could be archived from the Internet, so guidelines are needed to tailor the general collection practices of the organization. The collection policies answer questions related to selecting what to archive, determining extent, archiving links, and refreshing site contents.

Thus the following factors influence the collection policies:

Selecting what to archive - Both the National Library of Canada (NLC) and the National Library of Australia (NLA) acknowledge the importance of selection guidelines. The NLC's Guidelines state, "The main difficulty in extending legal deposit to network publishing is that legal deposit is a relatively indiscriminate acquisition mechanism that aims at comprehensiveness. In the network environment, any individual with access to the Internet can be a publisher, and the network publishing process does not always provide the initial screening and selection at the manuscript stage on which libraries have traditionally relied in the print environment... Selection policies are, therefore, needed to ensure the collection of publications of lasting cultural and research value."

Determining Extent - Directly connected to the question of selection is the issue of extent. What is the extent or the boundary of a particular digital work? This is particularly an issue when selecting complex Web sites.

Archiving Links - The extensive use of hypertext links to other digital objects in electronic publications raises the question of whether these links and their content should be archived along with the source item.

Refreshing the Archived Contents - In cases where the archiving is taking place while changes or updates may still be occurring to the digital object, as in the case of on-going Web sites, there is a need to consider refreshing the archived contents.

3.2.2. Gathering Procedures

There are two general approaches to the gathering of relevant information, especially internet-based. These are hand-selected and automatic. In the case of the NLA, the sites are reviewed and hand-selected. They are monitored for their persistence before being included in the archive. Alternatively, the Royal Library, the National Library of Sweden, acquires material by periodically running a robot to capture sites for its Kulturarw project without making value judgments.

3.2.3 Intellectual Property Concerns

Intellectual property remains a key issue in the acquisition process. The approaches to intellectual property vary based on the type of organization doing the information representation and archiving.

In the case of data centers or corporate archives where there is a close tie between the center and the owner or funding source, there is little question about the intellectual property rights related to acquisition. However, in the case of national libraries, the approaches to intellectual property rights differ from country to country. The differences are based on variant national information policies or legal deposit laws.

In many countries, such as Nigeria the law may either not exist or implemented, but in many others, the law has not yet caught up with the digital environment, and the libraries must make their own decisions. As an example, in the absence of digital deposit legislation, the PANDORA Project seeks permission from the copyright owner before copying the resource for the archive. In contrast, the Swedish and Finnish national library projects have an automated system and do not contact the owners.

3.3 Identification and Cataloging

Once the archive has acquired the information object, it is necessary to identify and catalog it. Both identification and cataloging allow the archiving organization to manage the information objects over time. Identification provides a unique key for finding the object and linking that object to other related objects. Cataloging in the form of metadata supports organization, access and curation. Cataloging and identification practices are often related to what is being archived and the resources available for managing the archive.

3.3.1 Metadata

All archives use some form of metadata for description, reuse, administration, and preservation of the archived object. There are issues related to how the metadata is created, the metadata standards and content rules that are used, the level at which metadata is applied and where the metadata is stored. These have been considered in the previous unit.

The majority of the projects created metadata in whole or part at the cataloging stage. However, there is increasing interest in automatic generation of metadata, since the manual creation of metadata is considered to be a major impediment to archiving, especially digital types. A variety of metadata formats are used by the selected projects, depending on the data type, discipline, resources available, and cataloging approaches used. Most national libraries use traditional library cataloging standards with some fields unable to be filled and others taking on new meaning.

There is even a greater variety of content standards used by the projects when entering data into the metadata fields. The national libraries tend to use traditional library cataloging rules such as those of Anglo-American Cataloguing Rules 2 (AACR2). However, work remains to identify the specific metadata elements needed for long-term preservation as opposed to discovery, particularly for non-textual data types like images, video and multimedia.

3.3.2 Persistent Identification

In the case of digital information, for those archives that do not copy the digital material immediately into the archive, the movement of material from server to server or from directory to directory on the network, resulting in a change in the URL, is problematic. The use of the

server as the location identifier can result in a lack of persistence over time both for the source object and any linked objects.

A multifaceted identification system is used by the American Astronomical Society (AAS). Name resolution is used instead of URLs. In addition, the AAS uses astronomy's standard identifier, called a "Bibcode", which has been in use for fifteen years.

3.4 Storage

Storage is often treated as a passive stage in the life cycle of information representation, but storage media and formats have changed with legacy information perhaps lost forever. Block sizes, tape sizes, tape drive mechanisms and operating systems have changed over time. Most organizations that responded to the question about the periodicity of media migration anticipate a 3-5 year cycle.

The most common solution to this problem of changing storage media is migration to new storage systems. This is expensive, and there is always concern about the loss of data or problems with the quality when a transfer is made.

The most rigorous media migration practices are in place at the data centers. For instance, the Atmospheric Radiation Monitoring (ARM) Center at the Oak Ridge National Laboratory plans to migrate to new technologies every 4-5 years. During each migration, the data is copied to the new technology. Each migration will require 6-12 months. This trend will likely become continuous as the size of the archive increases.

3.5 Preservation

Preservation is the aspect of information representation management that preserves the content as well as the look and feel of the information object. The time frame of long-term preservation can be thought of as long enough to be concerned about changes in technology and changes in the user community. Depending on the particular technologies and subject disciplines involved, the information managers interviewed estimated the cycle for hardware/software migration at 2-10 years.

3.5.1 *Hardware and Software Migration*

New releases of databases, spreadsheets, and word processors can be expected at least every two to three years, with patches and minor updates released more often. While software vendors generally provide migration strategies or upward compatibility for some generations of their products, this may not be true beyond one or two generations. Migration is not guaranteed to work for all data types, and it becomes particularly unreliable if the information product has used sophisticated software features. There is generally no backward compatibility, and if it is possible, there is certainly loss of integrity in the result.

Plans are less rigorous for migrating to new hardware and applications software than for storage media. In order to guard against major hardware/software migration issues, the organizations try to procure mainstream commercial technologies. For example, both the American Chemical Society and the U.S. Environmental Protection Agency purchased Oracle not only for its data

management capabilities but for the company's longevity and ability to impact standards development. Unfortunately, this level of standardization and ease of migration is not as readily available among technologies used in specialized fields where niche systems are required because of the interfaces to instrumentation and the volume of data to be stored and manipulated.

3.5.2 Preservation of the Look and Feel

At the specific format level, there are several approaches used to save the "look and feel" of material. For journal articles, the majority of the projects reviewed use image files: Tagged Image File Format (TIFF), Portable Document File (PDF), or Hyper Text Markup Language (HTML). TIFF is the most prevalent for those organizations that are involved in any way with the conversion of paper backfiles. The Optical Character Recognition (OCR), because it cannot achieve 100% accuracy, is used only for searching; the TIFF image is the actual delivery format that the user sees. However, this does not allow the embedded references to be active hyperlinks. HTML/SGML (Standard Generalized Mark-up Language) is used by many large publishers after years of converting publication systems from proprietary formats to SGML. The SGML version that is actually stored by the publisher is converted to HTML. PDF versions can also be provided by conversion routines.

For purely electronic documents, PDF is the most prevalent format. This provides a replica of the Postscript format of the document, but relies upon proprietary encoding technologies. PDF is used both for formal publications and grey literature. While PDF is increasingly accepted, concerns remain for long-term preservation and it may not be accepted as a legal depository format, because of its proprietary nature.

3.5.3 Preservation Formats

A key preservation issue is the format in which the archival version should be stored: Transformation or Native. Transformation is the process of converting the native format to a standard format. On the whole, much storage is in native formats. However, there are several examples of data transformation.

In some countries, there are intellectual property questions related to native versus transformed formats. According to Canadian Copyright Law, an author's rights are infringed if the original work is "distorted, mutilated or otherwise modified." After much discussion, the NLC decided that converting an electronic publication to a standard format to preserve the quality of the original and to ensure long-term access does not infringe on the author's right of integrity.

3.5.4 Standards and Interoperability

It is wonderful that in an environment that is so dynamic and open to change, there is a greater and greater emphasis on standards. In the political environment of those days, it was difficult to gain support for the standardization of word processing packages. However, documents are currently received in only a few formats. Text is received in SGML (and its relatives HTML and XML), PDF (Normal and Image), WordPerfect and Word. Images are received in TIFF Group 4 and PDF Image.

The market forces have reduced the number of major word processing vendors. To a lesser extent, consolidation has occurred in the number of spreadsheet and database formats. However, there is less consistency in the modeling, simulation and specific purpose software areas; much of this software continues to be specific to the project. Therefore, the emphasis in these areas is on the development of standards for interoperability and data exchange, realizing that perhaps the market forces will not play as large a role here as with more general purpose software applications.

3.6 Access

The previous life cycle functions that have been discussed are performed for the purpose of ensuring continuous access to the material in the archive. Successful practices must consider changes to access mechanisms, as well as rights management and security requirements over the long term.

3.6.1 Access Mechanisms

Most information managers consider the access and display mechanisms to be another source of change in the information environment, especially the digital one. Today it is the Web, but there is no way of knowing what it might be tomorrow. The electronic archive is used to create new access versions as the access mechanisms change. However, the originals are always retained. It has been shown, from the evolution of technology, that whatever level of detail is captured in the conversion process, it will eventually become insufficient. New hardware and software will make it possible to capture and display at higher quality over time. It is always desirable to capture and recapture using the original item.

3.6.2 Rights Management and Security Requirements

One of the most difficult access issues for information archiving involves rights management. What rights does the archive have? What rights do various user groups have? What rights has the owner retained? How will the access mechanism interact with the archive's metadata to ensure that these rights are managed properly? Rights management includes providing or restricting access as appropriate, and changing the access rights as the material's copyright and security level changes.

Security and version control also impact information archiving. There are many interesting questions concerning privacy and stolen information, particularly since the Internet Archive policy is to archive all sites that are linked to one another in one long chain.

Similarly, there is concern among image archivists that images can be tampered with without the tampering being detected. Particularly in cases where conservation issues are at stake, it is important to have metadata to manage encryption, watermarks, digital signatures, etc. that can survive despite changes in the format and media on which the digital item is stored.

4.0 Conclusion

In this unit you have studied the main issues involved in information representation.

5.0 Summary

In this unit, since information is fragile, the best way to ensure its durability is by emphasizing long-term archiving and preservation of the information.

6.0 Tutor-Marked Assignment

1. Discuss the issues involved in information representation
2. Enumerate briefly on the two main aspects to the acquisition of information objects
3. What factors will you consider when preserving the look and feel of an information object?

7.0 References/Further Readings

1. Jones K. S, P. Willett. (1997). *Readings in Information Retrieval*, Morgan Kaufmann.,
2. Kowalski G., M.T. Maybury. (2005). *Information Storage and Retrieval Systems*, Springer.
3. Understanding Metadata (2001) National Information Standards Organization (NISO), at <www.niso.org/standards/resources/UnderstandingMetadata.pdf>
4. Gail M. Hodge (2000) Best Practices for Digital Archiving: An Information Life Cycle Approach, D-Lib Magazine at <<http://www.dlib.org/dlib/january00/01hodge.html>>

Module 3 Information Organisation and Storage

Unit 1 Organisation of Bibliographic Information

Unit 2 Organisation of Digital Information

Unit 3 Data Representation and Organisation on Information System

UNIT 1 ORGANISATION OF BIBLIOGRAPHICAL INFORMATION**CONTENTS**

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Organization of Bibliographical Information
 - 3.1 What is Information Organisation
 - 3.1.1 Categories of Information By Content
 - 3.1.2 Transforming Unstructured Information to Structured Type
 - 3.2 Objectives of Information Organization
 - 3.3 Standards for Information Organization and Storage
 - 3.3.1 Catalogue Information Resources
 - 3.3.2 Classify Information Resources
 - 3.3.3 Create and Manage Databases
 - 3.3.4 Process Information Resources Physically
 - 3.3.5 Store Information Resources
 - 3.3.6 Analyze and Organize Specialized Information
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider how to organize bibliographic information. You will also learn the objective and standards of information organisation and storage.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- explain how to organize bibliographic information
- enumerate on the objectives of information organisation
- discuss the standards of information organisation and storage.

3.0 ORGANISATION OF BIBLIOGRAPHICAL INFORMATION

Organizing bibliographical information entails structuring the entire bibliographical information in the information life cycle. The principal attribute of information organisation in bibliographic systems (ie. document-based systems) consists in its classification. Classification brings like things together with respect to one or more specified attributes. Any number of attributes can be used to form classes of documents - size, color, subject, author, etc. Therefore, the most important attribute for information organisation is the attribute of "embodying the same work".

No other attribute can match it in collocating power because documents that share this attribute contain essentially the same information.

3.1 What is Information Organisation?

Information Organisation (IO) can be defined as the process of information structuring during content storage and also during search and display stages of an Information and Knowledge Access (IKA). This definition pre-supposes that for effective information organisation, the information content has to be structured.

Organizing information not only means bringing all the same information together, but also in pinpointing the differences and assist the user in selecting from the alternatives. The concept of 'work' of an author and its embodiment in different documents and the concept of different 'editions' of a document have been used and it is suggested that bibliographic systems should not only aim at bringing different works and editions together, but also point out the differences to help the user in selecting the right document and edition. This concept of 'work' may be expanded to cover works of different authors – it is expected that the bibliographic system should bring together all works which have same content. Again the concept of edition may be expanded to cover 'versions' of documents; this is important in digital versions of publications.

3.1.1 Categories of Information By Content

Broadly speaking, information can be categorized by their content into two types: *structured* and *unstructured information*. The ability to transform information from unstructured to structured type is the key to effective information organisation.

Structured Information

Structured information is information whose content is numerical and factual in nature, produced based on measurements, experimentation, survey, etc. Examples include scientific datasets, for example properties, formulae and structures of chemical substances, genetic sequences, etc.

In structured information, values for most attributes or fields is usually in numerical or codified form (e.g. name of a person, organization, designation, etc.). Search and retrieval for resources containing such information is usually by 'exact match'. This means that an information resource either satisfies or does not satisfy the query (ie. an information resource searched for a piece of structured information either possesses the value or does not possess it - binary, 'yes or no' response).

Unstructured Information

Unstructured information is information whose content is often non-numerical and not produced based on measurements, experimentation, survey, etc. Examples include textual content in resources like e-mail messages, publications of various type, presentations, lectures, expert profiles, etc. Much of these are available in multimedia formats.

Search and retrieval from unstructured information resources is usually by 'best match', i.e. to find documents which are 'most relevant' to a given query (e.g. web searching, bibliographic database searches, digital libraries).

3.1.2 Transforming Unstructured Information to Structured Type

Structured information in the scientific domain is usually referred to as 'hard databases' or 'scientific data sets', while unstructured information is sometimes referred as 'soft databases' or as 'documentary information'. Given that the volume of unstructured information is much larger than structured information and is growing very rapidly, organizing and leveraging content from unstructured information becomes a key concern.

A major challenge is how the two types of information content can be handled together during the process of information search and retrieval. One way of facing this challenge, some experts believe, is by bringing 'structure' to unstructured information. By so doing, the process of information organisation effectively handles structured information that its meant to.

3.2 Objectives of Information Organization:

The objective of information organization include:

- to provide 'structured' access to information
- to bring essentially like information together and to differentiate what is not exactly alike

To Provide 'Structured' Access to Information

Experts have often speculated that 'Information is only valuable to the extent that it is structured. Because of a lack of structure in the creation, distribution and reception of information, the information often does not arrive where it is needed and, therefore, is useless.'

It is often said that knowledge is power. However, there is an anonymous quote that *Knowledge is not power, power is structured access to knowledge*. So, a measure of how well information is organized is how well the access interface (like search and browse features for digital information) and the display interfaces are 'structured'.

Whatever information structuring is supported at the access and display interface is predominantly dependent on the quality of information structuring done while storing the content (information representations) into the repository, because it would be very difficult to provide structured access without structured representation of content. In practical terms, this means assigning appropriate metadata to information items, embedding appropriate markup, categorization, indexing, assigning identifiers, etc. at the content management level.

Information structuring also has implications for creators of information (e.g. researchers) and publishers of information, in embedding necessary structures (both syntactical and semantic) in the information items so that the user can quickly find and assimilate the most important findings in the document. For high quality information access, quality of information structuring is important both during content storage and also during access and display.

We note that the quality of information structuring is not the same for all information access. That is, the degree of information structuring (or 'structuredness') varies from one information to another. Moreover, information structuring is affected by factors such as the layout of the interface (e.g. web page design or Windows GUI screen design).

Information structuring process during content storage may be manual, fully automatic or semi automatic. It is usually automatic during the access and display stage, as this is often derived from the stored information structures.

To Bring Like Information together and Differentiate Unlike

The essential and defining objective of a system for organizing information is to bring essentially like information together and to differentiate what is not exactly alike. Designing a system to achieve this purpose is subject to various constraints. These constraints include:

- it should be economical
- it should maintain continuity with the past (documents that are already organized) and
- it should take full advantage of technologies.

3.3 Standards for Information Organization and Storage

For effective information organization and storage, the following standards must be followed:

- Catalogue information resources
- Classify information resources
- Create and manage databases
- Process information resources physically
- Store information resources
- Analyze and organize specialist information

3.3.1 Catalogue Information Resources

- Describe the principles of cataloguing and the suitability of AACR 2 standards to type of collection
- Locate and edit copy-cataloguing records for library and information resources
- Determine cataloguing standards for library and information resources
- Maintain catalogue records for library and information resources.

3.3.2 Classify Information Resources

Describe the way in which knowledge is organized in library classification systems

Describe the principles of classification and subject analysis of the item in hand

Apply a classification system to material held in library and information services

3.3.3 Create and Manage Databases

- Analyze a body of information to which access is required
- Select and develop a LIS database and collect and record information
- Maintain, monitor, and evaluate the database and implement improvements
- Review new systems to organize and access information

3.3.4 Process Information Resources Physically

- Ensure the processing of the item is correct in terms of the addition of library stationery, labels, ownership stamps and barcodes
- Ensure the covering of the item with protective covering is appropriate to required standards of presentation

3.3.5 Store Information Resources

- Prepare newly acquired resources
- Undertake basic processing of information
- Arrange material to facilitate access to resources for clients

3.3.6 Analyze and Organize Specialist Information

- Identify the requirements for descriptions of material
- Analyze material
- Describe material and format description
- Monitor and review analysis and description practices and procedures
- Contribute to enhancements of systems for describing material

4.0 Conclusion

In this unit you have studied the main points in the organisation of bibliographic information.

5.0 Summary

In this unit, organizing bibliographic information not only means bringing all the same information together, but also pinpointing the differences and assisting the user in selecting from the alternatives.

6.0 Tutor-Marked Assignment

1. Explain briefly how bibliographic information can be organized
2. Enumerate on the objectives of information organisation
3. Discuss the standards of information organisation and storage.

7.0 References/Further Readings

1. van Risjbergen C.J., (2004). *The Geometry of Information Retrieval*, Cambridge UP.
2. Chowdhury G.G., (2003). *Introduction to Modern Information Retrieval*, Neal-Schuman
3. Meadow C.T., B.R. Boyce, D.H. Kraft, C.L. Barry. (2007). *Text Information Retrieval Systems*. Academic Press.
4. Information and Knowledge Organization - Topic-8: Information organization in bibliographic systems (2006) at <www.ncsi.iisc.ernet.in/raja/is206/topic-8.htm>

UNIT 2 ORGANISATION OF DIGITAL INFORMATION

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Digital Information Organisation
 - 3.2 Digitization of Information
 - 3.3 Merits of Digital Information
 - 3.4 Demerits of Digital Information
 - 3.5 Digital vs Traditional libraries
 - 3.6. Features of a Digital Information Library
 - 3.7 Interoperability of Digital Information
 - 3.7.1 Formats and Standards
 - 3.7.2 Metadata
 - 3.7.3 Character Set Representations
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will study how to organize digital information. You will also learn digitization of information, merits and demerits of digital information, comparison of digital with traditional libraries, features of digital information library and interoperability of digital information.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- explain what digital information organisation is
- enumerate the merits and demerits of digital information
- Compare digital with traditional libraries
- Itemize the features of a digital information library
- discuss the interoperability of digital information.

3.0 MAIN CONTENT

3.1 Digital Information Organisation

Organization of Digital Information (or Digital Information Organisation) is the process of rendering large amounts of information into digital form so that it can be stored, retrieved and manipulated by computer. An example of this is the *digital library*, a storehouse of largely unstructured text documents. Another example is the *digital museum*, which contains pictorial and three-dimensional objects that are much more difficult to digitize and search than text, and that are susceptible to scanning and optical character recognition (OCR).

The rapid rise in computer and Internet use has resulted in the creation of vast quantities of digital information being created and transmitted. For example, virtually all business documents are now created in digital form, either by computers directly (in the case of machine-generated forms) or by humans using word processing software. The fact that this material is digitized makes it amenable to automated storage and retrieval.

The very health of institutions depends on their ability to manage information effectively, whether for educational, research, business, military or governmental purposes. Therefore, organisation of digital information is a critical technology for organisational entities of all sizes.

Systems employed in organisation of digital information are an amalgam of various technologies, including scanning, OCR, digital storage techniques, data compression, indexing and search algorithms, display devices and the Internet. These technologies must be integrated properly and scaled to enormous proportions to allow humans to deal effectively with the flood of digital information being made available.

3.2 Digitization of Information

It is true that ultimately, everything information that people are interested in accessing will have to be digitized. The reason is that digital searching is so easy, inexpensive, fast and ubiquitous that users will no longer tolerate, or access traditional materials. Capture requires a concerted, shared, worldwide effort. The cost of digitizing is not trivial, though, so it makes little sense for any work to be digitized more than once. Yet without any registry of digitized works, many books are digitized multiple times, while others are ignored.

Converting text, images and objects to digital form requires much more than digital photography or even high-resolution scanning. The general requirements for converting text, images and objects to digital form are the following:

- initial input, either scanning or keyboarding
- conversion to one of a set of standard formats
- optical character recognition (OCR) to capture text characters for searching
- OCR correction (since OCR is inherently error prone)
- creation and input of metadata and cataloging information
- special techniques for non-textual materials, such as music, images, videotape, etc.

3.3 Merits of Digital Information

The main advantages of digital information over traditional information include the following:

1. Everything Can Be Stored

The total number of different books produced since printing began does not exceed one billion. (It is known that the number of books now published annually is less than one million.) If an average book occupies 500 pages at 2,000 characters per page, then even without compression it can be stored comfortably in one megabyte.

2. *Very Large Databases*

A database of a billion objects, each of which occupies one megabyte, is large but not inconceivable. Once one is comfortable with sizes of this kind, it is feasible to imagine a thousand such databases, or to envision them all as portions of the same global collection. This amount of storage is sufficient to house not only all books, but photographs, legislative material, court decisions, museum objects, recorded music, theatrical performances, including opera and ballet, speeches, movies and videotape.

3. *Distributed Holdings*

When information is digitized and accessible over a network, it makes little sense to speak of its “location,” although it is technically resident on at least one storage device somewhere, and that device is connected to at least one computer. If the information is available at multiple mirror sites, it is even less meaningful to speak of it being in a “place.”

While traditional libraries measure their size by number of books, periodicals and other items held, the relevant statistic for a digital library is the size of the corpus its users may access. This means that digital libraries will want to expand their “holdings” by sharing digital links with other libraries. In the digital information world, ultimately, *all information* material should be accessible from *every* library.

3.4 Demerits of Digital Information

Although, the rapid growth in the creation and dissemination of digital objects by authors, publishers, corporations, governments, and even librarians, archivists and museum curators, has emphasized the speed and ease of short-term dissemination with little regard for the long-term preservation of digital information, it is known that digital information is fragile in ways that differ from traditional technologies, such as paper or microfilm. The main demerits of digital information include:

1. Corruption without Recognition

Digital information is more easily corrupted or altered without recognition. Digital storage media have shorter life spans, and digital information requires access technologies that are changing at an ever-increasing pace. Some types of information, such as multimedia, are so closely linked to the software and hardware technologies that they cannot be used outside these proprietary environments. Because of the speed of technological advances, the time frame in which we must consider archiving becomes much shorter. The time between manufacture and preservation is shrinking.

2. Lack of Tradition of Best Practices

While there are traditions of stewardship and best practices that have become institutionalized in the print environment, many of these traditions are inadequate, inappropriate or not well known in the digital environment. Originators are able to bypass the traditional publishing, dissemination and announcement processes that are part of the traditional path from creation to archiving and preservation. Groups and individuals who did not previously consider themselves to be archivists are now being drawn into the role, either because of the infrastructure and intellectual property issues involved or because user groups are demanding it.

3. Redundancy of Librarians and Archivists

Furthermore, librarians and archivists who traditionally managed the life cycle of print information from creation to long-term preservation and archiving, must now look to information managers from the computer science tradition to support the development of a system of stewardship in the new digital environment. There is a need to identify new best practices that satisfy the requirements and are practical for the various stakeholder groups involved.

3.5 Digital vs Traditional libraries

The shift from traditional libraries to the digital is not merely a technological evolution, but requires a change in the pattern by which people access and interact with information.

A traditional library is characterized by the following:

- emphasis on storage and preservation of physical items, particularly books and periodicals
- cataloging at a high level rather than one of detail, e.g., author and subject indexes as opposed to full text
- browsing based on physical proximity of related materials, e.g., books on science and technology are near one another on the shelves
- passivity; information is physically assembled in one place; users must travel to the library to learn what is there and make use of it

By contrast, a digital library differs from the above in the following ways:

- emphasis on access to digitized materials wherever they may be located, with digitization eliminating the need to own or store a physical item
- cataloging down to individual words or glyphs
- browsing based on hyperlinks, keyword, or any defined measure of relatedness; materials on the same subject do not need to be near one another in any physical sense
- broadcast technology; users need not visit a digital library except electronically; for them the library exists at any place they can access it, e.g., home, school, office, or in a car

3.6. Features of a Digital Information Library

The library for digital information has the following features:

- contain all recorded knowledge online (billions of items)
- distributed, maintained globally
- accessible by:
 - any person
 - in any language
 - any time
 - anywhere on earth
 - via the Internet
- act as the information resource for the 21st Century

3.7 Interoperability of Digital Information

Given that numerous libraries around the world will develop digital collections, the question becomes how it will be possible for a user of library A to access and view material housed in library B. The existence of numerous digital libraries will make it essential to share digitized items and ensure that cataloging, searching and retrieval tools at each one can be used readily with materials from others.

A number of essential elements required for interoperability of digital information are discussed below.

3.7.1 Formats and Standards

The first essential element in interoperability of digital information is standards. While standards may have the effect of inhibiting innovation, they are essential to interoperability. Agreement must be achieved on such fundamental issues as how text is to be stored. Is it straight ASCII, Microsoft Word, HTML, SGML, XML or something else? What kind of compression will be used? If text is compressed, how will searching be done? How are images, music and videotape to be represented? If agreement is not reached, at least the number of different ways in which works are digitized should be reduced to a number small enough to allow each library to support them.

Digital libraries must also have a second set of intake standards, going not to technology but to quality and reliability. Archivists question the permanence of digital materials since they note that electronic documents can be modified readily and the media on which they reside become obsolete at least once each decade. The question then is how an ever-expanding corpus of information will be converted to new media and formats as these evolve.

3.7.2 Metadata

This term is often used to mean information about an item, rather than the information in the item itself. Examples include the author, title, date of acquisition, price paid, donor, etc. It is particularly critical to capture metadata that is not present in or derivable from the item. For example, the author's date of birth is often not printed in a book but can be important in distinguishing among authors with similar names (particularly parents and children).

Libraries may *share* content by simply providing links, but uniform access to the content requires uniform metadata and a procedure for generating and storing it economically. It is of little importance to exchange documents at light speed if they must be held up for months until a human cataloger can prepare metadata.

3.7.3 Character Set Representations

This is not merely a question of different alphabets and writing systems, a major hurdle in itself, but also an issue of how characters are represented. For example, there are several widely differing mappings of Chinese characters into ASCII. There is some appeal to having a worldwide universal standard, such as Unicode, but the notion of attempting to list all of the world's glyphs and freeze them in a standard reduces flexibility and tends to overlook obscure or variant writing systems and restrict the development of new ones.

Possibly a standard should be developed that permits new character sets so long as the definition of the glyphs and the representation mapping is maintained in an accessible location.

4.0 Conclusion

In this unit you have studied the main ingredients of digital information organisation, how information can be digitized, the differences between digital and traditional libraries and interoperability of digital information.

5.0 Summary

In this unit, digital information organisation is the process of rendering large amounts of information into digital form so that it can be stored, retrieved and manipulated by computer. Systems employed in organisation of digital information are an amalgam of various technologies, including scanning, OCR, digital storage techniques, data compression, indexing and search algorithms, display devices and the Internet.

6.0 Tutor-Marked Assignment

1. Explain briefly what digital information organisation is
2. Enumerate on the merits and demerits of digital information
3. Compare digital with traditional libraries
4. Itemize the features of a digital information library
5. Discuss the interoperability of digital information.

7.0 References/Further Readings

1. Chowdhury G.G., (2003). *Introduction to Modern Information Retrieval*, Neal-Schuman
2. Meadow C.T., B.R. Boyce, D.H. Kraft, C.L. Barry. (2007). *Text Information Retrieval Systems*. Academic Press.
3. Jones K. S., P. Willett. (1997). *Readings in Information Retrieval*, Morgan Kaufmann.,
4. Raj Reddy et al (1999), Digital information Organisation in Japan, International Technology Research Institute, Baltimore, Maryland, USA at <www.wtec.org/pdf/dio.pdf>
5. Edward A. Fox, Overview of Digital Library Components and Developments at <<http://www.unm.edu/~jreenen/dlbook/chapter4.html>>

UNIT 3 DATA REPRESENTATION AND ORGANISATION ON INFORMATION SYSTEM

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Data Representation and Organisation on Information System
 - 3.2 Data Representation and Organisation on Web Information Space
 - 3.3 Metadata Creation
 - 3.3.1 Post-Publishing Representation
 - 3.3.1 Pre-Publishing Structuring
 - 3.4 Role of Metadata as Information Repositories
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider how data is represented and organised on information systems. You will also learn the importance of metadata and how they are created.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- explain what data representation and organisation on information system means
- describe the importance of metadata and how they are created
- discuss data representation and organisation on web information space.

3.0 MAIN CONTENT

3.1 Data Representation and Organisation on Information System

The notion of information representation and organization traditionally means creating catalogs and indexes for publications of any kind. It includes the description of the attributes of a document and the representation of its intellectual content.

Libraries in the world have a long history in recording data about documents and publications; such practice can be dated back to several thousand years ago. Indexes and library catalogs are created to help users find and locate a document conveniently. Records in the information searching tools not only serve as an inventory of human knowledge and culture but also provide orderly access to the collections.

The representation and organization of information in this network era has gone through dramatic changes in almost every stage of the process. The changes include not only the methods

and technology used to create records for publications, but also the standards that are central to the success and effectiveness of these tools in searching and retrieving information.

Today the library catalog is no longer a tool for its own collection for the library visitors; it has become a network node that users can visit from anywhere in the world via a computer connected to the Internet. The concept of indexing databases is no longer just for newspapers and journal articles; it has expanded into the web information space that is being used for e-publishing, e-businesses, and e-commerce.

3.2 Data Representation and Organisation on Web Information Space

The web is regarded as a universal information space. The heart of such a universal information space lies in the standards that make it possible for different types of data to be communicated and understood by heterogeneous platforms and systems. The well known Transmission Control Protocol/Internet Protocol (TCP/IP) allows different computer systems to talk to each other and to understand different dialects of networking language.

In the world of organizing information content, the content is represented by terms either in natural or controlled language or both. The characteristics of its container (book, journal, film, memo, report, etc.) will be encoded in certain format for computer storage and retrieval. The MACHine Readable Cataloging (MARC) has been used to encode information about their collections. In conjunction with cataloging rules, such MARC format standardized the record structure that describes information containers, i.e., books, manuscripts, maps, periodicals, motion pictures, music scores, audio/video recordings, 2-D and 3-D artifacts, and microforms.

Many organisation's library have embraced the Online Computer Library for their cataloging services. A well known online library is the Online Computer Library Center (OCLC) in Dublin, Ohio in the United States of America. It is about the largest and busiest cataloging service in the world. Over 30,000 libraries from 67 countries now use OCLC products and services. With the growth of the information space, e-publishing continues to thrive and libraries have expanded conventional cataloging of their collections into organizing the information on the Web.

3.3 Metadata Creation

The term "metadata" is used to mean the documentation about documents and objects. They describe resources, indicate where the resources are located, and outline what is required in order to use them successfully. An example of a metadata scheme is the OCLC's Metadata Initiative called *Dublin Core Metadata Initiative* inaugurated in 1995, which proposed a metadata scheme containing 15 data elements. Among them are title, creator, publisher, subject, description, format, type, source, relation, identifier, and rights.

Metadata schemes, such as Dublin Core, entail a group of codes or labels that describe the content and/or container of digital objects. When the metadata is embedded in hypertext documents, they can accommodate automatic indexing for digital objects and thus provide better aids in networked resource discovery. Several terms have been used interchangeably in describing the digital objects that a user views through various interfaces (e.g., a web browser).

They are given names such as Web document, Web object, digital object, hypertext, and hypermedia.

3.3.1 Post-Publishing Representation

Post-publishing representation is a method in which a special type of computer program generates metadata from digital objects already published. These programs are known as spiders, automatic robots, webcrawlers, wanderers, etc. Using these programs, metadata are extracted from the objects that were made available on the Internet.

Many of the Web search engines, e.g., Excite, Lycos, AltaVista, employ the post-publishing representation method to collect metadata and build their metadata bases for networked information discovery purposes. Although this fully automated process of metadata generation requires little or no human intervention, the methods used to extract metadata are too simple and far from effective in resource discovery.

The most appealing advantage of post-publication representation is probably that updating a metadata base can be done automatically and as frequently as one desires. This advantage makes it possible for popular search engines such as Yahoo! AltaVista, and HotBot to create dynamic metadata in response to queries. Since they do not generally retrieve the metadata content, results are created on the fly to answer users' queries. Another advantage comes with this automatic indexing process: the labour costs tend to be low because little or no human intervention is involved in the metadata harvesting process.

However, it is known that this automatic indexing is "less than ideal for retrieving an ever-growing body of information on the Web". Some of the factors that made this scheme a disadvantage, include:

- the inability to identify characteristics of a document such as its overall theme,
- lack of standards, and
- inadequate representation for images.

3.3.2 Pre-Publishing Structuring

One way to compensate for the shortcomings in post-publishing representation is through pre-publishing structuring. Pre-publishing structuring entails attaching structured metadata to the digital objects so that automated indexing programs can collect this information in a more efficient way. One of the earlier efforts in pre-publishing structuring of metadata is the Text Encoding Initiative (TEI) of the University of Virginia.

TEI is basically an encoding scheme consisting of a number of modules or Document Type Declaration (DTD) fragments, which include 3 categories of tag sets:

- core DTD fragments;
- base DTD fragments; and
- additional DTD fragments.

Another pre-publishing structuring project is the Encoded Archival Description (EAD) of the Library of Congress. EAD is an SGML document type definition for encoding finding aids for archival collections. Other domain-specific projects include the Content Standards for Digital Geospatial Metadata (CSDGM) of the Federal Geographic Data Committee, and the Government Information Locator Service (GILS). As of April 1998, there were over 40 projects in more than 10 countries that either use Dublin Core or are developing their own metadata element set that are based on Dublin Core.

The common element among these projects is that they embed the structured metadata into the Web objects prior to or after their publication. The structured metadata consists of components that allow establishing relationships among data elements with other entities, and these components are usually categorized into several different packages or layers.

A number of proposals have been made in structuring metadata. One of them is that "[meta]data elements must be described in a standard way as well as classified. Attribute standardization involves the specification of a standard set of attributes, and their allowable value ranges, independently of the application areas of data elements, tools, and implementation in a repository." Here five categories of attributes were outlined to include:

- identifying,
- definitional,
- relational,
- representational, and
- administrative,

all reflecting a complex structure in metadata elements.

Another proposal is that of a reference model for business-acceptable communication. This proposal defined clusters of data elements that would be required to fulfill a range of functions of a record. The functions of records are identified here as:

- The provision of access and use rights management
- Networked information discovery and retrieval
- Registration of intellectual property
- Authenticity, including handle, terms and conditions, structural, contextual content, and use history.

3.4 Role of Metadata as Information Repositories

Among the key concepts in digital information repositories, metadata plays two important roles:

- as a handler (i.e., identifier) and
- as points of access to data/document content.

As a handler, metadata also acts a locator, helping users to obtain the data or document by providing the exact location. As points of access to data/document content, metadata supplies information about the content of resources. The demand for effective organization of information does not diminish with powerful information technology, but rather, people nowadays have higher expectations for networked resources.

The success of a digital information repository in meeting such high expectations depends largely on the quality and scale of metadata, which, in turn, depends on a whole set of information processing standards and quality control management.

4.0 Conclusion

In this unit you have studied what data representation and organisation on information system entails. You also learnt the importance of metadata and how they are created.

5.0 Summary

In this unit, information representation and organization means creating catalogs and indexes for publications of any kind. The concept of indexing databases is no longer just for newspapers and journal articles; it has expanded into the web information space that is being used for e-publishing, e-businesses, and e-commerce. Metadata is the object that describes the information resources, indicates where the resources are located, and outline what is required in order to use them successfully.

6.0 Tutor-Marked Assignment

1. Explain briefly what data representation and organisation on information system entails
2. Describe the importance of metadata and how they are created
3. Discuss briefly data representation and organisation on web information space.

7.0 References/Further Readings

1. Baeza-Yates R., B. Ribeiro-Neto.(1999). *Modern Information Retrieval*, Addison-Wesley.
2. Jian Qin (2000), Representation and Organization of Information in the Web Space: From MARC to XML, *Information Science* Vol 3 No 2, pp 83 – 87.

Module 4 Information Retrieval Models

Unit 1 Retrieval of Bibliographic and Digital Information

Unit 2 Information Retrieval Models

Unit 3 Query Structure

Unit 4 User Profiles

UNIT 1 RETRIEVAL OF BIBLIOGRAPHIC AND DIGITAL INFORMATION**CONTENTS**

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Information Retrieval
 - 3.2 The Retrieval Process
 - 3.3 Retrieval of Bibliographic Information
 - 3.4 Evolution of Digital Libraries
 - 3.5 Retrieval of Digital Information
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider what information retrieval is, the retrieval process and how to retrieve bibliographic and digital information. You will also learn the evolution of digital information.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- explain how to retrieve both bibliographic and digital information
- enumerate on information retrieval
- itemize the retrieval process
- discuss briefly the evolution of digital libraries.

3.0 MAIN CONTENT**3.1 Information Retrieval**

Information Retrieval (IR) deals with the representation, storage, organisation of, and access to information items. The representation and organisation of information items should provide the user with easy access to the information in which he is interested.

Information retrieval is the science of searching for documents, for information within documents and for metadata about documents, as well as that of searching relational databases and the World Wide Web. Information retrieval is interdisciplinary, based on computer science,

mathematics, library science, information science, information architecture, cognitive psychology, linguistics, statistics and physics.

Automated information retrieval systems are used to reduce *information overload*. Many universities and public libraries use information retrieval systems to provide access to books, journals and other documents. Web search engines are the most visible information retrieval applications.

The effective retrieval of relevant information is directly affected both by the *user task* and the *logical view of the documents*. The user task may be information retrieval or information browsing. Classic information retrieval systems (such as web interfaces) normally allow information retrieval alone, while hypertext systems provide quick browsing. Modern digital library and Web interfaces might attempt to combine these tasks to provide improved retrieval capabilities.

The logical view of the document is provided by representative keywords or index terms, which are frequently used historically to represent documents in a collection. In modern computers, retrieval systems adopt a full text logical view of the document. However, with very large collections, the set of representative keywords may have to be reduced. This process of reduction or compression of the set of representative keywords is called *text operations* (or transformation).

Text operations can be accomplished through a number of processes including:

- the elimination of *stopwords* (such as articles and connectives)
- the use of *stemming* (which reduces distinct words to their common grammatical root)
- the identification of noun groups (which eliminates adjectives, adverbs and verbs)

Text operations reduce the complexity of the document representation and allow moving the logical view from that of a full text to that of a set of index terms.

3.2 The Retrieval Process

To describe the retrieval process, we use a simple and generic software architecture as shown in figure 1 below. Steps in the Retrieval Process are the following:

- define the text database (specification of document, operation and text model)
- text operations transform original documents and generate a logical view of documents
- database manager builds an index of the text (using the database Manager Module)
- the retrieval process is initiated
- the user specifies a *user need*, which is described and transformed by same text operations applied to the text
- *query operations* (which provides a system representation for the user need) is generated
- the query is then processed to obtain the *retrieved documents*
- the retrieved documents are ranked according to a *likelihood* of relevance
- user examines the set of ranked documents in the search for useful information
- user might pinpoint a subset of the documents of interest and initiate a *user feedback cycle*

- the system uses the documents selected by the user in such a cycle to change the query formulation (this modified query may be a better representation of the real user need)

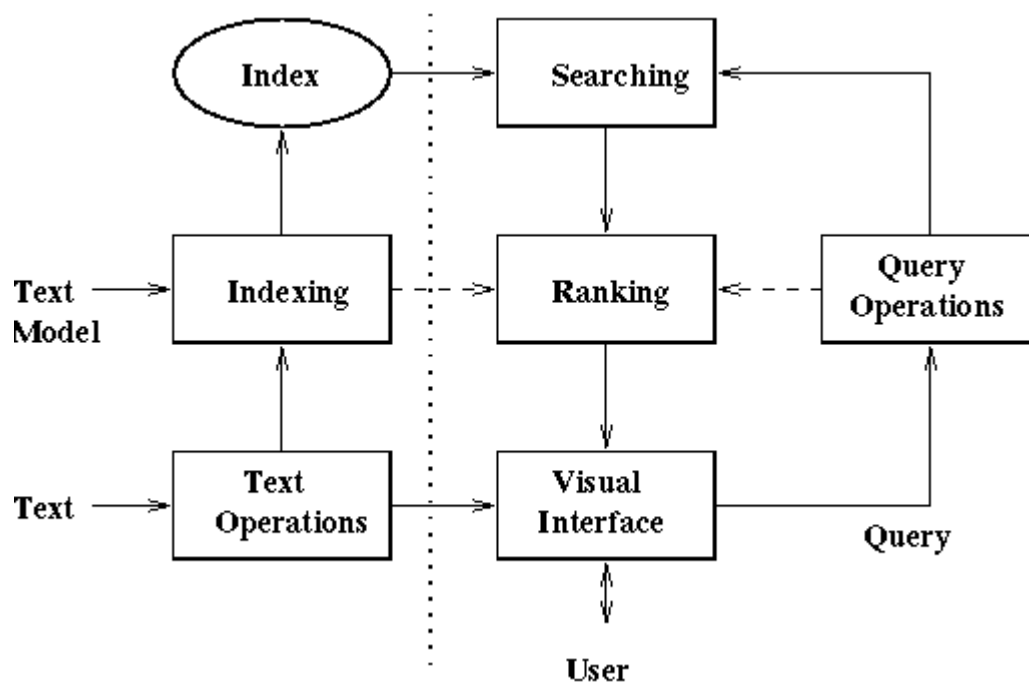


Figure 1: The Retrieval Process

3.3 Retrieval of Bibliographic Information

In the historical evolution of digital or online libraries, the mechanisms for retrieval of scientific literature have been particularly important. The first attempts at realization of a mechanism for retrieval of scientific literature occurred during the 1960s and centered around text search of technical citations. The content was the text of a bibliographic citation of a journal article, which included the title, author, journal, and keywords of the referenced article. A search query was matching specified words to words in the fields of the citation.

At this time, retrieval of bibliographic information was largely determined by the hardware capability. Therefore, the disk space constrained the collection of citations rather than the complete text of articles. The output device was a paper teletypewriter, mandating a short display, such as a title or citation. The network was a telephone line, used as one of the first packet-switched networks. Thus, the interface goal was to enable the specification of a precise query to retrieve a particular set of items from the citation database and return them to be printed on the teletypewriter terminal. The items were not yet actual documents but pointers to physical documents.

These were the bibliographic information retrieval systems and so well suited for generating a bibliography. Example, exhaustively printing all articles that contained the keywords "information retrieval" and "computer networks" for the references of a paper. The systems' speed was slow and queries precise. Therefore, their primary users were professional librarians generating bibliographies for scientists. Although the systems were intended for searching and locating desired items, their slow speed and precise queries limited their effectiveness in browsing.

The collections of citations that the retrieval systems handled were known as *bibliographic databases*. The bibliographic databases were extended over time to include searchable abstracts of the articles. Examples of these bibliographic databases are *MEDLINE* in biology and medicine and *Inspec* in electrical engineering and computer science.

3.4 Evolution of Digital Libraries

Organized collections of scientific materials are traditionally called "libraries," and the searchable online versions of these are called "digital libraries". The primary purpose of digital libraries is to enable searching of electronic collections distributed across networks, rather than merely creating electronic repositories from digitized physical materials. A digital library enables users to interact effectively with information distributed across a network. These network information systems support search and display of items from organized collections.

In the historical evolution of digital or online libraries, the mechanisms for retrieval of scientific literature have been particularly important. *Grand visions* (begun with a visionary article by Vannevar Bush published in 1945, just as World War II was ending) for the development of digital libraries in 1960 led first to the development of *text search*, from bibliographic databases to full-text retrieval. Next, research prototypes catalyzed the rise of *document search*, from multimedia browsing across local-area networks to *distributed search* on the Internet.

Traditionally, information retrieval has been a task for professional librarians. Nowadays, public computer networks have been used to access specialized information services. However, it has taken the recent rise of the Internet to make literature searching directly available to widespread groups of scientists.

Since the beginnings of online information retrieval more than 30 years ago, the base functionality has remained essentially unchanged. A collection of literature is maintained and indexed, which the user accesses by means of a terminal connected to a server across a network. The user specifies a query by a set of words, and all documents in the collection that contain those words are returned.

3.5 Retrieval of Digital Information

By the 1980s, full-text (or document) search rather than text (or citation) search had become established in retrieval systems. This was made possible by the digitization of information objects. This same era saw the initial deployment of bitmapped personal workstations and local-area networks in research laboratories and other industries.

With time, computer model changed from central shared mainframes to distributed personal workstations. This also changed information retrieval from text search to document search. As the research workstations of the 1980s turned into the personal computers of the 1990s and internet access became widespread, the research systems of the 1980s based on full-text technology became the internet services of the 1990s. Thus, full-text search coupled with multimedia browsing is today available to average scientists for their everyday needs.

The speed development of both the workstations and the networks increased. This brought an expansion in both the basic document and the basic retrieval. Multimedia gradually became possible, so that pictorial materials, such as graphics, images, and videos, could be included in the documents and accessed from collections across the network. For example, interactive display of color pictures from remote sources became technologically feasible.

The increased speed across the network meant that multiple sources could be searched within a single query while still maintaining effective user interaction for the return of results. Multiple collections could be stored in physically distributed locations, yet searched as a single, logically coherent collection. Thus the interactive information gateway technology pioneered in the 1970s was finally realized and then commercialized in the 1980s.

More profoundly, a different style of interaction became possible with the increased speeds. Rather than search, where a detailed query is made and comprehensive results returned, browsing enables broad queries to be used to quickly scan for appropriate sections of a digital library. This style resembles using the card catalog to locate a particular section of a physical library, then browsing those shelves in search of suitable materials. The underlying search mechanism is the same--full-text proximity--but any results returned can be scanned much more quickly.

4.0 Conclusion

In this unit you have studied how to retrieve both bibliographic and digital information. You also learnt the retrieval process and the evolution of digital libraries.

5.0 Summary

In this unit, information retrieval is the science of searching for documents, for information within documents and for metadata about documents, as well as that of searching relational databases and the World Wide Web. Automated information retrieval systems are used to reduce *information overload*. The effective retrieval of relevant information is directly affected both by the *user task* and the *logical view of the documents*.

6.0 Tutor-Marked Assignment

1. Enumerate on information retrieval
2. Explain briefly how to retrieve bibliographic
3. Itemize the retrieval process
4. Discuss briefly the evolution of digital libraries.

7.0 References/Further Readings

1. Baeza-Yates R., B. Ribeiro-Neto.(1999). *Modern Information Retrieval*, Addison-Wesley.
2. Bruce R. Schatz (1997) Information Retrieval in Digital Libraries: Bringing Search to the Net, *Science*, Vol. 275. No. 5298, pp. 327 – 334 at
<<http://www.sciencemag.org/cgi/content/full/275/5298/327>>
3. Information retrieval, *Wikipedia, the free encyclopedia* at
<http://en.wikipedia.org/wiki/Information_retrieval>
4. Ricardo Baeza-Yates and Ribeiro-Neto (1999), *Modern Information Retrieval*, ACM Press New York, USA at <www2.dcc.ufmg.br/livros/irbook/>

UNIT 2 INFORMATION RETRIEVAL MODELS

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Information Retrieval Models
 - 3.1 Introduction
 - 3.2 Boolean model
 - 3.3 Document similarity
 - 3.4 Probabilistic indexing
 - 3.5 Vector space model
 - 3.6 Probabilistic retrieval
 - 3.7 Fuzzy set models
 - 3.7.1 The Fuzzy Model of Information Retrieval
 - 3.8 Inference networks
 - 3.9 Language models
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider what information retrieval models are. You will also learn the various types of information retrieval models.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- explain what information retrieval models are
- enumerate on the various types information retrieval models

3.0 INFORMATION RETRIEVAL MODELS

3.1 Introduction

The goal of information retrieval (IR) is to provide users with those documents (including non-textual information, such as multimedia objects) that will satisfy their information need. Users have to formulate their information need in a form that can be understood by the retrieval system.

Information seeking is a form of problem solving. It proceeds according to the interaction among eight subprocesses:

- problem recognition and acceptance,
- problem definition,
- search system selection,
- query formulation,

- query execution,
- examination of results (including relevance feedback),
- information extraction, and
- reflection/iteration/termination.

To be able to perform effective searches, users have to develop the following expertise:

- knowledge about various sources of information,
- skills in defining search problems and applying search strategies, and
- competence in using electronic search tools.

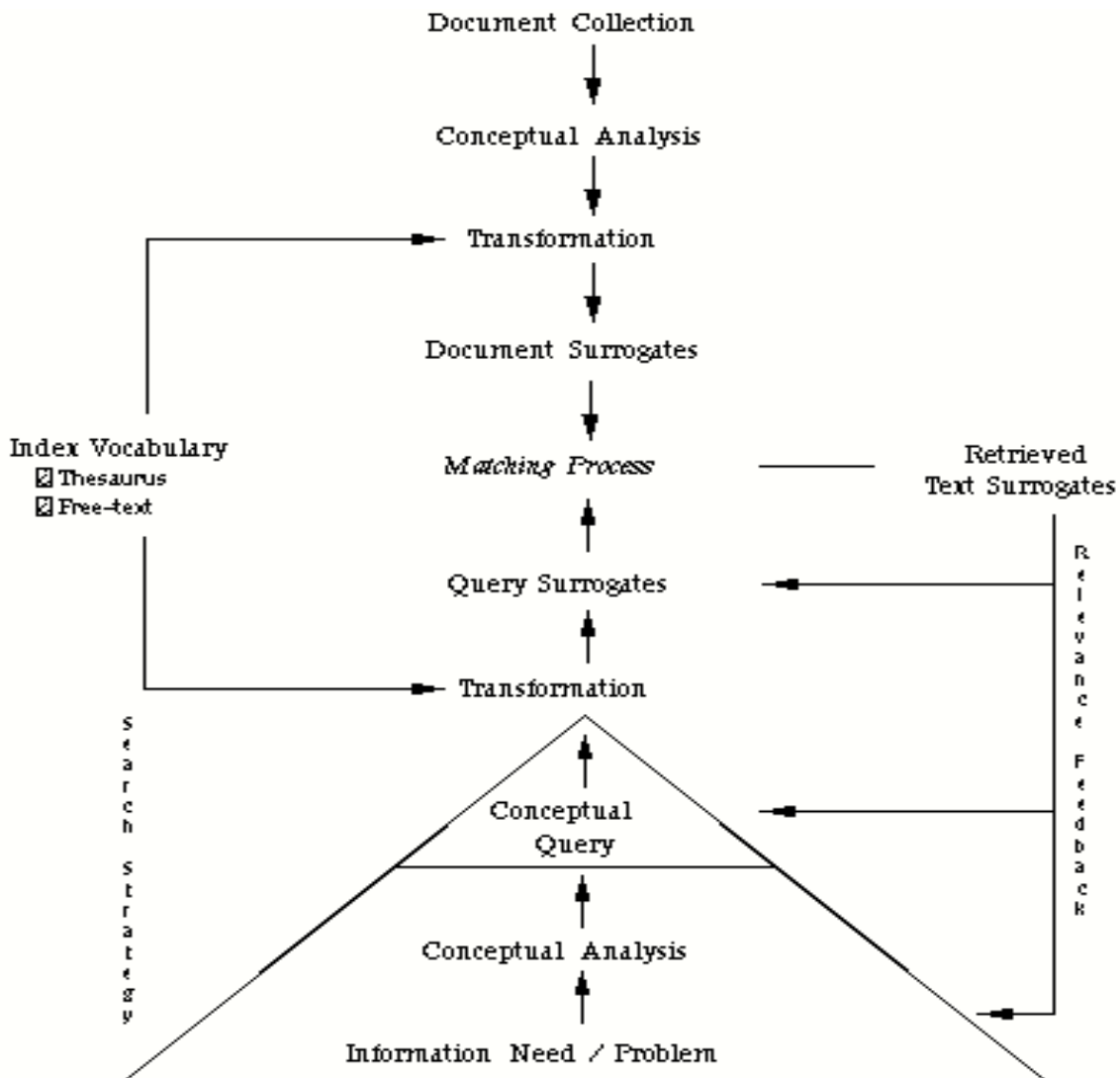


Figure 1: Representation of a General Model of the Information Retrieval process

The information need of the user can be understood as forming a pyramid, where only its peak is made visible by users in the form of a conceptual query (figure 1). The conceptual query captures the key concepts and the relationships among them. It is the result of a conceptual analysis that operates on the information need, which may be well or vaguely defined in the user's mind. This analysis can be challenging, because users are faced with the general "vocabulary problem" as they are trying to translate their information need into a conceptual query.

The vocabulary problem refers to the fact that a single word can have more than one meaning, and, conversely, the same concept can be described by surprisingly many different words. It has been shown that two people use the same main word to describe an object only 10 to 20% of the time. Further, the concepts used to represent the documents can be different from the concepts used by the user. The conceptual query can take the form of a natural language statement, a list of concepts that can have degrees of importance assigned to them, or it can be statement that coordinates the concepts using Boolean operators. Finally, the conceptual query has to be translated into a query surrogate that can be understood by the retrieval system.

In order to effectively retrieve information, a number of models have been developed. A model is an abstracts away from the real world, which uses a branch of mathematics and possibly a metaphor for searching. The well known information retrieval models with their year of proposal include:

- 4 Boolean model (± 1950)
- 5 Document similarity (± 1957)
- 6 Probabilistic indexing (± 1960)
- 7 Vector space model (± 1970)
- 8 Probabilistic retrieval (± 1976)
- 9 Fuzzy set models (± 1980)
- 10 Inference networks (± 1992)
- 11 Language models (± 1998)

3.2 Boolean Model

The *Boolean model* of information retrieval is a classical information retrieval (IR) model and, at the same time, the first and most adopted one. Proposed about 1950, it is used by virtually all commercial information retrieval systems today.

The Boolean information retrieval is based on *Boolean Logic* and classical *Sets Theory* in that both the documents to be searched and the user's query are conceived as sets of terms. Retrieval is based on whether or not the documents contain the query terms. Given a finite set

$$T = \{t_1, t_2, \dots, t_j, \dots, t_m\} \quad \dots 1$$

of elements called index terms (e.g. words or expressions, which may be stemmed describing or characterising documents such as– keywords given for a journal article).

Advantages of the Boolean model include:

- It is easy to implement and it is computationally efficient. Hence, it is the standard model for the current large-scale, operational retrieval systems and many of the major on-line information services use it.
- It enables users to express structural and conceptual constraints to describe important linguistic features. Users find that synonym specifications (reflected by OR-clauses) and phrases (represented by proximity relations) are useful in the formulation of queries.
- The Boolean approach possesses a great expressive power and clarity. Boolean retrieval is very effective if a query requires an exhaustive and unambiguous selection.
- The Boolean method offers a multitude of techniques to broaden or narrow a query.
- The Boolean approach can be especially effective in the later stages of the search process, because of the clarity and exactness with which relationships between concepts can be represented.

On the hand the disadvantages of Boolean model include the fact that users find it difficult to construct effective Boolean queries for several reasons. Users are using the natural language terms AND, OR or NOT that have a different meaning when used in a query. Thus, they will make errors when they form a Boolean query, because they resort to their knowledge of English.

3.3 Document similarity

The *Document Similarity* model of information retrieval is a statistical proposed in 1957 queries how similar each document is to every other document. The *Information Theoretic Measure* is applied. This measure stated mathematically is thus:

$$\text{sim}(A, B) = \frac{A \cap B}{A \cup B} \quad \dots 2$$

From the above equation, the model tries to answer the questions:

- 1) How is every pair of documents intersected without sacrificing efficiency?
- 2) What features should be intersected? Words or Phrases.

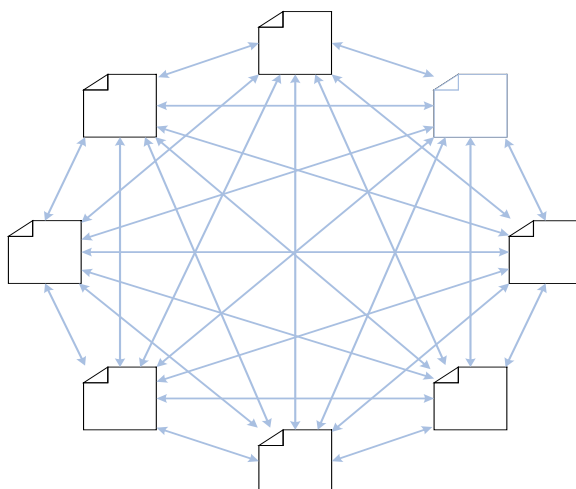


Figure 2: Document Similarity

In Document Similarity model, the system automatically clusters or groups similar documents together. In doing this, intra-cluster similarity is preferred to inter-cluster similarity (figure 3). Therefore in Document Similarity, the principle of similarity is applied. This principle states that *the more two representations agree in given elements and their distribution, the higher would be the probability of their representing similar information.*

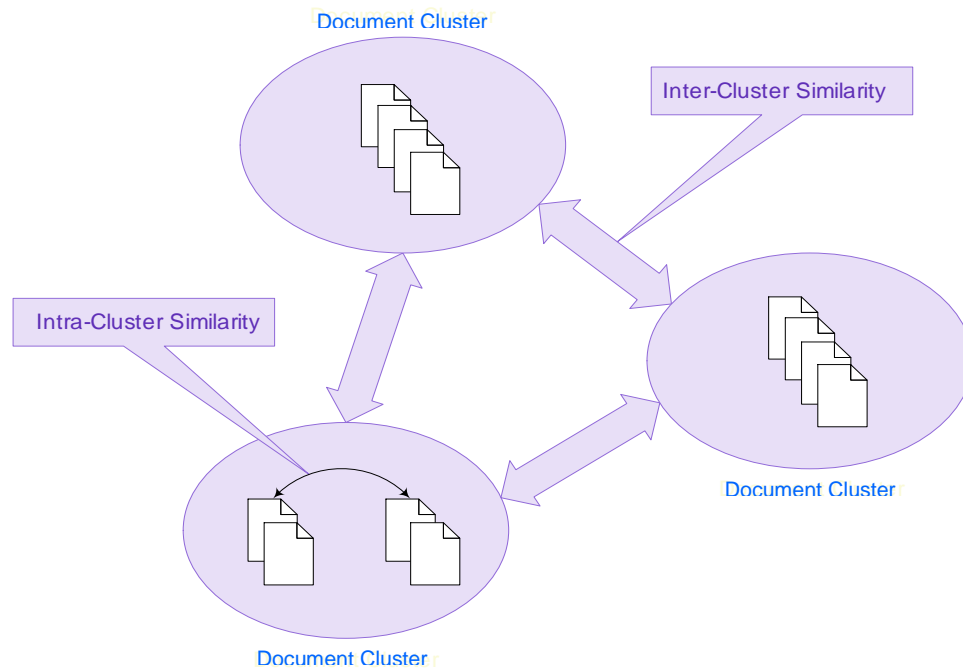


Figure 3: Document Clustering

3.4 Probabilistic indexing 1960

Probabilistic indexing is a retrieval model, proposed in 1960, in which the difference in the distributional behaviour of words was considered as a guide to whether a word should be assigned as an index term. It was shown that the statistical behaviour of 'speciality' words was different from that of 'function' words and that 'function' words were closely modelled by a Poisson distribution over all documents whereas specialty words did not follow a Poisson distribution.

Specifically, if one is looking at the distribution of a function word w over a set of texts then the probability, $f(n)$, that a text will have n occurrences of the function word w is given by

$$f(n) = e^{-x} x^n / n \quad \dots\dots 3$$

In general the parameter x will vary from word to word, and for a given word should be proportional to the length of the text. We also interpret x as the mean number of occurrences of the w in the set of texts.

The Probabilistic indexing model assumes that specialty words are 'content-bearing' whereas function words are not. What this means is that a word randomly distributed according to a Poisson distribution is not informative about the document in which it occurs. At the same time, the fact that a word does *not* follow a Poisson distribution is assumed to indicate that it conveys information as to what a document is about. This is not an unreasonable view. For instance, knowing that the specialty word WAR occurs in the collection one would expect it to occur only in the relatively few documents that are about WAR. On the other hand, one would expect a typical function word such as FOR to be randomly distributed.

3.5 Vector space model

The *Vector space model*, proposed in 1970, is a statistical retrieval model and represents the documents and queries as vectors in a multidimensional space, whose dimensions are the terms used to build an index to represent the documents. The creation of an index involves lexical scanning to identify the significant terms, where morphological analysis reduces different word forms to common *stems*, and the occurrence of those stems is computed.

Query and document surrogates are compared by comparing their vectors, using, for example, the cosine similarity measure. In this model, the terms of a query surrogate can be weighted to take into account their importance, and they are computed by using the statistical distributions of the terms in the collection and in the documents. The vector space model can assign a high ranking score to a document that contains only a few of the query terms if these terms occur infrequently in the collection but frequently in the document.

The vector space model, like all statistical retrieval models have the following advantages:

- They provide users with a relevance ranking of the retrieved documents. Hence, they enable users to control the output by setting a relevance threshold or by specifying a certain number of documents to display.
- Queries can be easier to formulate because users do not have to learn a query language and can use natural language.
- The uncertainty inherent in the choice of query concepts can be represented.

However, the vector space model as well as other statistical models has the following shortcomings:

- They have a limited expressive power. For example, the NOT operation cannot be represented because only positive weights are used. Again, the very common and important Boolean query ((A and B) or (C and D)) cannot be represented by a vector space query. Hence, the statistical approaches do not have the expressive power of the Boolean approach.
- The statistical approach lacks the structure to express important linguistic features such as phrases. Proximity constraints are also difficult to express, a feature that is of great use for experienced searchers.
- The computation of the relevance scores can be computationally expensive.
- A ranked linear list provides users with a limited view of the information space and it does not directly suggest how to modify a query if the need arises.

- The queries have to contain a large number of words to improve the retrieval performance. As is the case for the Boolean approach, users are faced with the problem of having to choose the appropriate words that are also used in the relevant documents.

The vector space model, on its own makes the following assumptions:

- The more similar a document vector is to a query vector, the more likely it is that the document is relevant to that query.
- The words used to define the dimensions of the space are orthogonal or independent. While it is a reasonable first approximation, the assumption that words are pairwise independent is not realistic.

3.6 Probabilistic retrieval

The *Probabilistic retrieval* model is a statistical model proposed in 1976 based on the Probability Ranking Principle, which states *that an information retrieval system is supposed to rank the documents based on their probability of relevance to the query, given all the evidence available.* The principle takes into account that there is uncertainty in the representation of the information need and the documents. There can be a variety of sources of evidence that are used by the probabilistic retrieval model, and the most common one is the statistical distribution of the terms in both the relevant and non-relevant documents.

Probabilistic retrieval model has the same general characteristic (advantages/shortcomings) as the other statistical retrieval models.

3.7 Fuzzy set model

The Boolean model had imposed a binary criterion for deciding relevance. Thus, the question of how to extend the Boolean model to accommodate partial matching and a ranking has been a problem for some time now. This question was answered by the introduction of the fuzzy set model in 1980.

The theory of fuzzy sets is a generalization of classical set theory for modeling vagueness and uncertainty. The *vagueness* can be modeled using a fuzzy framework, such that with each term is associated a *fuzzy* set and each document has a degree of membership in this fuzzy set. This interpretation provides the foundation for many models for information retrieval based on fuzzy theory.

The notion of fuzzy sets provides a convenient tool for representing classes whose boundaries are not well defined. The key idea here is to introduce the notion of a *degree of membership* associated with the elements of a set. This degree of membership varies from 0 to 1 and allows modelling the notion of *marginal* membership. That is to say that the Fuzzy set is a function that maps a value, which might be a member of a set, to a number between zero and one, indicating its actual degree of membership. A degree of zero means that the value is not in the set, and a degree of one means that the value is completely representative of the set.

A fuzzy set A , which is a subset of a universal set U is characterized by a membership function $\mu_A: U \rightarrow [0,1]$ which associates with each element u of U a number $\mu_A(u)$ in the interval $[0,1]$.

The complement of the fuzzy set is given by $\mu_{-A}(u) = 1 - \mu_A(u)$

The union of the fuzzy set is given by $\mu_{A \sqcup B}(u) = \max(\mu_A(u), \mu_B(u))$

The intersection of the fuzzy set is given by $\mu_{A \sqcap B}(u) = \min(\mu_A(u), \mu_B(u))$

3.7.1 The Fuzzy model of information retrieval

The Fuzzy model of information retrieval has as its basic idea to expand the set of index terms in a query with related terms (from the thesaurus) such that additional relevant documents can be retrieved. A thesaurus can be constructed by defining a term-term correlation matrix c whose rows and columns are associated to the index terms in the document collection.

A fuzzy model, like traditional Expert and Decision Support System, is based on the input, process, output flow concept (figure 4). A fuzzy model however differs from the traditional Expert and Support System in two important properties:

- what flows into and out of the process, and
- the fundamental transformation activity embodied in the process itself

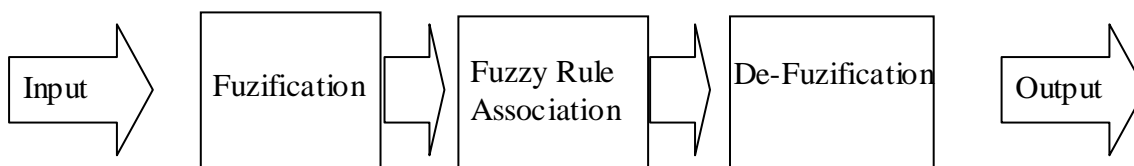


Figure 4: Information flow through a Fuzzy System

When a query is made for the address of a Person the archived data is clustered according to the various criteria, e.g., by similar street names, within the same zip code or by similar last name. It constructs and attaches to a window description of a set expression as for an example: $(\text{Cluster1} \cap \text{Cluster3}) \cup (\text{Cluster2} \cap \text{Cluster3})$. Here several properties of clusters are relevant. Each cluster entry is a key value followed by a set of archived record numbers.

3.8 Inference Networks

The *Inference Network* retrieval framework is a robust model from the field of information retrieval based on the formalism of Bayesian networks. The Bayesian network is an acyclic directed graph that encodes probabilistic dependency relationships between random variables. A directed graph is acyclic if there is no directed path $A \cdots Z$ such that $A = Z$. The presentation of probability distributions as directed graphs makes it possible to analyze complex conditional independence assumptions by following a graph theoretic approach. Probability theory ensures that the system as a whole is consistent.

The Bayesian network of figure 5 shows Turtle's simple model of the relevance of a document

given a query of three non-equal terms, say the query social political economic. All nodes in the network represent binary random variables with values $\{0, 1\}$. For the event “query is fulfilled” let $(Q = 1)$ has three possible causes:

- the subject referred to by the term social is true ($T_1 = 1$), or
- the subject referred to by the term political is true ($T_2 = 1$), or
- the subject of economic is true ($T_3 = 1$),

or a combination of the three causes.

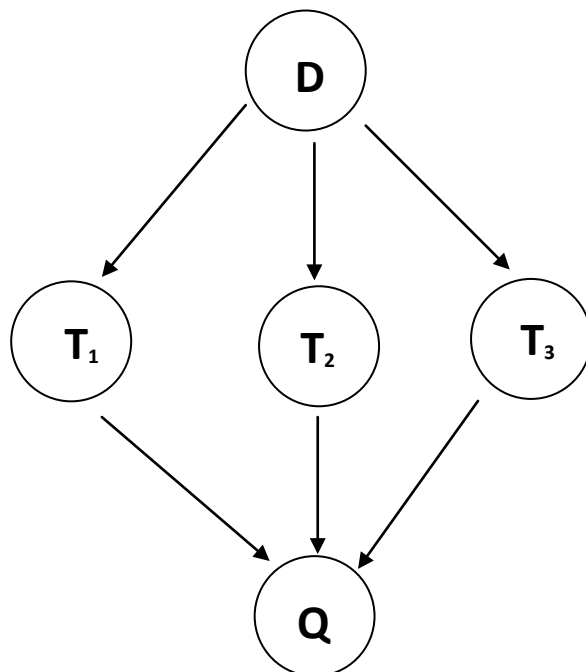


Figure 5: Simple Bayesian network

The three subjects in turn are inferred from the event “document is relevant” ($D = 1$). By the chain rule of probability, the joint probability of all the nodes in the graph above is:

$$P(D, T_1, T_2, T_3, Q) = P(D)P(T_1|D)P(T_2|D, T_1)P(T_3|D, T_1, T_2)P(Q|D, T_1, T_2, T_3)$$

The second, third and fourth term in the above equation are simplified because T_1 , T_2 and T_3 are independent given their parent D . The last term is simplified because Q is independent of D given its parents T_1 , T_2 and T_3 . The directions of the arcs suggest the dependence relations between the random variables. The model makes the following conditional independence assumptions:

$$P(D, T_1, T_2, T_3, Q) = P(D) P(T_1|D) P(T_2|D) P(T_3|D) P(Q|T_1, T_2, T_3)$$

Now, the network should be used as follows: If it is hypothesized that the document is relevant ($D = 1$), the probability of query fulfilment $P(Q = 1|D = 1)$ can be used as a score to rank the documents.

The main advantages of the model are that:

- it allows structured, weighted queries to be evaluated,
- multiple document representations, and
- efficient inference.

Two disadvantages of the Inference network models include:

- the models do not suggest how the probability measures $P(T_i|D)$, ($1 \leq i \leq n$) should be estimated. Instead, the approaches suggest the use of Bayesian probabilities. That is, the Bayesian probability of an event is a person's degree of belief in that event, which does not have to refer to a physical mechanism or experiment.
- the calculation of the probabilities generally takes exponential time in the number n of non-equal query terms. The introduction of the four canonical forms solves this problem, but it could have been solved by the network topology.

3.9 Language Models

A statistical language model assigns a probability to a sequence of m words $P(w_1, \dots, w_m)$ by means of a probability distribution. Language modeling is used in many natural language processing applications such as speech recognition and information retrieval. When used in information retrieval, a language model is associated with a document in a collection. With query Q as input, retrieved documents are ranked based on the probability that the document's language model would generate the terms of the query, $P(Q|M_d)$.

The root of statistical language modeling dates back to the beginning of the 20th century when Markov tried to model letter sequences in works of Russian literature. However, modern language models for information retrieval was effectively proposed in 1998. The first uses of language modeling approach for information retrieval focused on its empirical effectiveness using simple models. In the basic approach, a query is considered generated from an "ideal" document that satisfies the information need. The system's job is then to estimate the likelihood of each document in the collection being the ideal document and rank them accordingly.

The basic model has been extended in a variety of ways. For example, documents have been modeled as mixtures of topics and phrases are considered. Progress has also been made in understanding the formal underpinnings of the statistical language modeling approach, and comparing it to traditional probabilistic approaches.

Successful applications of the LM approach to a number of retrieval tasks have also been reported, including cross-lingual retrieval and distributed retrieval. Research has shown that the language modeling approach is a very effective probabilistic approach for studying information retrieval problems.

A statistical language model is a probability distribution over all possible sentences or other linguistic units in a language. It can also be viewed as a statistical model for generating text. The task of language modeling, in general, answers the question: how likely the i th word in a sequence would occur given the identities of the preceding $i-1$ words? In most applications of

language modeling, such as speech recognition and information retrieval, the probability of a sentence is decomposed into a product of *n-gram* probabilities.

The basic approach for using language models for information retrieval assumes that the user has a reasonable idea of the terms that are likely to appear in the “ideal” document that can satisfy his/her information need, and that the query terms the user chooses can distinguish the “ideal” document from the rest of the collection. The query is thus generated as the piece of text representative of the “ideal” document. The task of the system is then to estimate, for each of the documents in the collection, which is most likely to be the ideal document.

4.0 Conclusion

In this unit you have studied a number of information retrieval models developed to retrieve information. The characteristics of the various types of information retrieval models were considered with their merits and demerits.

5.0 Summary

In this unit, information retrieval models are approaches developed to effectively retrieve information. A model is an abstract away from the real world, which uses a branch of mathematics and possibly a metaphor for searching. The well known information retrieval models include Boolean model, Document similarity, Probabilistic indexing, Vector space model, Probabilistic retrieval, Fuzzy set models, Inference networks and Language models.

6.0 Tutor-Marked Assignment

1. Enumerate on information retrieval models
2. Explain briefly the various types of information retrieval models

7.0 References/Further Readings

1. Croft W.B., J. Lafferty. (2003). *Language Modeling for Information Retrieval*. Springer.
2. Jones K. S., P. Willett. (1997). *Readings in Information Retrieval*, Morgan Kaufmann.,
3. Kowalski G., M.T. Maybury. (2005). *Information Storage and Retrieval Systems*, Springer.
4. van Risjbergen C.J., (2004). *The Geometry of Information Retrieval*, Cambridge UP.
5. Djoerd Hiemstra, A tutorial to formal models of information retrieval at <<http://www-clips.imag.fr/mrim/essir03/PDF/5.Hiemstra.pdf>>
6. Gina-Anne Levow (2003) Information retrieval: Models and Methods at <<http://ttic.uchicago.edu/~dmcalister/course/lec8.ppt>>
7. InfoCrystal: A Visual Tool For Information Retrieval, Ph.D. Research and Thesis at MIT (1995) Chapter 2 - Information Retrieval Models at <http://comminfo.rutgers.edu/~aspoerri/InfoCrystal/Ch_2.html>
8. Advanced information retrieval Chapter. 02: Modeling (Set Theoretic Models) – Fuzzy model at <<ce.sharif.edu/courses/8485/2/ce324/.../root/.../Chap02cFuzzy.ppt>>
9. Fuzzy Logic Information Retrieval Model at <<www.tcnj.edu/~mmartin/.../FuzzyLogicIRM/FuzzyLogic.ppt>>
10. Xiaoyong Liu and W. Bruce Croft, Statistical Language Modeling for Information Retrieval at <<ciir.cs.umass.edu/pubfiles/ir-318.pdf>>

UNIT 3 QUERY STRUCTURE

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Query Structure
 - 3.1 What is a Query?
 - 3.2 Query Formulation
 - 3.2.1 Factors Affecting Query Formulation
 - 3.3 Query Expansion
 - 3.4 Structures of Queries
 - 3.4.1 Query Formulation and Expansion
 - 3.4.2 Weak Query Structures
 - 3.4.3 Strong Query Structures
 - 3.5 Retrieval of Digital Information
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider what constitutes a query. You will also learn the about how query is formulated, what it means to expand a query and finally discrete structures of queries.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- explain what constitutes a query
- enumerate how a query is formulated
- itemize the processes that lead to query expansion
- discuss briefly structures of queries.

3.0 QUERY STRUCTURE

3.1 What is a Query?

In general, a query is a question, often required to be expressed in a formal way. The word derives from the Latin *quaere* (the imperative form of *quaerere*, meaning to ask or seek). In the field of information retrieval, a query is what a user of a search engine or database enters.

A query can be a broad or a narrow query. Broad query is defined as a query in which the terms are more general than the terms in the search request or task. Example, for the search task *How much blood on average goes through the heart in one minute?* the query “heart structure” is a broad query. Again, if only half or less than half of the aspects of a multi-aspect task are included in the query, the query is defined as broad. In practice, the broad queries usually require the user

to browse through the results or refine the query in order to find an answer to the question or to improve the relevance of the retrieved documents. This is called query expansion.

On the other hand, narrow query is a query in which the terms are specific or precise to the terms in the search task. Narrow or precise queries are targeted at finding relevant documents immediately, without needing to navigate. Here, the aim is to maximize the relevance of the retrieved documents. For a query to be defined precise in a multi-aspect task, all of the main aspects of the task need to be covered in the query. For example, for the task *How large were the economical losses in the September 11th 2001 terrorist attack directed at Pentagon in the USA?* the aspects are “economics”, “losses”, “2001 terrorist attack”, “Pentagon in the USA”. If only terrorist attack to Pentagon is present in the query, the query is considered imprecise or broad. Sometimes the aspects are presented in the query incompletely.

Every query is represented in the Query panel as a tree consisting of several nodes. A valid query includes at least 2 nodes:

- the root node of the tree, and
- a concept node

The *root node* does not play any particular role in the query and is not editable. The *concept node*, however, constitutes the foundation upon which the query will expand and is essentially the root of the query (but not the root of the tree). This concept node represents a specific concept that we wish to search for. More specifically, we are interested in retrieving objects present in the knowledge base whose type is determined by the concept described in the concept node.

3.2 Query Formulation

Query Formulation is simply the process by which a user or searcher defines his/her information needs. This means that in every query formulation technique there is a human in the loop. From very simple or narrow queries to extremely complex or broad queries, there must be a person to define the information need in the form of a query.

Query formulation is an essential part of successful information search and retrieval. It is typically based on search keys given by the user and is a major step in the complex process of information search. It therefore poses a huge challenge to users to formulate effective queries for their web information search. This is more so, given that the web is used by a diverse population varying in their levels of expertise.

Information search consists of four main steps: problem identification, need articulation, query formulation, and results evaluation. This process of information search is affected by numerous factors including, environmental (e.g., the database and the search topic), user or searcher (e.g., online search experience), search process (e.g., commands used), and search outcome variables (e.g., precision and recall).

3.2.1 Factors Affecting Query Formulation

It is known from studies that three main factors affect query formulation:

- Media expertise, including familiarity with the search environment, search engine expertise, computer expertise and expertise in information retrieval
- Domain expertise
- Type of search task

In media (web) expertise, the more experienced web user the searcher is, the more likely he/she is to use a “straight to information” search style (or narrow query) rather than a broader “navigating to information” style (or broad query). It is assumed that the more experienced the user becomes with the web, the more the search engines become a tool, rather than something to spend time with. Thus, the users want to search information as efficiently as possible and not just to see if something interesting happens to come by. The web experience also makes the users understand the structure of the material and the usual style of writing in the web. These are important skills for successful information search

Domain expertise presumably helps people in query formulation by giving them a possibility to use either more terms in their queries (synonyms), or possibly fewer, but more accurate terms. Thus, domain expertise is not directly expected to lead to longer queries, but the quality of the selected terms is expected to be high.

Search task is divided into three broad categories: fact-finding, exploratory, and comprehensive search tasks. In *fact-finding*, the source of information is not a key issue, but precision of the result set is a key issue for efficient search. On the other hand, in *exploratory* search tasks, the searcher’s aim is to obtain a general idea of the search topic or possibly to retrieve a couple of documents as an example. So here, high precision of the result set is not necessarily the most important thing. When the task is to find as many documents as possible on a given topic, then *comprehensive* search task is indicated. In this case, the recall should be as high as possible for the search to be successful.

3.3 Query Expansion

Query expansion is the process of reformulating a seed query to improve retrieval performance in information retrieval operations. In the context of web search engines, query expansion involves evaluating a user's input (what words or other types of data were typed into the search query area) and expanding the search query to match additional documents. This is because the original query does not always give satisfactory results.

Typical query expansion involves techniques such as:

- Finding synonyms of words, and searching for the synonyms as well
- Finding all the various morphological forms of words by stemming each word in the search query
- Fixing spelling errors and automatically searching for the corrected form or suggesting it in the results
- Re-weighting the terms in the original query

It has been shown that many information users or searchers are not always clear about what they are looking for and even if they are, they are not always sure about how to formulate that information need. In information retrieval systems one of the ways in which this can be accommodated to some degree is allowing the user to expand the initial query (query expansion), possibly after some relevant documents have been found, in order to enrich the query specification.

The purpose of query expansion is to make the query resemble more closely the relevant documents and thus, to retrieve those relevant documents. Therefore, query expansion could mean adding or deleting terms from the original query or even changing terms. This can be done using information from relevance feedback with relevant documents identified manually by the user or by assuming the top-ranked documents from an initial ranking are relevant.

3.4 Structures of Queries

Query structure means the use of operators to express the relations between search keys. The selection of good search keys is difficult but crucial for good search results. The structure of queries may be described as *weak* (queries with a single operator, no differentiated relations between search keys) or *strong* (queries with several operators, different relationships between search keys). Strong structured queries are known to perform better. More precisely, strong query structures are based on facets. Each facet indicates one aspect of the information request or task, and is represented by a set of concepts, which, in turn, are expressed by a set of search keys.

The main attribute of query structures lie in the weightings of the search keys. Several operators are used to indicate several query structures. For instance, weak query structures are indicated by the following:

- SUM, which is an unexpanded query, means average of the weights of search keys
- WSUM1, which is an expanded query of SUM, means weighted average of the weights of keys, and
- WSUM2, another expanded query of SUM, means weighted average of the weights of keys.

On the other hand, strong query structures are indicated by the following:

- BOOL, a query with Boolean operator, which is a facet based query structure
- SSYN, a SUM-of-synonym-groups-query; here, each facet formed a clause with the SYN operator
- ASYN, similar query to SSYN queries, but with SYN groups combined with the probabilistic AND operator, ie. a product of facet was calculated instead of an average
- XSUM, a modification of the Boolean query structure: facets were combined with SUM operator instead of AND operator
- OSUM, is a combination of Boolean queries consisting of term keys of different facets and SYN clauses consisting of expansion keys for each facet.

3.4.1 Query Formulation and Expansion

In query formulation, generally three levels - a conceptual, linguistic and string level - can be differentiated. To demonstrate the different levels of query formulation and expansion, let us assume that the following is a query task or request:

Storage of radioactive waste produced in nuclear power plants, examples of risks and accidents.

At the conceptual level, about four facets and seven concepts can be recognized from this request. Therefore a typical query “plan” can be as follows:

nuclear power plants AND radioactive waste AND storage AND (risk OR accident)

At the string level, terms are replaced by search keys. The query is expressed with syntax of the query language (the linguistic level). So sample query is first formulated into the Boolean query structure (depending on the information retrieval model used). An unexpanded query will contain the search concepts selected on the basis of the request and given thus:

#and(#3nuclear power plant) #3(radioactive waste) storage #or(risk accident)

With query expansion, synonyms of the terms are added to the query. The sample query with the synonym expansion is the following:

*#and(#or(#3nuclear power plant) #3(nuclear station) #3(atomic power plant))
#or(#3(radioactive waste) #3(nuclear waste))
#or(storage store)
#or(risk accident danger hazard))*

3.4.2 Weak Query Structures

The weak query structures used to combine the search keys of the above search request can be represented by the *SUM*, *WSUM1* and *WSUM2* queries. An unexpanded *SUM* query can be constructed by the system from the original concept of the request, and each concept is represented by a single key or a set of keys corresponding to the term but without phrases.

The unexpanded query structure is as follows:

SUM

*#sum(nuclear power plant nuclear station atomic power plant atomic reactor nuclear reactor
radioactive waste nuclear waste storage store risk accident danger hazard disaster catastrophe)*

In the query expansion, all expressions will be as single words, ie. no phrases were included. The original keys were usually weighted higher than the expansion keys. Thus in *WSUM* queries expansion keys and the keys of equivalent expressions were given smaller weights than the keys of the original concepts. A typical *WSUM1* query structure will be given as follows:

WSUM1

#wsum(1 2 #3(nuclear power plant) 1 #3(nuclear station) 1 #3(atomic power plant) 1 #3(atomic reactor) 1 #3(nuclear reactor) 2 #3(radioactive waste) 1 #3(nuclear waste) 2 storage 1 store 2 risk 2 accident 1 danger 1 hazard 1 disaster 1 catastrophe)

In WSUM2, query keys will be weighted according to the type of expansion they belonged to. A typical WSUM2 query structure will be given as follows:

WSUM2

#wsum(1 10 #3(nuclear power plant) 9 #3(nuclear station) 9 #3(atomic power plant) 7 #3(atomic reactor) 7 #3(nuclear reactor) 10 #3(radioactive waste) 9 #3(nuclear waste) 10 storage 9 store 10 risk 10 accident 9 danger 9 hazard 7 disaster 7 catastrophe)

3.4.3 Strong Query Structures

Similarly, queries can be constructed from the search request of above using the various operators representing the strong query structures. Using the same query request, the strong query structures will be the following:

SSYN

*#sum(#syn (#3(nuclear power plant) #3(atomic power plant) #3(nuclear station) #3(atomic reactor) #3(nuclear reactor))
#syn(#3(radioactive waste) #3(nuclear waste))
#syn(storage store)
#syn(risk accident danger hazard disaster catastrophe))*

ASYN

*#and(#syn(#3(nuclear power plant) #3(atomic power plant) #3(nuclear station) #3(atomic reactor) #3(nuclear reactor))
#syn(#3(radioactive waste) #3(nuclear waste))
#syn(storage store)
#syn(risk accident danger hazard disaster catastrophe))*

XSUM

*#sum(#or(#3(nuclear power plant) #sum(#3(atomic power plant) #3(nuclear station) #3(atomic reactor) #3(nuclear reactor)))
#or(#3(radioactive waste) #sum(#3(nuclear waste)))
#or(storage #sum(store))
#or(risk accident #sum(danger hazard disaster catastrophe)))*

OSUM

*#sum(#and(#3(nuclear power plant) #3(radioactive waste) storage #or(risk accident))
#syn(#3(atomic power plant) #3(nuclear station) #3(atomic reactor) #3(nuclear reactor))
#syn(#3(nuclear waste))
#syn(store)
#syn(danger hazard disaster catastrophe))*

4.0 Conclusion

In this unit you have studied what a query is. You also learned query formulation and expansion. Finally, you considered the discrete structures of a query.

5.0 Summary

In this unit, a query, which can be broad or narrow, is what a user of a search engine or database enters. A query must be formulated based on search keys given by the user as an essential part of successful information retrieval. In the course of the search, this query can be expanded if the original query does not give satisfactory results.

6.0 Tutor-Marked Assignment

1. Explain what constitutes a query
2. Enumerate on how a query is formulated
3. Itemize the processes that lead to query expansion
4. Discuss briefly structures of queries.

7.0 References/Further Readings

1. o van Risjbergen C.J., (2004). *The Geometry of Information Retrieval*, Cambridge UP.
2. o Voorhees E.M., D.K. Harman (2005). *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press.
3. Jaana Kekalainen and Kalervo Jarvelin, The Impact of Query Structure and Query Expansion on Retrieval Performance at <citeseer.ist.psu.edu/545794.html>
4. Anne Aula, Query Formulation in Web Information Search at <anne.aula.googlepages.com/questionnaire.pdf>

UNIT 4 USER PROFILES

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 User Profiles
 - 3.2 User Profiling Standards
 - 3.2.1 vCard (version 3)
 - 3.2.2 IMS Learner Information Package (LIP)
 - 3.2.3 IEEE Public and Private Information (PAPI) Specification
 - 3.2.4 Global TV-Anytime Specification
 - 3.3 Representation and Maintenance of User Profiles
 - 3.4 Problems with User Profiles
 - 3.5 Personalization of User Profiles
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider what user profile stands, the various standards, representations, problems of user profile. You will also learn the need to evolve a personalized user profile.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- describe in detail what a user profiles is
- enumerate on the existing user profiling standards
- itemize the problems with ordinary user profile
- discuss briefly the need to personalize user profile.

3.0 MAIN CONTENT

3.1 User Profiles

Due to the characteristics of retrieving methods, the conventional Information Retrieval Systems often suffers from inaccurate and incomplete queries as well as inconsistent document relevance. The user profile is thus used to satisfy a user's information needs by retrieving information quickly that he/she needs.

User profile keeps user's data that are used for query expansion and optimally. They are used for a wide variety of applications and have varying levels of descriptivity and complexity. At its most basic level the user profile is a simple form with predefined fields, for example string entries giving values for given fields. However, the categorization of the fields together with the use of the data contained therein varies widely from application to application.

One common and important use for a user profile is found in the field of online searching for information retrieval, where the profile is used to filter results to yield only relevant information to a certain user as search results.

The user profile can be distinguished in two kinds: static and dynamic. An example of a static user profile is his demographic data such as gender, age and profession. A dynamic user profile is his domain knowledge, goals, and preferences.

3.2 User Profiling Standards

Some of the existing user profiling standards are described below.

3.2.1 vCard (version 3)

The vCard specification from the Internet Mail Consortium is a means of Personal Data Interchange (PDI), which automates the traditional business card. It can be used to store vital directory information (name, addresses, telephone numbers, email, URLs), geographic and time zone information, and can include graphics and multimedia (photo, logos, audio clips).

3.2.2 IMS Learner Information Package (LIP)

The IMS Learner Information Package (LIP) specification offers a data model that describes characteristics of a user needed for the general purpose of recording and managing learning related history, goals and accomplishments; engaging the user in a learning experience; discovering learning opportunities for user.

The main elements in LIP are:

- Accessibility: In terms of language, disabilities, and preferences.
- Activity: Any complete learning-related (e.g. self-reported, formal/informal education, training, work experience, and military or civic service).
- Affiliation: Membership of professional organisations.
- Competency: Skills, knowledge, and abilities acquired in the cognitive, affective, and/or psychomotor domains.
- Goal: Learning, career and other objectives and aspirations.
- Identification: Biographic and demographic data relevant to learning.
- Interest: Information describing hobbies and recreational activities.
- Qualifications, Certifications and Licenses.
- Relationship: Relationship between components.
- Security key: The set of passwords and security keys assigned to the learner for transactions with learner information systems and services.
- Transcript: A record that is used to provide an institutionally based summary of academic achievement. The structure of this record can take many forms.

3.2.3 IEEE Public And Private Information (PAPI) Specification

PAPI was created to represent student records and its development is moving towards harmonization with IMS. It specifies data interchange formats, facilitating communication between cooperating systems. User records are divided into personal information and performance information and these are maintained separately. A key feature of the standard is

the logical division, separate security, and separate administration of several types of learner information.

The current specification splits the learner information into the following areas:

- Learner personal information: name, address, and telephone number (private to learner).
- Learner relations' information: learner's relationship to other users of learning technology systems, such as teachers, instructors, and other learners.
- Learner security information: learner's security credentials, such as: passwords, challenge/responses, private keys and public keys. This is private to the learner (with the exception of public information).
- Learner preference information: describes information that may improve human-computer interactions. This type of information is similar to personal information except that it may be public.
- Learner performance information: relates to the learner's history that is created and used by learning technology components to provide optimum learning experiences. Generally, learner performance information is created and used by automated learning technology systems.
- Learner portfolio information: is a collection of a learner's accomplishments and works that is intended for illustration and justification of his/her abilities and achievements.

3.2.4 Global TV-Anytime Specification

The TV-Anytime Forum is an association of organisations that seeks to develop specifications to enable audio-visual and other services based on mass-market high volume digital storage in consumer platforms. The TV-Anytime Metadata specification employs metadata to describe content, user preferences, consumption habits, for targeting a specific audience. In particular, the consumer metadata section of the specification is interesting, as it describes how to define usage history description schema and user preferences description schema.

3.3 Representation and Maintenance of User Profiles

User profiles are a representation of the user's interests. Most Web Assistant build profiles non-invasively by observing which Web pages users visit over a period of time. They generally use the profile to suggest related Web pages to the users as they browse.

Some search engines have developed a three-descriptor representation to monitor user interest dynamics. This model maintains a long-term interest descriptor to capture user's general interests and a short-term interest descriptor to keep track of user's more recent faster changing interests. Some learn user's interests by looking at more than just the pages themselves. They also observe and measure user mouse and scrolling activity in addition to user browsing activity.

The user profile is constructed by observing the information consumption patterns of the user such as the browser web page cache. The profile may also be compiled using manual methods. It can be used to drive information sourcing and to match and filter information obtained from information sources. The matched content is ranked and presented. Implicit and explicit feedback from the consumption behavior is used to update the profile which then drives information sourcing and filtering in future.

The several different ways to specify user profiles include:

- *manual*: users specifies topics of interests (and weights) explicitly
- there is selection of predefined terms or query
- there is the problem of maintenance
- *user feedback*: user collects relevant documents
- the terms in selected document are regarded as important
- the problem here is how to motivate the user to give feedback
- a similar approach to this is used by spam filters
- *heuristics*: observing user behaviour
- example, if a user has opened a document for long time, it is assumed that he/she reads it and therefore it might be relevant
- the problem is that heuristics might be wrong.

3.4 Problems with User Profiles

The problem of user profiling is multi-faceted, and the issues one must address include the choice of information to store and the decision of whether to use existing standards or to create new ones.

Studies have shown that people are unwilling to explicitly specify their interests or give accurate information about themselves. Yet a lot of accurate personal data exists on user desktops that can be used to create richer profiles. As a result many research focus on creating implicit user profiles such that user profiles are created using frequently occurring document words to represent the profile.

User Profiles created in this manner suffer from the following problems:

1. Irrelevant words – Words frequently occur in documents or web pages without being related to the contents of the page.
2. Polysemy and synonymy – A word can have multiple meanings and multiple words can have the same meaning. A word based profile does not have sufficient context to disambiguate the meanings of individual words.
3. Size of the profile – The size of the profiles built using words grow very fast, larger profiles reduce precision.
4. Words in the profile may represent a mixture of information, transactional and recreational needs of the user. For instance, the term camera might appear in the profile because a user read a review for a camera (in a transactional context).

3.5 Personalization of User Profiles

The huge amount of information available on the Internet is widely shared primarily due to ability of Web search engines to find useful information for users. However, many search engines are known to lack a personalization mechanism that would understand the information needs of the user at a particular instance of time and return custom results.

Personalized information retrieval and search promises to improve the Internet experience. An important requirement for building personalized web applications is to build user profiles that

represent the users' interests. The key issue in personalization is the building and maintenance of a user profile that records the user's interests, preferences and other information, and the use of this profile to aid in the personalization of information retrieval.

Personalization broadly involves the process of gathering user-specific information during interaction with the user, which is then used to deliver appropriate content and services; tailor-made to the user's needs. When applied to search, personalization would involve the following steps:

1. Collecting and representing information about the user, to understand the user's interests.
2. Using this information to either filter or re-rank the results returned from the initial retrieval process, or directly including information into search process itself to select personalized results.

One way to design a pragmatic personalized user profile is to personalize web search engines using ontology-based contextual user profiles. In contrast to long-term user profiles, contextual user profiles are constructed that capture what the user is working on at the time they conduct a search. The results of a popular search engine, say Google, is post-processed making use of contextual information and this compared to standard systems.

So rather than building long-term user profiles, the contextual systems try to adapt to the user's current task. They monitor users' tasks/request, anticipates task-based information needs, and proactively provide users with relevant information. The user's tasks are monitored by capturing content from the web browsers and other applications. One application developed in 2003 at Microsoft Research, indexes the content seen by a user and uses the index to provide easier access to information already seen by the user and also to provide rich contextual information for Web searches.

4.0 Conclusion

In this unit you have studied what user profile is. You also learned the various existing standards in user profiling and therefore the need to make for changes.

5.0 Summary

In this unit, user profile is a system used to keep user's data that are used for query expansion and optimally for information retrieval. Many user profiling standards exist, but they seem not to satisfy the relevance need of the user. Therefore, there is the need to personalize user profiles to tailor them to their personal needs

6.0 Tutor-Marked Assignment

1. Describe in detail what a user profiles is
2. Enumerate on the existing user profiling standards
3. Itemize the problems with existing standard user profiles
4. Discuss briefly the need to personalize user profile.

7.0 References/Further Readings

1. Hersch W.R. (2002). *Information Retrieval: A Health and Biomedical Perspective*, Springer..

2. Voorhees E.M., D.K. Harman (2005). *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press.
3. Croft W.B., J. Lafferty. (2003). *Language Modeling for Information Retrieval*. Springer.
4. Jones K. S., P. Willett. (1997). *Readings in Information Retrieval*, Morgan Kaufmann.,
5. Kowalski G., M.T. Maybury. (2005). *Information Storage and Retrieval Systems*, Springer.
6. Contextual Information Retrieval Using Ontology Based User Profiles at
<citeseer.uark.edu/publications/vishnu.pdf>
7. K Ramanathan, et al, (2008) Creating hierarchical user profiles using Wikipedia at
<www.hpl.hp.com/techreports/2008/HPL-2008-127.pdf>
8. B Rousseau et al, User Profiling for Content Personalisation in Information Retrieval at
<<http://www.eprints.wit.ie/653/>>

Module 5 Information Retrieval Systems

Unit 1 Information Retrieval System

Unit 2 Web Information Retrieval

Unit 3 Bibliographic Information Retrieval System and Evaluation

UNIT 1 INFORMATION RETRIEVAL SYSTEM**CONTENTS**

1.0 Introduction

2.0 Objectives

3.0 Main Content

3.1 Information Retrieval System

3.2 Search Engines

3.3 The Architecture of a Search Engine

3.4 Example of an Information Retrieval System: MEDLINE

3.4.1 MEDLINE Database and Retrieval System

4.0 Conclusion

5.0 Summary

6.0 Tutor-Marked Assignment

7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider what information retrieval system is, search engines and a popular example of information system.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- describe in detail what an information retrieval system is
- discuss the technology behind information retrieval system
- describe the architecture of a search engine

3.0 MAIN CONTENT**3.1 Information Retrieval System**

Before an information retrieval system can actually operate to retrieve some information, that information must have already been stored inside the system. This is true both for manual and computerised systems. Originally it will usually have been in the form of documents. The retrieval system is not likely to have stored the complete text of each document in the natural language in which it was written. It will have, instead, a document representative which may have been produced from the documents either manually or automatically.

The starting point of the text analysis process may be the complete document text, an abstract, the title only, or perhaps a list of words only. From it the process must produce a document representative in a form which the system can handle or which the user can understand.

Figure 1 is the diagram of a typical Information Retrieval System. It shows three main components: input, processor and output. From the input side, the need is to obtain a representation of each data and query suitable for a computer to use. Most computer-based retrieval systems store only a representation of the data (or query) which means that the text of a document, for instance, is lost once it has been processed for the purpose of generating its representation.

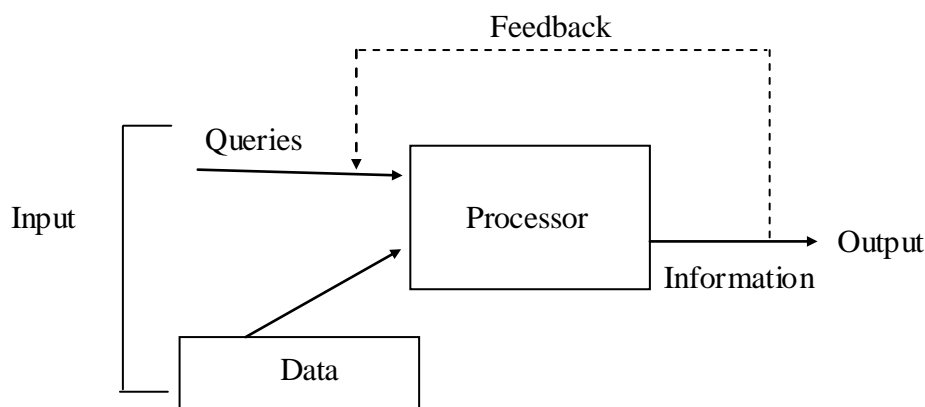


Figure 1: A Typical Information Retrieval System

A *data representative* can be a list of extracted words considered to be significant. Rather than have the computer process the natural language, an alternative approach is to have an artificial language within which all queries and documents can be formulated. Of course it presupposes that a user is willing to be taught to express his information need in the language.

When the retrieval system is on-line, it is possible for the user to change his request (query expansion) during one search session in the light of a sample retrieval, thereby, it is hoped, improving the subsequent retrieval run. Such a procedure is commonly referred to as *feedback*. An example of a sophisticated on-line retrieval system is the MEDLINE system.

The processor is that part of the retrieval system concerned with the retrieval process. The process may involve structuring the information in some appropriate way, such as classifying it. It will also involve performing the actual retrieval function. That is, executing the search strategy in response to a query. In the diagram, the data have been placed in a separate box to emphasize the fact that they are not just input but can be used during the retrieval process in such a way that their structure is more correctly seen as part of the retrieval process.

Finally, the output, which is usually a set of citations or document numbers. In an operational system the story ends here. However, in an experimental system it leaves the evaluation to be done.

3.2 Search Engines

A *search engine* is a tool designed to search for information. When this occurs on the World Wide Web it is then called a web search engine. The search results are usually presented in a list

and are commonly called *hits*. The information may consist of web pages, images, information and other types of files.

Some search engines also mine data available in databases or open directories. Unlike Web directories, which are maintained by human editors, search engines operate algorithmically or are a mixture of algorithmic and human input. Common search engines include: Google , Yahoo, AOL Search, Ask.com, Bing and Looksmart.

3.2.1 The working of a search engine

Web search engines work by storing information about many web pages, which they retrieve from the WWW itself. These pages are retrieved by a Web crawler (sometimes also known as a spider) — an automated Web browser which follows every link it sees. Exclusions can be made by the use of robots.txt. The contents of each page are then analyzed to determine how it should be indexed (for example, words are extracted from the titles, headings, or special fields called meta tags). Data about web pages are stored in an index database for use in later queries.

Some search engines, such as Google, store all or part of the source page (referred to as a cache) as well as information about the web pages, whereas others, such as AltaVista, store every word of every page they find. This cached page always holds the actual search text since it is the one that was actually indexed, so it can be very useful when the content of the current page has been updated and the search terms are no longer in it. This problem might be considered to be a mild form of linkrot, and Google's handling of it increases usability by satisfying user expectations that the search terms will be on the returned webpage. This satisfies the principle of least astonishment since the user normally expects the search terms to be on the returned pages. Increased search relevance makes these cached pages very useful, even beyond the fact that they may contain data that may no longer be available elsewhere.

When a user enters a query into a search engine (typically by using key words), the engine examines its index and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text. Most search engines support the use of the boolean operators AND, OR and NOT to further specify the search query. Some search engines provide an advanced feature called proximity search which allows users to define the distance between keywords.

The usefulness of a search engine depends on the relevance of the *result set* it gives back. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another. The methods also change over time as Internet usage changes and new techniques evolve.

Most web search engines are commercial ventures supported by advertising revenue and, as a result, some employ the practice of allowing advertisers to pay money to have their listings

ranked higher in search results. Those search engines which do not accept money for their search engine results make money by running search related ads alongside the regular search engine results. The search engines make money every time someone clicks on one of these ads.

3.3 The Architecture of a Search Engine

A search engine contains three components: an indexer, a crawler, a query server. The crawler collects pages from the Web. The indexer processes the retrieved documents and represents them in an efficient search data structure. The query server accepts the query from the user and returns the result pages by consulting with the search data structures.

Figure 1 shows the system architecture of Google. Most of Google is implemented in C or C++ language for efficiency and can run on either Solaris or Linux. Here, the web crawling (downloading of web pages) is done by several distributed crawlers. There is a URL server that sends lists of URLs to be fetched to the crawlers. The web pages that are fetched are then sent to the store server. The store server then compresses and stores the web pages into a repository.

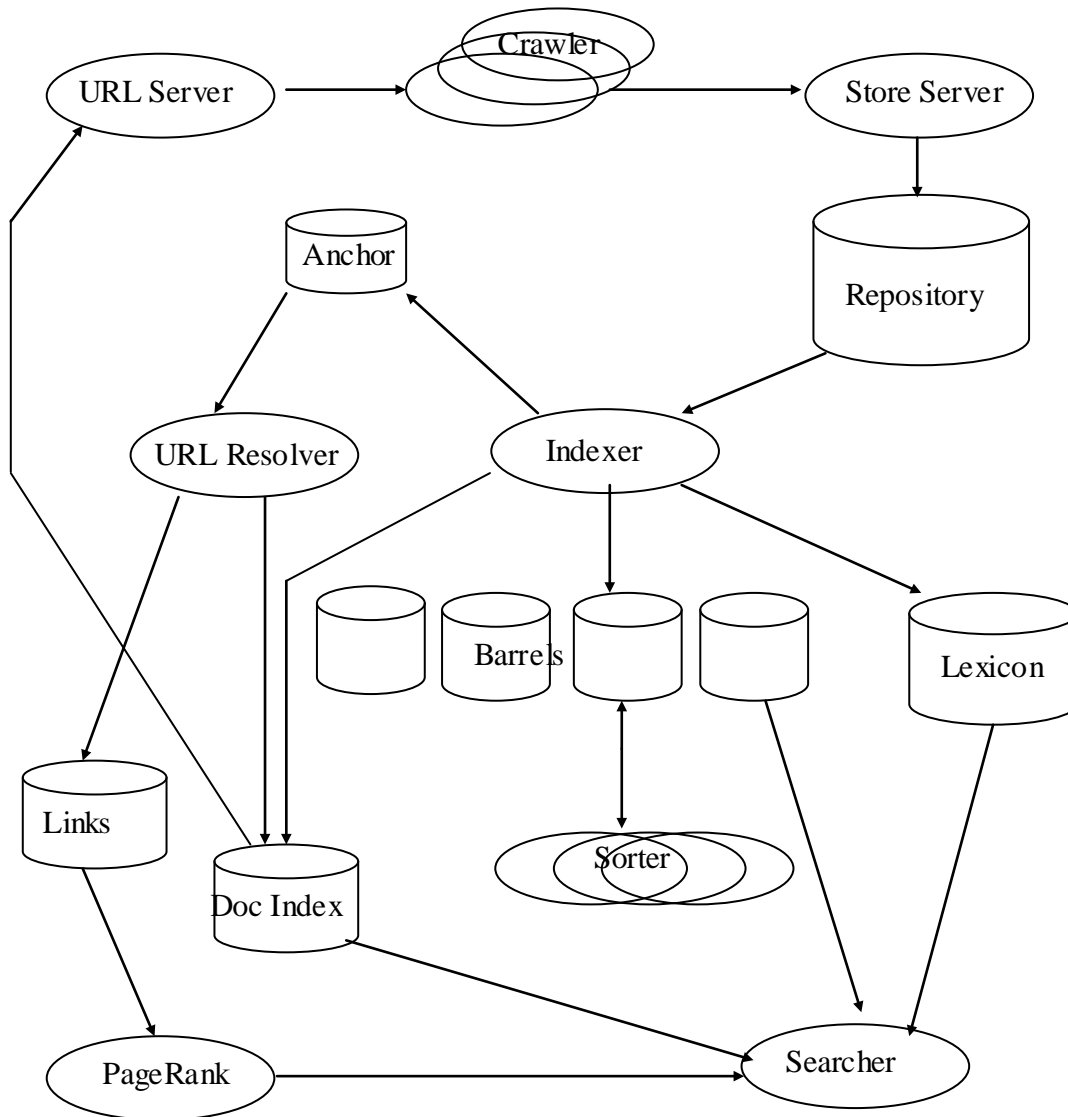


Figure 1: High Level Google Architecture

Every web page has an associated ID number called a doc ID which is assigned whenever a new URL is parsed out of a web page. The indexing function is performed by the indexer and the sorter. The indexer performs a number of functions. It reads the repository, uncompresses (a kind of decompression) the documents, and parses them.

Each document is converted into a set of word occurrences called hits. The hits record the word, its position in document, an approximation of font size and capitalization. The indexer distributes these hits into a set of "barrels", creating a partially sorted forward index. The indexer performs another important function. It parses out all the links in every web page and stores important information about them in an anchors file. This file contains enough information to determine where each link points from and to, and the text of the link.

The URL resolver reads the anchors file and converts relative URLs into absolute URLs and in turn into docIDs. It puts the anchor text into the forward index, associated with the docID that the anchor points to. It also generates a database of links which are pairs of docIDs. The links database is used to compute PageRanks for all the documents.

The sorter takes the barrels, which are sorted by docID and resorts them by wordID to generate the inverted index. This is done in place so that little temporary space is needed for this operation. The sorter also produces a list of wordIDs and offsets into the inverted index. A program called DumpLexicon takes this list together with the lexicon produced by the indexer and generates a new lexicon to be used by the searcher. The searcher is run by a web server and uses the lexicon built by DumpLexicon together with the inverted index and the PageRanks to answer queries.

3.4 A Typical Information Retrieval System: MEDLINE

MEDLINE (Medical Literature Analysis and Retrieval System Online) is a bibliographic database and information retrieval system of life sciences and biomedical information. It includes bibliographic information on articles from academic journals covering biology, biochemistry, molecular evolution, medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care.

MEDLINE is compiled by the United States National Library of Medicine (NLM) and is freely available on the Internet and searchable via PubMed and NLM's National Center for Biotechnology Information's Entrez system. This free nature is the secret of its popularity.

3.4.1 MEDLINE Database and Retrieval System

MEDLINE database contains over 18 million records from about 5,000 publications covering biology, medicine and health from 1950 to date. The database is freely accessible on the Internet via the PubMed interface and new citations are added daily except Sunday and Monday. About 48% of the database are for cited articles published in the U.S., about 88% are published in English, and about 76% have English abstracts written by authors of the articles.

The system uses Medical Subject Headings (MeSH) for information retrieval. Engines designed to search MEDLINE (such as Entrez and PubMed) generally use a Boolean expression combining MeSH terms, words in abstract and title of the article, author names, date of publication, etc. Both Entrez and PubMed allow also to find articles similar to a given one based on a mathematical scoring system that takes into account the similarity of word content of the abstracts and titles of two articles.

There are tutorials for instruction on the PubMed interface to MEDLINE. Unlike using a typical internet search engine, PubMed searching of MEDLINE requires a little investment of time. Using the MeSH database to define the subject of interest is one of the most useful ways to improve the quality of a search. Using MeSH terms in conjunction with limits (such as publication date or publication type), qualifiers (such as adverse effects or prevention and control), and text-word searching is another.

Finding one article on the subject and clicking on the "Related Articles" link to get a collection of similarly classified articles can expand a search that yields few results. In addition to the National Library of Medicine's tutorials, there are several other guides to effective searching, such as pages from a book on MEDLINE usage.

4.0 Conclusion

In this unit you have studied what information retrieval system is. You also learned search engines the technology behind information retrieval system as well as MEDLINE, a common example of information retrieval system.

5.0 Summary

In this unit, an information retrieval system can operate to retrieve some information only when that information have already been stored inside the system. The retrieval system will not store the complete text of each document, but a document representative which may have been produced from the documents either manually or automatically.

6.0 Tutor-Marked Assignment

1. Describe in detail what an information retrieval system is
2. Discuss briefly the technology behind information retrieval system
3. Explain briefly the working of search engines.

7.0 References/Further Readings

1. Manning C.D., P. Raghavan, H. Schütze. (2008). *Introduction to Information Retrieval*, Cambridge UP.
2. Baeza-Yates R., B. Ribeiro-Neto.(1999). *Modern Information Retrieval*, Addison-Wesley.
3. Grossman D.A., O. Frieder. (2004). *Information Retrieval: Algorithms and Heuristics*, Springer.
4. van Rijsbergen C. J. (1979), INFORMATION RETRIEVAL: Chapter 1 at <<http://www.dcs.gla.ac.uk/Keith/pdf/Chapter1.pdf>>
5. Bruce R. Schatz, et al., (1997) Information Retrieval in Digital Libraries: Bringing Search to the Net, *Science* 275, 327 at <www.sciencemag.org/cgi/content/full/275/5298/327>

UNIT 2 WEB INFORMATION RETRIEVAL

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 The Web
 - 3.2 Web Information Retrieval and Systems
 - 3.2.1 Web Information Retrieval Tools
 - 3.3 Uniqueness of Web Information Retrieval
 - 3.4 Classic Information Retrieval vs Web Information Retrieval
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will consider what the web is. Then you will also study web information retrieval and systems, comparing it with classic information retrieval system.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- describe in detail what the web is
- discuss web information retrieval and its system
- compare and contrast web information retrieval and classic information retrieval

3.0 MAIN CONTENT

3.1 The Web

The *World Wide Web*, known as *the Web* for short, is a system of interlinked hypertext documents accessed via the Internet. With a software called the web browser (such as internet explorer), one can view Web pages that may contain text, images, videos, and other multimedia and navigate between them using hyperlinks.

Sometimes, the terms Internet and World Wide Web are used in every-day speech without much distinction. However, the Internet and the World Wide Web are not one and the same. The Internet is a global system of interconnected computer networks. In contrast, the Web is one of the services that runs on the Internet. It is a collection of interconnected documents and other resources, linked by hyperlinks and URLs.

Viewing a webpage on the World Wide Web normally begins either by typing the URL of the page into a web browser, or by following a hyperlink to that page or resource. The web browser then initiates a series of communication messages, behind the scenes, in order to fetch and display it.

Initially, the server-name portion of the URL is resolved into an IP address using the global, distributed Internet database known as the domain name system, or DNS. This IP address is necessary to contact the Web server. The browser then requests the resource by sending an HTTP request to the Web server at that particular address. In the case of a typical Web page, the HTML text of the page is requested first and parsed immediately by the web browser, which then makes additional requests for images and any other files that form parts of the page.

While receiving these files from the Web server, browsers may progressively render the page onto the screen as specified by its HTML, CSS, and other Web languages. Any images and other resources are incorporated to produce the on-screen web page that the user sees. Most Web pages will themselves contain hyperlinks to other related pages and perhaps to downloads, source documents, definitions and other web resources. Such a collection of useful, related resources, interconnected via hypertext links, is called a "web" of information. Making it available on the internet created what is called the World Wide Web.

The Web is becoming a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before. Its success is based on the conception of a standard user interface which is always the same no matter what computational environment is used to run the interface. As a result, the user is shielded from details of communication protocols, machine location, and operating systems.

Further, any user can create his own Web documents and make them point to any other Web documents without restrictions. This is a key aspect because it turns the Web into a new publishing medium accessible to everybody. As an immediate consequence, any Web user can push his personal agenda with little effort and almost at no cost.

3.2 Web Information Retrieval and Systems

Retrieving information from the web is becoming a common practice for internet users. The huge size and heterogeneity of the web is no longer in doubt. Therefore, the web poses a dire challenge to the effectiveness of classical information retrieval systems. A critical goal of successful information retrieval on the web, though, is to identify which pages are of high quality and relevance to a user's query.

The success of the web lies in the many software tools that are available for its information retrieval. These software include the search engines (Google, AltaVista etc), hierarchical directories (Yahoo), many other software agents and collaborative filtering systems.

3.2.1 Web Information Retrieval Tools

The main tools used by the web in its information retrieval include the following:

1. General-purpose search engines: can be Direct (examples, AltaVista, Excite, Google, Infoseek and Lycos,) or Indirect or Meta-search (examples, MetaCrawler, DogPile, AskJeeves, and InvisibleWeb)

2. Hierarchical directories: including Yahoo and all portals
3. Specialized search engines: these deals with heterogeneous data sources include the Home page finder such as Ahoy, the Shopping robots such as Jango and Junglee, whose database is mostly built by hand, and Applet finders.
4. Search-by-example: Examples are Alexa's "What's related", Excite's "More like this", Google's "Googlescout", etc.
5. Collaborative filtering: Examples include Firefly and GAB
6. Meta-information: These are Search Engine Comparisons and are used for Query log statistics.

Hierarchical directories can be manual, that is, the database is mostly built by hand or automatic. Examples of manual hierarchical directories are Yahoo!, LookSmart and Open Directory. Automatic hierarchical directories are now populating hierarchy. Here, for each node in the hierarchy, you formulate fine-tuned query and run modified HITS algorithm. Another feature of automatic hierarchical directories is Categorization: For each document the "best" placement is found in the hierarchy. The techniques used in automatic hierarchical directories are connectivity and/or text based.

3.3 Uniqueness of Web Information Retrieval

With the fast growth of the Internet, more and more information is available on the web and as a result, web information retrieval has become a fact of life for most Internet users. The uniqueness of web information retrieval is listed as the following:

- Bulk - The bulk size of the Internet is over 400 million documents as measured in the year 2000, which is growing at the speed of 20M per month.
- Dynamic Internet - The Internet is changing everyday while most classic information retrieval systems are designed for mostly static text databases.
- Heterogeneity - The Internet contains a wide variety of document types: pictures, audio files, text and scripts etc.
- Variety of Languages - The types of languages used in the Internet is more than 100.
- Duplication - Copying is another important characteristic of the web, as it is estimated that about 30% of the web pages are duplicates.
- High Linkage - Each document averagely has more than 8 links to other pages.
- Ill-formed queries - web information retrieval systems are required to service short and not particularly well represented queries from the Internet users.
- Wide Variance in Users - Each web user varies widely in their needs, expectations and knowledge.
- Specific Behavior - It is estimated that nearly 85% users only look at the first screen of the returned results from search engines. 78% users never modify their very first query.

3.4 Classic Information Retrieval vs Web Information Retrieval

Classic information retrieval constitutes all previous information retrieval techniques before and other than the web information retrieval. The input of classic information retrieval is mainly for document collection and the goal is to retrieve document or text with information content that is relevant to user's information need. Classic information retrieval involves two main aspects:

- processing the document collection and
- processing queries (searching).

To determine the query results (ie. which documents to return), information retrieval models like the Boolean and Vector models are used.

On the other hand, the input of web information retrieval is the publicly accessible web while the goal is to retrieve high quality pages that are relevant to user's information need. Classic information retrieval involves two main aspects: processing the document collection and processing queries (searching).

Web information retrieval can be static, in which files like text, audio and videos are retrieved, or dynamic, which is mainly database access generated on request. Two aspects of the web information retrieval are processing and representation of the document collection and processing queries. Processing and representation of document collection involves either gathering the static pages or learning about the dynamic pages.

Web information retrieval has the following advantages over classic information retrieval:

1. User

- Many tools are available to the user
- Personalization of information result given a query is better
- Interactivity: for instance the query can be refined or expanded as desired

2. Collection/System

- Hyperlinks are available to link one document to the other
- Statistics is easy to gather even in large sample sizes
- Interactivity: the system makes the users explain what they want

4.0 Conclusion

In this unit you have studied what web information retrieval system is. You also learned the similarities and differences between classic and web information retrieval.

5.0 Summary

In this unit, the fast growth of the Internet has made more and more information to be available on the web. Thus, web information retrieval has become a fact of life for most Internet users. The present success of the web lies in the many software tools available for its information retrieval.

6.0 Tutor-Marked Assignment

1. Describe in detail what the web is
2. Discuss web information retrieval and its system
3. Compare and contrast web information retrieval and classic information retrieval

7.0 References/Further Readings

1. Jones K. S., P. Willett. (1997). *Readings in Information Retrieval*, Morgan Kaufmann.,
2. Chowdhury G.G., (2003). *Introduction to Modern Information Retrieval*, Neal-Schuman
3. Sahami M et al (2004), The Happy Searcher: Challenges in Web Information Retrieval at <robotics.stanford.edu/users/sahami/papers-dir/PRICAI-2004.pdf>
4. Baeza-Yates R et al (1999), *Modern Information Retrieval*, ACM Press, New York, USA.
5. Huang L, (2000), A Survey *On Web* Information Retrieval Technologies at <www.nlp.org.cn/docs/docredirect.php?doc_id=78>

UNIT 3 BIBLIOGRAPHIC INFORMATION RETRIEVAL SYSTEM AND EVALUATION

CONTENTS

- 1.0 Introduction
- 2.0 Objectives
- 3.0 Main Content
 - 3.1 Indexing Applications for Information Retrieval
 - 3.1.1 Document Linearization
 - 3.1.2 Filtration
 - 3.1.3 Stemming
 - 3.1.4 Weighting
 - 3.2 Bibliographic Information Retrieval System
 - 3.3 Evaluation of Information Retrieval
 - 3.3.1 Relevance as Factor in Evaluation of Information Retrieval
- 4.0 Conclusion
- 5.0 Summary
- 6.0 Tutor-Marked Assignment
- 7.0 References/Further Readings

1.0 INTRODUCTION

In this unit you will study bibliographic information retrieval system and evaluation. Before then, you will consider indexing and its applications in information retrieval, which constitute the foundations of bibliographic information retrieval.

2.0 OBJECTIVES

After going through this unit, you should be able to:

- discuss indexing applications for information retrieval
- describe bibliographic information retrieval system
- evaluate information retrieval as a whole.

3.0 MAIN CONTENT

3.1 Indexing Applications for Information Retrieval

Documents used for retrieval in the information retrieval systems are first indexed to make it retrievable. Therefore *indexing* is a process of preparing the raw document collection into an easily accessible representation of documents. Transforming a document into an indexed form involves the use of:

- a set of regular expressions
- parsers
- a set of stop words (a stop list)
- other miscellaneous filters

Indexing is normally done in five steps:

1. markup & format removal
2. tokenization
3. filtration
4. stemming and
5. weighting.

Markup & format removal with tokenization constitutes document linearization. If no markup removal and weighting is required, then this transformation involves only tokenization, filtration and stemming. This type of indexing is frequently found in bibliographic databases that merely sort text files and raw data. On the Web, however, the above five steps are used especially since documents are created in different formats and relevance scores are needed.

3.1.1 Document Linearization

Document Linearization is the process by which a document is reduced to a stream of terms. This is usually done in two steps and as follows.

1. *Markup and Format Removal* - During this phase, all markup tags and special formatting are removed from the document. Thus, for an HTML document all tags and text inside these are removed. This normally would include all element attributes, scripts, comment lines and text placed into these.
2. *Tokenization* - During this phase, all remaining text is parsed, lowercased and all punctuation removed. Hyphenation rules must be invoked. For instance, some systems may elect to retain hyphens while others may be designed to either ignore hyphens or interpret these as spaces or as join tokens.

During linearization all Cascading Style Sheets (CSS) instructions are removed. This means that without a clear understanding of text linearization processes, the arbitrary repositioning of content using CSS, nested division tags or tables can indeed be detrimental in the sense that what users perceive as relevant content is not what the search engine is actually "reading" and scoring as relevant. In fact, after document linearization, three things are certain:

1. the effective flow of the text should describe a coherent stream of terms
2. this stream of text must capture the intended semantics, theme, topics, subtopics, etc of the document.
3. the position of terms in the text stream is determined by how the markup lines (e.g., HTML tags) were declared in the source code.

All this underscores the fact that users's perception of relevancy (e.g., the information and its meaning displayed before users) and machine perception of relevancy (relevance scores) are two different things. More than often the way search engines perceive and interpret content and the way users and browsers read that content does not match. This is why it is so important to conduct document linearization -as part of a GAP analysis- before and after optimizing a document.

3.1.2 Filtration

Filtration refers to the process of deciding which terms should be used to represent the documents so that these can be used for:

- describing the document's content
- discriminating the document from the other documents in the collection.

Frequently used terms cannot be used for this purpose for two reasons. First, the number of documents that are relevant to a query is likely to be a small proportion of the collection. The second reason is that terms appearing in many contexts do not define a topic or sub-topic of a document.

For these reasons, frequently used terms or stopwords are removed from text streams. However, removing stopwords from one document at a time is time consuming. A cost-effective approach consists in removing all terms which appear commonly in the document collection, and which will not improve retrieval of relevant material.

3.1.3 Stemming

Stemming refers to the process of reducing terms to their stems or root variant. For instance, "computer", "computing", "compute" can be reduced to "comput" and "walks", "walking" and "walker" reduced to "walk". Not all systems use the same type of stemmer.

Over the years, many have considered the advantages and disadvantages of using stemming. For instance, there is no doubt that stemming insures that documents containing variations of a given queried term are considered in the final answer set. Stemming also reduces the size of an inverted file. However, too much stemming is not practical and can annoy users.

3.1.4 Weighting

Weighting is the final stage in most Information Retrieval Indexing applications. Terms are weighted according to a given weighting model which may include local weighting, global weighting or both. If local weights are used, then term weights are normally expressed as term frequencies, *tf*. If global weights are used, the weight of a term is given by *IDF* values. The most common (and basic) weighting scheme is one in which local and global weights are used (*weight of a term = tf*IDF*). This is commonly referred to as *tf*IDF* weighting.

3.2 Bibliographic Information Retrieval System

The basic technology for searching bibliographic databases is the primary method for large-scale information retrieval. When there is a large collection (say a million documents) to be searched, the retrieval methods used today are still those developed for bibliographic databases 30 years ago. These text retrieval methods rely on indexing the documents so that selected items can be quickly retrieved. A user sends a query from his terminal across the network to a "server." Software at the server searches the index, locates the documents matching the query, and returns these documents to the user terminal or the output.

To support retrieval by word matching (example, find all documents containing the word "fiber"), an inverted-list index is built. The documents are scanned for words, omitting a few noise words (such as "the" and "of"), and a list is built for every word. These lists are called

"inverted" because for each word they contain pointers to the documents that contain that word. The index consists of the inverted lists in alphabetical order by word. It can be used for fast search of a specified word by scanning the index for that word and then using the attached document pointers to retrieve the matching documents.

This word-matching search often uses word stemming to increase its retrieval effectiveness. Words are shrunk to a canonical form, so that, for example, "comput" represents "computer," "computers," and "computing." If multiple words are specified, the resulting sets of documents can be merged (logical AND results in an intersection; OR yields the union).

As it became technologically and economically feasible to provide faster networks and larger disks, it became possible to store and retrieve more than just a citation, so that the scale of documents for information retrieval became greater. First, the abstract was added, and this is the economic level that today remains the standard for scientific literature. Then, video terminals became the mode of display, so that text could be viewed more rapidly than with teletypewriters. This led to the extension from abstracts to "full text." An online full-text article contains all words within an article but may exclude non-textual materials such as figures, tables, and equations.

The searching technology also increased in scope while staying fundamentally the same in function. And because there was now a full article instead of an abstract (20 versus 2 kilobytes), there were more words per document. Individual words thus became less discriminating in searches, and phrases became more useful. Internally, this change in focus implied that Boolean operators became less useful (for example, finding "fiber" and "optics" anywhere in the same document often happens coincidentally), whereas proximity operators became more useful (for example, "fiber" within two words of "optics" finds such intended phrases as "fiber network optics").

To implement these new proximity operators, additional information was needed in the index. An inverted-list index contains all of the words along with pointers to all documents containing each word. To compute proximity, the word position within the document is also specified. When proximity search is desired, the modified word lists can be intersected as they were for Boolean ANDs, followed by comparison of the word positions within the same document.

Full-text retrieval was driven by demands in the professions, particularly law. Bibliographic retrieval was pioneered in medicine, where it might be argued that abstracts were satisfactory for identifying the content of an article. This was less so in law, and the Lexis system of U.S.A. court records provided in 1973 the first large-scale commercial system demonstrating the practicality of full-text documents. Today, full text is common for the majority of popular materials.

3.3 Evaluation of Information Retrieval

In evaluating an information retrieval system the main concern is with providing data so that users can make a decision as to:

1. whether they want such a system and
2. whether it will be worth it.

Furthermore, these methods of evaluation are used in a comparative way to measure whether certain changes will lead to an improvement in performance. In other words, when a claim is made for say a particular search strategy, the yardstick of evaluation can be applied to determine whether the claim is a valid one.

The next question to consider is what can be measured that will reflect the ability of the system to satisfy the user. There are six main measurable quantities when evaluating an information retrieval system:

1. The *coverage* of the collection, that is, the extent to which the system includes relevant matter
2. The *time lag*, that is, the average interval between the time the search request is made and the time an answer is given
3. The form of *presentation* of the output
4. The *effort* involved on the part of the user in obtaining answers to his search requests
5. The *recall* of the system, that is, the proportion of relevant material actually retrieved in answer to a search request
6. The *precision* of the system, that is, the proportion of retrieved material that is actually relevant.

Quantities (1)-(4) are readily assessed. Therefore, recall and precision, because they are not readily assessed are known as the *effectiveness* of the retrieval system. Therefore, effectiveness is used to mean a measure of the ability of the system to retrieve relevant documents while at the same time holding back non-relevant one. It is assumed that the more effective the system the more it will satisfy the user.

In classic information retrieval, the performance of an Information Retrieval system is evaluated by assessing recall and precision. In web Information Retrieval, the quality of pages varies widely. Thus just being relevant is not enough. The goal is to return both high-relevance and high-quality (in other word, valuable) pages.

The final question is what technique should be used for evaluation. It is important to note that the technique of measuring retrieval effectiveness has been largely influenced by the particular retrieval strategy adopted and the form of its output. For example, when the output is a ranking of documents, an obvious parameter such as rank position is immediately available for control.

Using the rank position as cut-off, a series of precision and recall values could then be calculated, one part for each cut-off value. The results could then be summarized in the form of a set of points joined by a smooth curve. The path along the curve would then have the immediate interpretation of varying effectiveness with the cut-off value.

3.3.1 Relevance as Factor in Evaluation of Information Retrieval

Different users may differ about the relevance or non-relevance of particular documents to given questions. Therefore, *relevance* is a subjective notion. Several experiments and researches have been done to assess relevance. And it is a general assumption in the field of Information Retrieval that should a retrieval strategy fare well under a large number of *experimental*

conditions then it is likely to perform well in an *operational* situation where relevance is *not* known in advance.

A document is relevant to an information need if and only if it contains at least one sentence which is relevant to that need. This is the true evaluation of the effectiveness of a document, since effectiveness is purely a measure of the ability of the system to satisfy the user in terms of the relevance of documents retrieved.

4.0 Conclusion

In this unit you have studied bibliographic information retrieval system and evaluation. You also learned the process of indexing with its applications as a foundation for bibliographic information retrieval.

5.0 Summary

In this unit, indexing is the process of preparing the raw document collection into an easily accessible representation of documents, which is readily retrievable. The bibliographic information retrieval system relies on the process of indexing the documents so that selected items can be quickly retrieved. In order to assess the effectiveness of retrieval systems, an evaluation of the information retrieval system is done. In evaluating an information retrieval system the main concern is with providing data so that users can make a decision as to whether they want such a system and whether it will be worth it.

6.0 Tutor-Marked Assignment

1. Discuss indexing applications for information retrieval
2. Describe bibliographic information retrieval system
3. Evaluate on information retrieval as a whole.

7.0 References/Further Readings

1. Chowdhury G.G., (2003). *Introduction to Modern Information Retrieval*, Neal-Schuman
2. Meadow C.T., B.R. Boyce, D.H. Kraft, C.L. Barry. (2007). *Text Information Retrieval Systems*. Academic Press.
3. van Rijsbergen C. J. (1979), INFORMATION RETRIEVAL: Chapter 7 at <<http://www.dcs.gla.ac.uk/Keith/pdf/Chapter1.pdf>>
4. DRAFT: 8 - Evaluation in Information Retrieval (2009), Cambridge University Press at <<http://www.dcs.gla.ac.uk/Keith/pdf/Chapter7.pdf>>