COURSE GUIDE

LIS 303

INFORMATION RETRIEVAL (CATALOGUING II)

Course Team: Ilo, Promise I. (PhD) (Course Writer/Course Developer) University of Nigeria, Nsukka & Nkiko, Christopher (PhD) (Course Writer/Course Developer) -Elizade University Ilara-Mokin
Prof. R. U. Ozioko (Course Editor)-Michael Okpara University of Agriculture, Umudike

NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters
University Village
Plot 91, Cadastral Zone
NnamdiAzikiwe Expressway
Jabi, Abuja

Lagos Office
14/16 Ahmadu Bello Way
Victoria Island, Lagos

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

Printed 2022

ISBN: 978-978-058-233-3

CONTENTS **PAGE**

INTRODUCTION

Library collections consist of print holdings, databases, electronic journals and books, digital curated resources, documents, multimedia objects, open educational resources, video documentaries, institutional repositories and Internet resources which are usually processed and stored to meet the teaching, learning and research needs of diverse library patrons. The avalanche of information resources in the library will be mere decorations if not properly accessed and utilised. The whole essence of classification and cataloguing, the hallmark of librarianship, will be tantamount to exercise in futility and wasted efforts if library patrons cannot ferret out specific and desired materials from the maze of collections with ease and within the shortest time – frame or response rate. It is against this background that LIS 303, information retrieval is conceptualised as critical for providing appropriate skills and knowledge- base to guarantee effective and efficient library services delivery. The course explored an overview and definition of the concept of information retrieval, importance of information retrieval, information representation and retrieval, OBJECTIVES of information retrieval, approaches to information representation, retrieval techniques and query representation; information retrieval models; information retrieval systems; problems and challenges of information representation and retrieval in libraries and information centres in Nigeria.

COURSE LEARNING OUTCOME

The course is designed to inculcate and imbue in the students the ability to:

i. Define information retrieval,
ii. Appreciate the significance and necessity for information retrieval,
iii. Understand the different approaches to information representation,
iv. Be conversant with traditional and modern retrieval techniques and query representation,
v. Be aware of the basic components of information retrieval system,
vi. Be aware of the basic characteristics of the different types of information retrieval systems,
vii. Be able to differentiate between types of indexing and abstracting languages,
viii. Discuss different information retrieval models,
ix. Identify the different information retrieval techniques;
x. Differentiate between the different information retrieval techniques,

xi. Explain the concept of precision and recall,
xii. Highlight the functional process of information retrieval,
xiii. Identify the major information retrieval models;
xiv. Understand Boolean logic and relevance ranking,
xv. Differentiate between vector space model and Boolean model,
xvi. Explain Probabilistic Model of information retrieval,
xvii. Identify the different search engines in the World Wide Web;
xviii. Understand the indexing of documents in the World Wide Web,
xix. Explain the different web search services,
xx. Discuss the taxonomy of web search tools,
xxi. Explain the different categories of users of information retrieval system,
xxii. Distinguish between the Digital natives and immigrate users,
xxiii. Identify professional users, information specialist users, novice users,
xxiv. Discuss challenges of information retrieval in Nigerian context,
xxv. Identify general search problems encountered by users,
xxvi. Proffer solution to improving the information retrieval system.

STUDY UNITS

There are 13 study units in this course. They are divided into five modules. The modules and units are presented as follows:

Module 1 **Concept of Information Retrieval, objectives and Importance of information retrieval**

Unit 1 Definitions, Tools and Generation for Information Retrieval
Unit 2 Objectives and Functions of Information Retrieval

Module 2 **Information Representation and Retrieval**

Unit 1 Approaches to Information Representation
Unit 2 Language in information Retrieval

Module 3 **Information Retrieval Techniques and Models**

Unit 1 Information Retrieval Techniques
Unit 2 Information Retrieval Models

Module 4 **Multimedia Information Retrieval Systems and Information Retrieval on the World Wide Web**

Unit 1 Multimedia Information Retrieval Systems

Unit 2 Information Retrieval on the World Wide Web
Unit 3: Digital Libraries and Retrieval Imperatives
Unit 4: Evaluation of Information Retrieval Techniques and Processes

Module 5 Information Retrieval in Nigerian Libraries and Information Centres

Unit 1 Users and Information Retrieval
Unit 2 Evaluation of Information Sources
Unit 3 Problems and challenges of information representation and retrieval in libraries and information centers in Nigeria

WORKING THROUGH THIS COURSE

To achieve effective mastery and excellence in this course, students are implored to diligently read through the materials for comprehension. Students are to avoid rote learning or memorisation and regurgitation. There is need to develop operational knowledge of the course as it exposes students to the kernels of information retrieval which is a requisite skill for professional practice as a library and information expert. Each unit has embedded OBJECTIVES to guide students understanding of the thematic issues. They represent the intended learning outcomes. Students should endeavor to personally answer the tutor-marked assignments to gauge the extent to which they have internalised the discourse. It is also advisable to have a note-book for personal jottings during the study. Class attendance and participation remain a non-negotiable requirement for success in this course. Continuous assessment which constitutes 30% of the total score for the course must be taken seriously.

HOW TO GET THE MOST FROM THIS COURSE

Access to internet connectivity is crucial to enable students participate in the online real time facilitation. Students should endeavor to revisit the recorded lectures at their pace. Submit all assignments on good time and seek for feedback and classifications on unclear or ambiguous issues. Adhere strictly to official regulations at all times.

MAIN COURSE

CONTENTS PAGE

Module 1 Concept of Information Retrieval, Objectives and Importance

Unit 1: Definitions, Tools, and Stages of Information Retrieval

Unit Structure

1.1 Introduction

This unit will introduce students to an overview of the concept of information retrieval, various definitions of information retrieval with particular emphasis on Library and Information Science perspective as well as establish consensus among the divergent perspectives.

1.2 Intended Learning Outcomes

By the end of this unit, you will be able to:

- define the concept Information Retrieval
- identify the basic characteristics of Information Retrieval
- discuss the Information Retrieval Tools
- identify the different information retrieval tools
- define and discuss indexes, catalogues, abstracts and bibliographies as information retrieval tools
- discuss the different types of bibliographies
- differentiate between catalogues and indexes.

1.3 Overview of Definitions and Concept of Information Retrieval

An Overview of Information Retrieval - Storage and retrieval of information have always been major issues of concern among library and information science professionals over the years. This stems from the necessity to keep pace with information explosion and satisfy the information needs of library patrons. The idea of retrieval is predicated on the premise that library is a growing organism characterised by exponential growth in information resources of diverse formats requiring a mechanism that ensures corresponding rate of accessibility and searchability to forestall underutilisation and user frustration.

In the views of Vickery & Vickery (1992) the information required by users or enquirers may be factual and conceptual –the value of physical property, the details of a technical method, the description of a device, an equation for a relation between variables, the ideas behind a physical theory, etc. In contrast to this is the information stored in a retrieval system in the form of messages, physical records bearing graphic markings (i.e., number, text, drawing, etc.,) which carry meaningful content that the recipient can interpret.

Records in retrieval systems can be of several kinds. For example:

- Quantitative and qualitative data about variables of interest;
- Texts (including illustrations) on every kind of subject;
- Drawing, graphs, charts, maps, and other graphic materials;
- Computer programmes;
- Descriptions of objects (for example, of minerals, laboratory apparatus, industrial equipment, etc.)
- Names and locations – of people, institutions, manufacturers;
- Bibliographic references – (i.e. indicators of the identity and location of texts where any of the above types of information may be found).

The total process of information retrieval is often multi-stage. Example, a search for some quantitative data on the properties of a manufactured material might require a series of steps namely;

i. Search bibliography for reference to texts about the material;
ii. Locate the texts and find the one that gives the name of a manufacturer and another that mentions a computer databank that could include data on the material;
iii. Search directories to locate the manufacturer and the databank;
iv. contact manufacturer and receive a brochure containing relevant data;
v. Access databank and retrieve further data.

1.3.1 Definition of Information Retrieval (IR)

The effectiveness of a library is determined by the speed with which its users can locate and retrieve accurate and relevant information for their multifaceted needs. Chimah, Unagha and Nwokocha (2010) defined information retrieval as "a process that involves extraction of information from a collection or database in response to an information problem".

Edom (2012) averred that information retrieval is "a mechanism or apparatus that aids library users to locate, retrieve and utilise needed documents, information or books from the library collection". Rashid (2020) described information retrieval as "the activity of obtaining the right information, to the right user at the right time and that it deals mainly with the representation, storage, organisation and access to information".

It is also the activity of obtaining information resources relevant for a user's information need from a collection of information resources (Manning, Raghavan and Schutize, 2009). Chowdhury (2017) opined that information retrieval is "concerned with all activities related to the organisation of, processing of, and access to, information of all forms and formats. That it allows people to communicate with an information system or service in order to find information – text, graphic images and sound recordings or video that meets their specific needs".

The foregoing definitions reflect divergent and varied perspectives in the conceptualisation of information retrieval. The definitions however presuppose the existence of users' information needs requiring satisfaction, the availability of stored information in diverse formats, availability of manual or automated/system capable of ferreting out the required information with appropriate speed and accuracy. The key issue is that the retrieval process is very dependent on indexing and storage. To a large extent they determine the optimum strategy for searching an information retrieval system. For Rowley (1998), information retrieval systems have become almost synonymous with computers, but paper-based systems such as card and document filing systems, still exist and were in existence before the advent of computers.

1.3.2 The Concept of Information Retrieval

The phrase "information retrieval" was first coined in 1952, according to Spack and Willet (1997), and it acquired prominence in the research community in 1961. At the time, the organisational role of information retrieval was considered as a significant advancement in libraries, which were no longer merely book stores, but also locations where information

was cataloged and indexed. The idea behind information retrieval is that there are documents or records holding data that have been structured in a way that makes it easy to discover. As a result, an information retrieval system is implemented to obtain the documents or information that the user community requires. It should provide the appropriate information to the appropriate user. As a result, an information retrieval system tries to collect and organise information in one or more topic areas so that it can be provided to the user as soon as it is requested. Lancaster (1968) comments in Chowdhury (1999) Unrealistic that an information retrieval system does not inform (i.e. alter the user's knowledge of the subject of his inquiry); instead, it just notifies user of the related documents relevant to his request. The information retrieval system acts as a link between the world of knowledge authors and the world of information users.

Two broad categories of information retrieval have been identified:

- In-house Information retrieval and
- Online Information retrieval.

In-house information retrieval systems are design and developed by the individual library or information centre to serve primarily users within its environs. The library catalogue is an example of an in-house database. For illustration, the Online Public Access Catalogue (OPAC) allows library patrons to do online catalog searches and subsequently check the availability of the requested item(s). Online information retrieval systems are designed to give remote access to databases to every legitimate user irrespective of the user's location. Such services are generally provided on a commercial basis, and a variety of vendors offer them. In the past, scholars are the opinion that an effective and dependable information retrieval system should have the following:

- Prompt dissemination of information,
- Filtering of information,
- The right amount of information at the right time,
- Browsing,
- Getting information in an economical way,
- Current literature,
- Interpersonal communication, and
- Personal help.

There are two types of users in a conventional library system: library and information staff and end-users. Library and information staff frequently serve as intermediates, as well as end-users seeking information for personal use or decision-making. All information retrieval systems should be designed with the user in mind. As a result,

every user's interests should be taken into account at every stage of data storage and retrieval.

1.3.3 Elements or Components of Information Retrieval

Information Retrieval mainly consists of four elements which are briefly discussed below:

Information Carrier: This serves the purpose of transmitting information from one system to another. It is also storage device used to store information for future use. Examples are magnetic tapes, CD-ROM, DVD, and Memory stick, etc.

Descriptor: This is a term or terminology that is used to search for information from storage known as a descriptor. It can be a name, number, signal, etc.

Document address: Every document must have an author that identifies the location of those documents. In Library and Information Science profession, document address involves a call number, class number, ISBN, shelf number, accession number, file number, etc. all these help users in retrieving their required information.

Transmission of Information: Transmission of information means to supply any document that is required by the user. Information retrieval system uses various communication channels to do this. It works like source → transmitter → media → receiver → destination.

1.3.4 Functions of Information Retrieval

These include the following:

Acquisition: Acquisition is the first function of information retrieval. It is the process of selecting, ordering and receiving materials for library or archival collections. It can be through purchase, donation or gift and the resource can be books, documents, journals, etc.

Content Analysis: The second step is to evaluate the effectiveness of the information acquired.

Content Presentation: Information should be presented clearly and effectively so that users will be able to understand them easily. For this purpose (catalogue, bibliography, index, CAS) will help a lot.

Creation of Store: In this stage the library authority creates a new file for storing their information already collected and ready for presentation. Such information is always organised in a systematic way.

Creation of Search Method: The authority will decide what kind of search logic they may use for searching and retrieving information.

Dissemination: This is the last stage of information retrieval. It is an act of spreading information systematically. So the whole function happens like Acquisition → Content analysis → Information presentation →Creation of store → Creation of search logic → Dissemination retrieval result → Stop.

1.3.6 Types of Information Retrieval Systems (IRS)

Types	User Characteristics	Environment	Tasks	Technology
Online search services	Expert users and information managers	Office, academic library, corporate information centre	Retrieval of information, downloading of information and integrating into other documents	Range of different workstations ; earlier configurations with direct link to service, more state-of-the-art application links through the internet
CD-ROM	Depends on database. This can include children, general public library users, professional users and others	Library, airport, home, office	Retrieval of information, downloading of information, and integrating information into other documents	Often multimedia, graphical user interfaces(GUI), mouse

The Internet	Internet surfers- preponderance of academics and students	Study/work place, home	E-mail communication, shopping, file transfer , screen and mouse	Desktop and portable PCs with keyboard
OPACs	Library users – profile depends on type of library	Library, office/home, other public venues	Narrowly defined – identifying book availability, searching for information	Sometimes larger screen, special-purpose keyboard, but also accessed through standard office equipment; remote and local access may use different workstations
Document management systems	Corporate users with some shared experience of the system and shared OBJECTIVESand tasks	Office-based, but may also extend to mobile operation and use in production units and in, trains and cars	Consultation of corporate archives in the pursuit of work-based asks	Workstations linked to powerful central computing resources; some applications will be state-of-the-art

Figure 1.1 Types of Information Retrieval Systems (IRS)
Source: *Jennifer Rowley 1998, p160.*

1.3.6 Stages in the Development of Information Retrieval

In the last 30 years, computer-based information retrieval systems have gone through a number of phases (generations) (Rowley, 1998:162). These generations have had hands-on experience with a wide range of

information systems. While the third generation is currently the focus of attention, prior generations are as essential because:

a.) A variety of operational systems may be built on second-generation technology.
b.) They are the cornerstones of today's modern cutting-edge system.

Each generation is built on more advanced technology, which has ramifications for how the systems are used. When a result, as one generation succeeds the previous, the type of data held in systems, their connectivity, the user interface, and the type of the user group have all changed. Some of the characteristics of these generations are summarised in the table below.

Table 1.1
Table summarising generation of information retrieval

First generation	Metadata	Command-based interfaces, expert and intermediary users; a limited number of online systems within organisations and libraries and available externally through online search services.
Second generation	Full-text data	Menus and command-based interfaces; additional retrieval facilities, such as hypertext and full-text search facilities; DOS-based interfaces; end-user access intended but not always possible or archived; online systems, with early CD-ROM-based system.
Third generation	Multimedia	Graphical user interfaces GUI,; focus on end-user access; market orientation, and emphasis on product packages; storage and distribution on CD-ROM or over high-capacity networks; multimedia; intermediary as trainer; greater use in the home and other public access environments.

1.4 Information Retrieval Tools

Libraries have existed since the dawn of writing and have acted as repositories for society's intellectual resources. As a result, libraries have always been concerned with information storage and retrieval.

Libraries were required to make optimum use of information retrieval technologies to assist the storage and retrieval process as the amount of information rose enormously, such as cataloguing and metadata. Printed materials of many kinds have been and continue to be the traditional methods of information retrieval:

a. Books with chapter headings and indexes;
b. Handbooks and manuals with section headings and indexes;
c. Catalogues and bibliographies of books and other printed materials; abstracting and indexing publications, arranges by topic and with indexes, giving references to journal articles, technical reports, patent, etc.; and printed directories of people, institutions, manufacturers, etc.

Information retrieval systems may use the computer itself to store both the document files and index files, and to maintain data-bases.

Libraries provide appropriate tools to enhance the capability of users to navigate through their avalanche of resources for effective retrieval. The tools sometimes referred to as retrieval aids facilitate the location of bibliographic records and actual retrieval of associated books, journals, articles, newspapers and electronic resources either from the library shelves or through the world wide web. The relevant information retrieval tools are listed and discussed as follows:

- Indexes
- Abstracts
- Catalogues
- Metadata
- Classification
- Bibliographies

Indexes- An index is a detailed alphabetical list of topics, names of persons or places, treated in the volume, usually by page number but often by section or entry number. It serves as a systematic guide to the location of words, concepts, and other items in publications, documents and other records (Adeyoyin, 2011). It consists of entries appearing in some logical order, which enables the user to locate them easily relying on the information provided. It is a guide or facilitator to the source of information. Indexes provide surrogates or pointers or indicators to the resource the library patron is searching for. They do not retrieve the actual materials but point to the attributes of information located within a document for example, book index directs attention to pages where appropriate information can be obtained about certain subjects in the book. Indexes also provide lists of existing publications in a subject area

over a specific period of time (Dalrymple, 2001). Typical examples are the periodical indexes, science index, social science index, and so on. Indexes therefore help with appropriate information aids to track the needed article or topic in any given subject. They do not provide detailed contents but necessary guides to accessing the contents.

The process of assigning index terms or keys to a record or document is known as indexing. These index terms or keys aid in document and record retrieval in the future. The assignment of indexing terms can be either intellectual (i.e. done by a person) or automated (i.e. done by a computer) (Rowley, 1998). The nature of computer-generated indexes varies widely, and searchers will have a better chance of finding what they're looking for if they understand some of the inherent constraints.

Xin Lu (1990) in Onwuchekwa & Jegede (2011),) writes that in an ideal document retrieval environment, a document or query statement is represented by a set of distinct index terms as well as the semantic relationships between these terms, allowing retrieval to be based on a semantic relationship structure. Documents are retrieved based on the correspondence between search terms expressed in the query and index terms in the document, according to Macleod (1990), referenced in Onwuchekwa & Jegede (2011). Indexing systems that help with document retrieval work by manually or automatically designating index terms to the examined subject of each document.

The most complex part of indexing is that phase where two different indexers analyse the content of a given document in two different ways resulting in two distinct index entries. Information retrieval experts, on the other hand, have noticed that indexing is more consistent when the vocabulary used is controlled, because indexers are more likely to agree on the terms needed to describe a particular topic if they are chosen from a pre-determined list rather than being given a free hand. It is also easier for users/searchers to identify the terms that are relevant for the information requirement if the terms must be chosen from a predetermined list. Both subject heading list and thesauri contain alphabetically arranged terms with necessary cross references and notes that can be used for indexing or searching in an information retrieval environment. The different kinds of vocabulary control tools include subject heading list and thesauri.

A thesaurus, according to Rowley (1992) in Onwuchekwa & Jegede, is a collection of words and phrases that displays synonyms, hierarchical and other linkages and dependencies, and serves as a standardised vocabulary for information systems. Thesaurus terms are used to manage terminology in indexing and to help searching by alerting the searcher to the index terms that have been used. The thesaurus facet is a

specialised retrieval language that includes both a thesaurus and a classification type, each of which has a unique item not found in the other. A thesaurus facet's obvious benefit is that it may be used to organise books on shelves in a special library as well as index terms in a database. Some of these technologies, particularly the subject heading list in catalogues, have been used to organise internet resources such as the Infomine, Biome, and Sosig, among others. Automatic indexing, which has become more common as technology has advanced, is when content identifiers are assigned using current computer facilities. The lack of human expertise in an automatic indexing environment can be solved by judicious use of common vocabulary in stored records and information requests. Additional benefits of automatic indexing include maintaining indexing uniformity, saving indexing time, producing index items at a lesser cost, and improving retrieval effectiveness.

Abstracts - Abstract is a summary of the essential points of a publication or article, accompanied by an adequate bibliographical description to enable the publication to be located; a short summary of a longer work such as a Theses/Dissertations or research paper. The abstract concisely reports the aims and outcomes of your research so that readers know exactly what the paper is about. A concise summary of an academic text, it serves two main purposes: to help potential readers determine the relevance of a paper for their own research; and to communicate key findings to those who do not have time to read the whole paper. An abstract may be indicative, informative, critical, and subject specific. Indicative abstract preponderantly points to the original documents with summary of key issues and appropriate pages. Informative abstract goes beyond direction to providing comprehensive and concise information about the publication. Critical abstract engages in value judgment or a critique of the document. Subject abstract is concerned with and limited in scope to specific subjects, for example, chemical abstracts, science abstracts, Library and Information Science. An abstract is a retrieval reference source in which books that are mainly contribution to periodicals are summarised. Abstracts provide succinct overview of any publication. It is a concise summary or condensation of the work, highlighting the major points and findings of the publication. The purpose of abstract is to help potential readers determine the relevance of the material without retrieving and reading through the entire journal or book. The appropriateness of the information to the user's task or assignment as gleaned from the abstract will determine the need to locate and retrieve the work or opt for an alternative. There exist databases with online abstracts to enhance the searching capability of the library patron to retrieving myriads of relevant information through a mere click of the mouse. Examples of such abstracts include Library and Information Science Abstracts (LISA); Biological Abstracts; Chemical Abstracts; Physical Abstracts;

Environmental Abstracts; Thesis and Dissertation Abstracts; Library, Information Science and Technology Abstracts, Science Abstract, etc.

The abstract has been defined by the authors from various perspectives. An abstract, according to Lancaster (2003), is a succinct but accurate depiction of the contents of a document, as opposed to an extract, annotation, or summary. An abstract, according to Rowley (1996), is a brief and accurate portrayal of a document's content in a style close to the original material. It usually follows the style and layout of the parent document and includes all of the important points made in the original document. Abstracts are always short texts that accompany the original document or are included in its surrogate as documentary output. Other information scientists have used various criteria to identify distinct types of abstracts, and these are included below:

Abstract by writer*:* These are abstracts written by authors, subject experts, or by professional abstractors.

- **Abstracts by purpose:** They are written to serve different purposes for example informative abstract, indicative abstract, critical and special purpose abstracts.

- **Abstract by form** is another type of abstract and can be categories as structured abstract, mini abstract and telegraphic abstract. With the rapid increase in the availability of full text and multimedia information in digital form, the need for automatic abstracts or summaries as filtering tool is becoming extremely important. Craven (2000) in his works proposes a hybrid abstracting system in which some tasks are performed by human abstractors and others by an abstractor assistance software. A number of library classification and catalogue codes have been devised to help in the process of organising library items systematically so that information may be retrieved quickly.

Library catalogue provides bibliographical records of all formats of items in the library collection. It serves as the directory of all holdings in the Library. It represents a major retrieval tool which assists the user to locate materials either by author, subject, and title. There are basically two kinds of library catalogues:

- Single Library Catalogue which provides an index to all holdings in a particular library.
- Union catalogue which refers to joint catalogue of group of networked or participating libraries.

- Provides a complete bibliographical description of every item in the collection to show its uniqueness relative to other items for easy identification and retrieval.
- Shows the Library holdings in respect of particular authors, subjects and genres.
- Enables patrons to locate an item on the shelves, provided the item has not been borrowed or miss-shelved or stolen.
- Enables a patron to know if a library has a particular item in its collection that meets the information need.

Svenonius (2000) lists the following major LEARNING OUTCOME of a catalogue, identified by Cutter, as:

- To allow a person to search for a book by author, title, or subject.
- To show what books the library possesses by a given author on a given subject and in a specified genre.
- To aid in the selection of a book based on edition or character.

However, Cutter's stated aims were slightly amended in 1997, and the catalog's OBJECTIVES were rewritten as follows:

- In order to be compatible with automated cataloguing environments,
- To locate entities that match the user's search criteria,
- To locate an entity,
- To choose an entity that is suited for the user's needs, and
- To inquire about or gain access to the entity specified (IFLA, 2011).

In the works of **Large and Behesti (1997),** Online Public Catalogues (OPAC) were first utilised in the mid-1970s, according to Large and Behesti (1997), but it was not until the beginning of the previous decade that a substantial number of libraries transitioned from card-catalogues to automated-catalogues. However, the initial catalogues were mainly modules tied to the automated circulation system, with limited capability and brief catalogue records.

Online public catalogues (OPAC) are user interfaces that allow users to communicate with a library's collection(s). OPACs allow users to search the library's catalog as well as perform other tasks like checking borrower records, reserving reading materials, and receiving library news bulletins. Since then, several modifications have occurred, and OPACs have vastly improve.

The content (text, numeric, audio, image video, etc.) of Internet resources varies, as does the file type, availability, uniform resource

locator (URL or the address of a web page), and so on. To assist catalogers in making information retrieval simple and successful, some new rules and principles are required.

A catalogue therefore, provides a comprehensive and analytical bibliographic description of every item in a library collection. It is the commonest retrieval tool in Nigerian libraries. It provides surrogates or bibliographical records of all materials in the library which are arranged in a predetermined order, especially author, title and subject. It describes other elements such as publisher, edition, place of publication, series, illustrations, pagination, ISBN of each of the holdings.

The broad adoption of online catalogue has propelled the library into the digital age, placing libraries at the forefront of information access at academic institution.

Metadata: Schwartz (2001) The term metadata, been used largely in the field of database management, first appeared in LIS literature in the mid-1990s, according to Schwartz (2001). However, within a short amount of time, the topic had become a very prominent area of research, with several hundred publications. Metadata has been divided into five categories based on how it is used:

- Administrative metadata, which is used to manage and administer data resources.
- Descriptive metadata is used to describe or identify an information resource.
- Preservation metadata, which is relevant to the management of information resources' preservation.
- Use metadata, related to the level and types of use of information resources Metadata can be used for a variety of purposes, and metadata consumers can be humans or computer systems.

Metadata's main purpose is to make it easier to identify, locate, retrieve, manipulate, and utilise digital things in a networked environment. Since the introduction of the internet and the web, metadata has become a crucial issue for information organisation.

The rise of digital libraries has raised awareness of the necessity for metadata for a wide range of digitally available materials

Classification - The first library classification Scheme was developed by Melvil Dewey in1876. Universal Decimal Classification (UDC) was the second major classification scheme to appear. Another classification scheme is the Library of Congress Classification Scheme (LCCS). It is an example of a semi-enumerative categorisation scheme, in which a

long list of all the classes in the universe of subjects is provided. Although library classification schemes were originally created to organise bibliographic items on library shelves, many librarians and information professionals have adapted them to organise electronic information resources. The following are some example:

BUBL LINK (www.bubl.ac.uk): provides access to a catalogue of over 11,000 carefully chosen online resources, all of which are cataloged using DDC on all academic sources. Users can search the catalogue by selecting a Dewy class, for as 300 Social Sciences, or by using the alphabetical index to find a term/phrase. As a result, a user can access digital resources through a categorised list or an alphabetical list of subjects.

CYBER DEWEY: This is another example of DDC being used to organize digital information resources, and it dates back to 1995, when David Mundie utilised it to arrange internet content. Similar to the Bubllink, the Cyber Dewey allows users to search the catalogue for material by selecting a specific Dewey class, word, or phase from the alphabetical index.

Bibliography - It is a systematic list of materials which provides description and identification of the edition, dates of issue, publisher, authorship and type of publication. It gives a list of publications of a specific publisher or author. It is a useful tool in literature search on any subject of interest. In general, a bibliography should include: the author's names, the titles of the works, the names and locations of the publisher, the dates of publication and the page numbers of the source.

Bibliographies - They are compilations of publications on a specific topics, subjects or authors. They can also refer to totality of published works within a geographical entity in a specific period of time (Lanning and Bryner, 2009). They include a list of journal articles, books, technical reports, conference proceedings, theses, and dissertations on a specific topic or by a specific author. The given items are organised by authors, titles, and subjects as a reference source. Bibliographies provide relevant avenue for retrieving existing published works on any topic, subject or authors. There are different types of bibliographies, namely:

- Annotated bibliography: a descriptive summary of each publication aimed at assisting users determine the relevance for retrieval and subsequent utilisation.
- Critical bibliography: This provides a critique of the listed publications in addition to annotations.
- Trade bibliography: It provides requisite information for potential buyers to aid their selection of publications for

acquisition. It is usually issued by booksellers and commercial publishers to the general public to create awareness for the materials, prices, agents or bookstores where the materials can be procured. The book seller, books-in-print, and publishers trade list annual are only a few examples. Non-trade publications, such as theses, dissertations, and government documents are normally excluded (Penka, 2001).

- Subject Bibliography: This is a compilation of bibliographies that is restricted to a particular subject, specialisation or topic. It has the target audience as researchers and others engaged in specialised scholarship. Economic development in Nigeria, a bibliography of librarianship, and a bibliographic guide to technology are some examples of subject bibliographies.
- National Bibliography: This refers to all published works within a national geographical entity or country. It sometimes has an enlarged scope to include publications written about the country, or in the language of the country, or by the citizens of the country, irrespective of the place of publication. It also includes complete cataloguing data and is usually compiled by the National Library of respective countries. Some examples of national bibliographies include: National Bibliography of Nigerian (NBN), British National Bibliography (BNB), and Ghanaian National Bibliography, American Books Publishing Records (ABPR), British Books Publishing Records (BBPR). Bibliographies are therefore very relevant aids in information retrieval. They assist to locate materials on any topic or subject and provide means of verifying bibliographic records as well as evaluative comments on the suitability of the material.

1.4.1 Techniques of Information Retrieval

Two techniques are used to retrieve information effectively and they are briefly described in below:

Traditional System*:* Traditional system involves the following segments:

- Catalogue: A catalogue is a list of books that libraries use to locate items. It includes information such as the author's name, title, edition, number of volumes, publication date, page number, series name, and ISBN. It is a necessary tool for information retrieval.
- Index: An index is a list of entries organised in a systematic way to help library patrons find information in a document. It is an important tool for retrieving information.

- Abstract: This is an exact and succinct representation of the original material. It allows users to get a general idea of the content without having to read it completely. It can also be used to retrieve information.
- Bibliography: A bibliography is a list of books that is not restricted to a single collection.
- Authority File: Library Authority will sometimes use its own methods to retrieve information rather than using current methods. This strategy is frequently written down in books or files for future use,

Non-traditional System: Non-traditional systems are divided into two main categories namely:

- Semi-automatic System: A semi-automatic system combines human and machine retrieval of data. This method necessitates human intelligence and labor in addition to the usage of machines. Semi-automatic systems include catalog cards, punch cards, notes cards, apache cards, and so on.
- Automated System: This system is mostly dependent on a single computer and networking technology. In the 1970s, an offline network was used, but from 1980 to the present, real-time sharing has been used. Computer, modem, CD-ROM, hard disk, internet, and other automatic tools are examples. The following are their descriptions:
- **Computer:** An electronic machine that can store and work with large amounts of information.
- **CD-ROM:** A compact disc used as a read-only optical memory device for a computer system.
- **Hard Disk:** A right non-removable magnetic disk with a large data storage capacity.
- **Floppy Disk:** A flexible removable magnetic disk (typically encased in a hard-plastic shelf) for storing data.
- **The Internet:** A global computer network made up of interconnected networks that use standardized communication protocols to provide a variety of information and communication services.

1.5 Summary

Information retrieval focuses on the provision of reliable mechanism that facilitates speedy access to required information from the total collection with highest level of accuracy and precision. It guarantees user satisfaction and optimal utilisation of information resources. The effectiveness of libraries as a reservoir of knowledge and the memory of society lies on the functionality and capabilities of the information

retrieval framework. The usefulness of acquisition, organisation through classification and cataloging as well as storage of information resources can only be appreciated if users can retrieve the material with ease as and when needed. Information retrieval is pervasive and thus applicable to all formats of information.

This unit examined the various definitions of information retrieval and noted that it involves gamut of intricate activities that deal with the representation, storage, organisation and access to relevant information. It recognises that exponential growth in information resources and demand for a corresponding rate of accessibility. Information retrieval is all-embracing, transcending print or manual–based information to multimedia objects, computer-based, and Internet resources.

The pattern of using information retrieval systems or information seeking behavior is influenced by a number of factors:

- The users' awareness of and capacity to access different information sources,
- The relationship between the user and the information unit in question,
- The simplicity with which the information unit can be accessed,
- The working conditions of the users,
- The time available to the user for consulting information systems,
- The amount of competition that exist in the users' field of activities,
- The users past experience or knowledge,
- How easily the user gets on with other people,
- How friendly, knowledgeable, and efficient are the members of the information unit, and the various products and services of the information.

Self-Assessment Exercise(s)

1. What do you understand by the term "information retrieval"? (b.) Justify your answer with at least five (5) relevant authorities in this regard.
2. Information retrieval as a concept is pervasive and applicable to all formats and sources of information. Discuss.
3. In a tabular form highlight the types, features environment and the technology of information retrieval tools.
4. Briefly discuss the stages of information retrieval and indicate why the first stage is still very important today.
5. The effectiveness of libraries as reservoirs of knowledge and the memory of society lie on the functionality and capabilities of information retrieval techniques, Discuss.
6. Discuss at least five basic tools that aid information retrieval?

1.6 References/ Further Reading/Web Resources

Craven,T.C. (2000). Abstracts Produced using Computer Assistance, *Journal of the American Society for Information science*, 745-56.

Chimah, J. N., Unagha, A. O. and Nwokocha U. (2010). Information Retrieval in Libraries and Information Centres: Concept, Challenges and Search Strategies.

International federation of Library Association (IFLA) (2011). (International Federation Vol. 5 (6), Serial No. 23, Pp. 108-120 Copyright © IAARR 2011: www.afrrevjo.com 112 Indexed African Journals Online: www.ajol.info of Library Associations and Institutions) study group

Journal of Applied Information Science and Technology, 4. Retrieved from http://www.jaistonline.org/chimahunaghaNwokocha_2k10.pdf.

Chowdhury, G.O. (2017). Introduction to modern Information Retrieval, (3rd ed.). Facet Publishing.

Dalrymple, P. W. (2001). Bibliographic Control, Organization of Information, and Search Strategies. In R.E. Bopp, R. E., and L. C. Smith (eds.), *Reference and Information Services: An Introduction*. Englewood, Colorado: Libraries Unlimited, pp. 69 - 96.

Edom, B. O. (2012). Principles of the use of library. Owerri: Springfield Publishers Ltd.

Lancaster, F., Elliker, C. and Colonell, T. H. (20030. *Subject Analysis*. Annual *review of information science and technology*, 24, 34-84.

Lanning, S. & Bryner, J. (2010). Essential Reference Services for Today's School Media Specialists. Santa Barbara, CA: Libraries Unlimited.

Large, A. and Behesti, J. (1997). 'OPAC: a research review'. Library and information science research, (19), 111-33.

Manning, C. D., Raghavan, P and Schutze, H. (2009). An Introduction to Information Retrieval. Online Edition, Cambridge University Press, Cambridge, England. https://nip.standford.edu/IR-book/ pdf/irbookonlinereading.pdf

Ojedokun, A.A. (2007). *Information Literacy for Tertiary Education Students in Africa*. Third World Information Services Limited, Ibadan, Nigeria. Pp 65-77

Onwuchekwa, E.O and Jegede, O.R. (2011). Information Retrieval methods in libraries and information centers. *African Research Reviews: An International Multidisciplinary Journal, Ethiopia* 5 (6), 108-120

Rashid, H. A. (2020). Information Retrieval. Library & Information Management. Retrieved from https://nip.standford.edu/IR book/pdf/irbookonlinereading.pdf

Rowley, Jennifer (1998). The Electronic Library: fourth edition of Computers for libraries. London: Library Association Publishing.

Schwartz, C. (2001). Sorting out the web: Approaches to Subject Access, Westport, Ablex Publishing.

Svenonius, E. (2000). The intellectual foundation of Information Organization, Cambridge, MA: MIT Press.

Vickery, B. & Vickery A. (1992). Information Science in theory and practice - Revised edition. London: Bowker-Saur.

Unit 2: Objectives and Functions of Information Retrieval

Unit Structure

2.1 Introduction

The previous Unit explored the definition and tools of information retrieval as well as underscored the exponential growth in information resources of diverse formats which necessitate the need for accurate mechanism for ensuring corresponding rate of accessibility and search ability to forestall underutilisation and user frustration. It is against this background that unit 2 will examine specifically the Objectives And functions of information retrieval with the aim of stimulating interest and deeper appreciation of the concept and its versatility in human endeavors.

2.2 Intended Learning Outcomes

By the end of this unit, you will be able to:

- discuss eight Objectives of Information Retrieval as enunciated by Onwuchekwa
- identify the six functions of information retrieval.

2.3 Objectives and Functions of Information Retrieval

It was US Government cited in Nazlin (n.d) E-learning and Libraries by that "stated our nation's educators and institutions of learning must be aware of … and adjust to … these new realities … the ability to seek, find, and decipher information can be applied to countless life decisions, whether financial, medical, educational or technical". The above quote aptly illustrates the extensiveness and application of information retrieval. Since information is critical to every sphere of human endeavor, in a knowledge–driven world characterised by information explosion, the utility of information retrieval is enormous. Beyond locating information from bibliographic records of manual library catalogues, we can now readily access information for diverse purposes

from the global information network within seconds, thus bridging information inequalities among people and nations. We can navigate through library collections through online public access catalogue (OPAC). There is now the capability of retrieving information from subscription-based databases, institutional repositories, websites, social networking sites, YouTube and from the internet leveraging on plethora of search engines.

3.1.1 Objectives and Functions of Information Retrieval

The core objective of information retrieval is to ferret out relevant information or information surrogates as and when needed from the mass of collection or database in response to a user's query/request. Onwuchekwa (2011) highlighted eight Objectives an information retrieval must seek to actualise as follows:

1. Prompt dissemination of information;
2. Filtering of information;
3. The right amount of information at the right time;
4. Browsing capability;
5. Getting information in an economical way;
6. Provision of current literature;
7. Interpersonal communication; and
8. Personal help.

1. **Prompt Dissemination of Information:** This is predicated on the information retrieval system meeting set goals within a timely frame. The response rate to the user's enquiry must be at such speed that guarantees effectiveness. Any form of sluggishness or downtime tends to frustrate the retrieval effort. The system is expected to execute commands without delay and perform regularly within the specified time limit.

2. **Filtering of Information:** This involves removal of unwanted information from the maze of information using digital mechanism. It seeks to tailor retrieval to the preferences of the user and search specification. Every good retrieval system must therefore have the capability of sorting and discriminating against or restricting the flow of undesirable information. It selects the most valuable to meet the user's search queries.

3. **The right amount of information at the right time:** Corollary to the issue of filtering is that the system should provide the right amount of information at the right time to avoid information overload. Excessive amount of information could be overwhelming and makes comprehension as well as utilization

pretty difficult. Since retrieval of too much information tends to be counterproductive, the objective of the system to ensure that the right amount information is retrieved at the right time.

4. **Browsing capability:** It provides an embedded application or interface which enables users to locate, view and retrieve relevant information. There is usually a dialogue box through which search queries are sent to the system. It crawls through the databases to ferret out information that matches the search terms and the output is displayed as feedback to the patron's queries. Browsing capability is therefore a core characteristic of any retrieval system. The user needs to understand the peculiarities of each system's browsing capability and features for optimal utilisation.

5. **Getting information in an economical way:** Some of the barriers to information are cost, language and distance. If users pay prohibitive fees to retrieve information or pay to translate the information into intelligible language or format, it is therefore not economical. All forms of barriers that make it expensive to retrieve appropriate information must be eliminated for effectiveness. This is particularly important in provision of information to the visually impaired persons and other disadvantaged persons in society.

7. **Provision of current literature:** Information retrieval systems should enhance retrieval of up-to-date information in any field of choice. This helps to ensure reliability and authenticity of contents. It should also have the capability to enable users delineate the scope or coverage or period desired for retrieval. For example, the user can request that only information covering the period of 2019-2021 alone should be retrieved. It stands to reason that the system must be constantly up-to-date with latest information or literature to maintain relevance as users abhor obsolete or stale information.

8. **Interpersonal communication:** Modern retrieval systems provide for interactive interface. This allows for online chats, newsgroups, discussion forums, video conferencing and email. Online chat which is the most common involves users communicating using the keyboard and receiving feedback as well. It could be synchronous or asynchronous. The former depicts simultaneous and instant communication. The latter involves leaving a message and receiving feedback later.

8. **Personal help:** Personal help features are integral part of information retrieval system. It provides an interface for users having difficulties in navigating the system to request for specific helps or services. It makes available detailed procedures to resolve retrieval problems or trouble-shooting guidance. This helps to minimise or eliminate user frustrations in the process of searching for requisite information.

The Objectives Anticipate that library users' interest must be paramount in the storage of information and the design of a retrieval system that allows users unfettered interaction with the different information sources to facilitate appropriate retrieval and variety of ancillary operations. It has been observed that library users are not homogeneous in their information seeking behavior. Some users know exactly and specifically the kind of information they are looking for while others do not know exactly until they find same. Information retrieval is expected to aggregate both perspectives and provide several access points to meet the diverse information needs.

Chowdhury (2019) summarised the six functions of information retrieval as detailed below. They are to:

(a) Identity the source of information relevant to the areas of interest of the target users' community:
(b) Analyse the contents of the sources;
(c) Represent the contents of the analysed sources for matching with the users' queries;
(d) Match the search statement with the stored databases;
(e) Retrieve the information that is relevant; and
(f) Make necessary adjustments in the system based on feedback from the users.

In the same vein, Rashid (2020) identified six functions of information retrieval as constituting the following:

(i) Acquisition: the need to collect information from diverse sources to provide robust, comprehensive and up to date contents. The acquisition could be through outright purchase, subscription, donation / gift, Curation, etc.
(ii) Content Analysis: This involves distilling the contents to generate appropriate descriptors or taxonomies for effective classification based on prevailing international standards.
(iii) Content presentation: It involves the creation of representations that sufficiently characterise the information resources. The

presentation should be explicit and intelligible for users to comprehend.

(iv) Creation of store: provision of appropriate databases, files, shelves, hard Disks, websites etc. for the storage of the systematically organised information resources for later retrieval / dissemination.

(v) Creation of search method: This involves creation of search logic for browsing, searching and retrieving information from the store. It provides appropriate interface for users to navigate through information sources.

(vi) Dissemination: This last stage deals with retrieval results in response to the request entered into the system by the users. The result is expected to meet or exceed the users' expectations for accurate, relevant, precise and comprehensive information within the highest possible time.

The functions of information retrieval as enunciated by Rashid (2020) resonate perfectly with those earlier expressed by Chowdhury (2019) and both are tangential to actualising the eight Objectives of information retrieval.

2.3 Summary

We discussed the ubiquity of information to all facets of life and the current reality of a knowledge-driven world characterised by information explosion. The dire necessity to ferret out relevant information from a plethora of sources is the major thrust of information retrieval. The objects of information retrieval include prompt dissemination of information, filtering of information, providing the right amount of information at the right time, providing a browsing capability, getting information in an economical way, provision of current literature, interface for interpersonal communication and personal help. Chowdhury summarised the six functions of information retrieval as follows:

(a) Identify the source of information relevant to areas of interest of the target users' community.

(b) Analyse the contents of the sources.

(c) Represent the contents of the analysed sources for matching users' queries.

(d) Match the search statement with the stored databases.

(e) Retrieve the information that is relevant.

(f) Make necessary adjustments in the system based on feedback from the users. Rashid also labeled the six functions as:
(i) acquisition, (ii) content analysis, (iii) Content presentation,

(iv) creation of store, (v) creation of search method, and (iv) Dissemination.

Self-Assessment Exercise(s)

1. Discuss the OBJECTIVES and functions of information Retrieval
2. The importance of information retrieval lies in its ubiquity and enormous application to all spheres of human endeavour. Discuss.
3. Identity the six functions of Information Retrieval as Propounded by Chowdhury (2019).

2.4 References/Further Reading/Web Resources

Nazlin B. (2016). "E-learning and Libraries". In The Sage handbook of E-learning Research edited by Haythornthwaite, C. A. Retrieved https://discovery.ucl.ac.uk/id/eprint/10101930/1/Bhimani_ELearning%20and%20Libraries_%20Accepted%20Manuscript%20.pdf

Onwuchekwa, E. (2011). Information Retrieval Methods in Libraries and Information Centers. African Research Review 5(6) pp. 118-176, 108-120Doi:10.4314/afrrev.v516.10

Chowdhury, G.G. (2019). Introduction to Modern Information Retrieval, (3rd Ed). London: Facet Publishing.

Rashid, H.A. (2020). Information Retrieval. Library & Information Management. Retrieved from https://limbd. Accessed on April 21, 2021.

Module 2 Information Representation and Retrieval

Unit 1 Approaches to Information Representation
Unit 2 Language in information Retrieval

Unit 1: Approaches to Information Representation

Unit Structure

1.1 Introduction

Having learnt in the precious unit the Objectives And function of information retrieval as the cornerstone for providing requisite access points to the diverse sources of information with the aim of meeting the search queries of the library users, this unit will explore another layer of information retrieval known as information representation and its different approaches. Information representation of all formats of information provides unique identification for items in the collection leading to effective retrieval devoid of trial and error syndrome. The unit will therefore explain the concept of information representation, discuss the different approaches to information representation and discuss the three stages of subject indexing approach of familiarisation, content analysis and translation of terms into controlled vocabulary.

1.2 Intended Learning Outcomes

By the end of this unit, you will be able to:

- explain the concept of information representation
- discuss different approaches to information representation
- discuss the three stages of subject indexing approach.

1.3 Information Representation

Information in this context connotes all library materials which have been hitherto stored for users' access, retrieval and utilisation. It encompasses text, document, multimedia objects including audio, video, image, curated digital resources, Internet resources, databases, books, journals and reports. The format of information notwithstanding, it must be represented to enhance retrieval. Chu-Heting (2003) defined information representation as "" the extraction of some elements (keywords or phrases) from a document or the assignment of terms (descriptors or subject headings) to a document so that its essence can be characterised and presented. It is tagging or labeling or symbolic identification of information resources to guarantee their recoverability from the maze of collections. This is creation of attributes to aid discovery of information beyond mere casual browsing and reliance on chance or serendipity. It is about providing unique labels for directed location of library materials.

Information representation can be done via any combination of the following means: abstracting, indexing, categorisation, summarisation and extraction". Roshdi and Roohparvar (2015) and Djoerd (2009) noted that information representation is called the indexing process which is done behind the scene and excludes the involvement of end-users. It is chiefly aimed at the creation of terms, words, notations, tags that significantly and sufficiently characterise the information resources in the collection for uniqueness thereby facilitating retrieve ability.

1.4 Approaches to Information Retrieval

Chu-Heting (2003) identified the different approaches for information representation in library service as follows:

- Alphabetical subject approach
- Hierarchical or Subject Indexing approach
- Coordinate Representation approach
- Computer-based application approach
- Statistical Methods approach

1.4.1 Alphabetical Subject Approach

This involves the grouping of information and surrogates under subjects and further arranged into alphabetic order for easy retrieval. The library professional must clearly resolve the issue of synonyms, homographs, singular or plural forms, complex and compound words or subjects in the alphabetical subject approach. It also recognises and caters for relationships among subjects and terms; these include syntactic and semantic relationships. The former refers to relationships among words

and phrases as depicted by their arrangement. For example, a keyword search for "photographs and Albums" should permit patrons to indicate if they want "photographs of Albums" or "Albums of photographs". The latter related with the meanings of words. For example, there is semantic difference between mercury (planet) and mercury (metal) irrespective of similarity in sound and spelling. It is a departure from the hitherto classified order or classified catalogue used by libraries. The classified system only allowed broad subject grouping contrary to the detailed subject specification of the alphabetical subject approach pioneered by Melvil Dewey. This made possible the arrangement of books on shelves and arrangement of entries in catalogues and bibliographies.

1.4.2 Hierarchical or Subject Indexing Approach

This is one of the traditional approaches which rely on subject and classification schemes to provide content or information representation for documents and other materials. It is the mechanism of identifying and selecting descriptors, taxonomic categories, notations or terms which adequately express the contents of the information resources. This helps to facilitate the retrieval of information based on its subject content. Subject indexing approach involves three steps, to wit: familiarisation, content analysis and translation.

- *Familiarisation:* The indexer establishes a general overview of the contents of the document or materials through reading the title, preface, forward, table of contents or skimming through the chapters. Beyond generating keywords, phrases or taxonomic categories from the material, the indexer is expected to distil the target users and delineate the scope with the aim of aligning with the information seeking behaviour of the potential users.

- *Content Analysis:* Based on professional judgment the indexer generates concepts, terms covered in the book or document as index terms. The terms must represent the thrust of the book or document. Indexers with requisite subject backgrounds are more versatile with the terminologies of the discipline involved, and thus find content analysis easier and fascinating. For standardisation, only terms derived from the controlled vocabulary are used as index terms or access point for retrieval. The concepts or terms generated from the material in the process of content analysis will be translated into an indexing language. For purposes of information representation, the indexer assigns a descriptor or subject heading to match the concept or term identified during content analysis. The descriptor is chosen from related terms, narrower or broader terms listed in the controlled vocabulary or thesaurus. Such controlled vocabularies include the Library of

Congress Subject Heading List (LCSH) Sears List of Subject Headings (SLSH).

- *Translation of terms into controlled vocabulary*: The indexer selects the most appropriate descriptor from the controlled vocabulary that matches the terms or concepts derived from the book or document as a perfect representation of the contents of the material. The assigned descriptor is used as access point for retrieval by potential users.

- *Coordinate Representation*: This involves the application of uniterms and Boolean logic for information representation and retrieval. The concept of uniterms is premised on the belief that documents and materials could be broken into different facets, where each facet stands for a unique idea or theme of the document. Uniterms are therefore "individual terms selected by indexers to represent different facets of a book or document." The terms or keywords are differently derived from the original document without recourse to controlled vocabulary.

- Coordination also thrives in the belief of the Boolean logic that the entire subject of any document or book can be a summation of concepts or classes of objects, to form complex concepts, or the subdivision of complex concepts into individual, simple concepts. The former is regarded as the AND operator, and the latter as OR and NOT operators. The coordination representation utilises conjunction of AND, OR, NOT operators as well as combination of terms from the document to produce narrower descriptor or index term for enhanced access point to the material.

1.4.3 Computer-based Application Approach

This approach deals with the deployment of computer applications or software for indexing or recording information. It is a departure from manual subject indexing to automatic indexing and abstracting. The application collocates frequently occurring keywords to represent plural facets of a document. Two notable computer applications in this regard are keyword In Context (KWIC) and Keyword out of Context (KWOC). The former identifies high-frequency keywords and the latter; proximity of non- significant keywords outside the document. The combinations of keywords in context and out of context are used for constructing appropriate information representation.

1.4.4 Statistical Methods

This is advancement over the computer – based application approach. It leverages on Machine Readable Cataloging (MARC) standards and other algorithms to represent internet resources and other global networked information resources. The use of transmission control protocol and the Internet protocol (TCP/IP) help to mainstream MARC records of library collection (sounds, videos, audios, images and other multimedia objects) into the cyberspace, thus library collections which had been converted into MARC records can be virtually represented on the web. Automatic indexing of databases and institutional repositories are also made possible. Internet algorithms have facilitated the emergence of diverse formats of information to be represented and recognised by heterogeneous platforms and systems.

3.0 Summary

Information representation is one of the major preconditions for effective retrieval system. It is the creation of significant descriptors, taxonomic categories, and notations or terms that adequately express the contents of information resources or their surrogates. Alphabetical subject approach engages in detailed subject specification and further arrangement into alphabetic order. Subject indexing approach assigns the most appropriate descriptor from the controlled vocabulary to represent the information. It involves three stages of familiarisation, content analysis and translation of terms into controlled vocabulary. The coordination representation used Boolean logic operators and uniterms to produce index terms for multiple access point to the material. Computer-based application uses two notable indexing software, Keyword In Context (KWIC) and Keyword Out of Context (KWOC) for automatic indexing and abstracting. The statistical methods facilitated through Internet algorithms ensure representation of information on the cyberspace and other global networked information resources.

Self-Assessment Exercise(s)

1. What is information representation?
2. Discuss five (5) approaches to Information Representation.
3. Subject indexing approach involves three stages of familiarisation, content analysis and translation of terms into controlled vocabulary. Discuss

1.6 References/Further Reading/Web Resources

Chu-Heting (2003). Information Representation and Retrieval in the Digital Age. Medford; NY: Information Today for the American Society for Information Science and Technology

Djoerd, H. (2009). Information Retrieval Models in Goker, A. and Davies J. Information Retrieval: searching in the 21st Century. London: John Wiley & sons.

Roshdi, A. and Roohparvar, A. (2015). Information Retrieval Techniques and Applications. *International Journal of Computer Networks and Communications Security*3 (9), 373 – 377.

Unit 2: Language in Information Retrieval

Unit Structure

2.1 Introduction

Language plays pivotal soles in both information representation and retrieval. The professional indexer relies on words, phrases and terms to accurately describe the content of any document. Similarly, the degree of retrieval success or failure is predicated on the choice of words as search terms by the library patrons. This unit explores the concept of indexing language, natural indexing language, free indexing language and controlled indexing language.

2.2 Intended Learning Outcomes

By the end of this unit, you will be able to:

- explain the concept of indexing language
- differentiate between controlled and natural indexing languages
- explain the meaning of free indexing language.

2.3 Language in Information Retrieval

An indexing language is a set of terms or codes that can be used as an index's access point. In contrast, a searching language refers to the terms that a searcher uses to express a search criteria. During a search, the same terms or codes can be used as access points to records. The ultimate objective of information retrieval is locating accurate and relevant information with ease and within the quickest timeframe to meet the patron’s need. This is only made possible when the language of searching matches the indexing language used in information representation. Indexing language is the fulcrum upon which information representation and retrieval thrive.

The National Institute of Open Schooling (2012) defined indexing language as “as the set of terms used in an index to represent topics or

features of documents and the rules for combing or using those terms". It provides lexical elements which act as clues to the subject of the document. The basic rudiments of indexing language are the vocabulary, syntax and semantics. Vocabulary relates to the terms, phrases, descriptors used to represent the contents of the documents. Syntax deals with the arrangement of words and phrases to form index term. Semantics convey the contextual meaning of words, indicating synonyms, preferred spellings, hierarchical and associative relationships.

There are basically three types of indexing languages, to wit:

(a) Natural Indexing Language
(b) Free Indexing Language
(c) Controlled Indexing Language

Each of these is briefly discussed below.

2.3.1 Natural Indexing Languages

Natural Indexing Languages or Derived Term Systems: Natural indexing languages or derived term systems are not really a distinct or ordinary language of the document being indexed, rather than a natural or ordinary language. To put it another way, this is not artificial language but language used in normal communication or ordinary context of the document. This means that it is derived from terms and phrases contained in the document. A derived term system is one in which all descriptors are derived from the indexed document. It is premised that the terms or phrases used are discipline specific or terminologies that are familiar to the subject hence not strange to users for information retrieval. Natural indexing language could be created manually or automatically by computer indexing. The computer may index every phrase in the text, with the exception of a restricted stop-list of all common terms, or it may only index terms from a computer-held thesaurus. That is to say the indexer or computer assigns the descriptor derived from the title, abstracts or the full-text of the document. Indexing using natural language is not subject to restricted vocabulary and enjoys wide latitude of terms.

The library patron is also at liberty to formulate the search query in the ordinary language parlance for possible retrieval. Natural indexing language is said to enjoy the advantage of high specificity, no update required, absence of incompatibility between information representation and retrieval, high recall as well as exhaustively. However, it does not cater adequately for synonyms, where different terms refer to the same entity; homographs also introduce ambiguity in natural language indexing as terms with the same spelling and pronunciation but different meanings

present problematic scenario, for example, fine (a good quality or a levy) or bat (sports equipment or an animal).

2.3.2 Free Indexing Language

Free indexing language differs from natural indexing language in that the latter is limited by the language of the document being indexed, whilst the former is not (i.e., any appropriate term can be assigned). Unlike natural indexing language that is restricted by the terms and phrases contained in the document, free indexing language draws from anywhere adjudged suitable for representing the contents of the document. This means that free language indexing is dependent on the prerogative and skills of the indexer. Furthermore, free language indexing might be done manually by a human indexer, with the quality of the index relying heavily on the indexer's understanding of the subject and terminology. Because the computer must have some basis on which to assign terms, computerised free indexing is, for all intents and purposes, the same as natural-language indexing.. Terms found adequate and relevant are assigned as index terms.

2.3.3 Controlled Indexing Language

This is an artificial language that is clearly defined and provides prescribed terms and notation for indexing of documents and retrieval. Many information retrieval applications make extensive use of natural language indexing and controlled language indexing**.** It provides array of terms as authority list for selecting approved terms for representing documents. This system allows for standardisation and consistency among indexers as the choice of subject descriptors must be in tandem with the prescribed guidelines. It involves assigning descriptors that match the concepts of the document as specified from the authority list. Controlled indexing languages are said to be more consistent, making searchers' jobs easier and more efficient. Controlled language indexing, on the other hand, is considered as beneficial in a supporting environment for experienced users since it eliminates the need to negotiate all of the differences inherent in natural language.

We have two kinds of controlled indexing languages, alphabetical indexing languages and classification schemes. The former includes thesauri and subject heading lists. This provides alphabetical subject terms and rules of their usage. Classification schemes are used for assigning notation which helps to categorise subjects within a context of classes for unique identification. It is therefore only approved terms and notations that can be used in information representation and retrieval.

2.4 Summary

Language plays very essential functions in information representation and retrieval. The professional indexer utilises words, phrases and terms to provide appropriate descriptors for representing documents. Information representation is about indexing which relies on three indexing languages to create terms, tags or notations that uniquely identify each document in the collection for retrieval purposes. The three languages discussed are natural indexing language, free indexing language and controlled indexing language. Natural indexing language is the language used in normal communication or ordinary context of the document. They are derivatives of terms and phrases used in the document. They are subject terminologies that potential users from the discipline are conversant with and will be most prone to utilising same as search terms. Free indexing language is not limited to terms and phrases contained in the document but are at liberty to utilise any terms considered suitable, adequate and relevant as index terms or descriptors. Controlled Indexing Language is an artificial language that is prescribed for standardisation and consistency among indexers. It involves assigning descriptors that match the concepts of the document as specified from the authority list and prescribed guidelines. It is also to be noted that the language of information representation must approximate to the language used as search terms for a seamless retrieval dynamics.

Self-Assessment Exercise(s)

1. What do you understand by the concept “Indexing Language”?
   (b.) How relevant is it to information retrieval?
2. Differentiate between controlled indexing language and natural indexing language.
3. What is Free Indexing Language?

2.5 References/Further Reading/Web Resources

National Institute of open schooling (2013). Information Retrieval System: Concern & scope. Retrieval fromhttp://digitalnios.ac.in/topic.php?id=339en5B15

Aina, L.O. (2004). Library and Information Science Text in Africa. Ibadan: Third world

Cleveland, D.B. and Cleveland, A.D. (2001). Introduction to Indexing and Abstracting. Greenwood: libraries unlimited.

Module 3 Information Retrieval Techniques and Models

This module introduces students to the various techniques involved in the retrieval of information. These techniques include Boolean searching, stemming and algorism. The method of evaluation of information retrieval systems, i.e. Precision and Recall were also discussed in this module. Efforts were made to identify and describe the different information retrieval models, which include the Boolean model, vector space model and probabilistic models.

Unit 1: Information Retrieval Techniques

Unit Structure

1.1 Introduction

Today's world is characterised by an influx of information in different media. This explosion in information has further increased the challenge of information retrieval as one will have to sieve through an ocean of information to get the needed information. As such, it has become imperative for individuals, especially Library and Information Science graduates to possess the necessary information retrieval skills. To this end, this unit will introduce students to information retrieval techniques, with specific reference to Boolean, stemming, algorism and rooting systems. The unit will also explain the concept of 'precision and recall' as well identify the functional process of information retrieval.

1.2 Intended Learning Outcomes

By the end of this unit, you will be able to:

- identify the different information retrieval techniques
- differentiate between the different information retrieval techniques
- explain the concept of precision and recall
- highlight the functional process of information retrieval.

1.3 Understanding Information Retrieval Techniques

Today, information can be found in the different media and in different form. Aside storing information in printed document, publishers of information now store information in non-print media like electronic databases, Meta-data, and the internet. The majority of web sites include semi-structured and dynamic material that is intertwined with links and difficult to obtain. Searching the World Wide Web differs greatly from searching static and centralised databases. A variety of query languages have been created, all of which are based on semi-structured data models and are typically represented as labeled graphs. The main issue is figuring out how to incorporate knowledge into information mining algorithms. Although the majority of Web pages are text-oriented, a significant quantity of information is not immediately accessible by typical search methods, therefore documents cannot be retrieved without viewing each one separately (Lasic-lazic, Seljan, & Stancic, 2000). This has generally lead to an influx of information, otherwise known as information explosion. This explosion, coupled with the complexity of many information sources has reawakened the need to adopt some strategies or techniques in retrieving precise information.

Several advanced methods for Web information retrieval or mining are analyzed: 1) Syntax analysis, 2) Metadata-based search using RDF (Resource Description Framework), 3) Knowledge annotation by use of CGs (Conceptual Graphs), 4) KPS: Keyword, Pattern, Sample search techniques, 5) Techniques of obtaining descriptions by Fuzzification and Back-propagation (Lasic-lazic, Seljan, & Stancic, 2000). The problem of choosing proper words and indexation is also stressed out. Consequently, scientists and other scholars have established different information retrieval techniques. Among those techniques is the Boolean search technique, term truncation, stemming, amongst others.

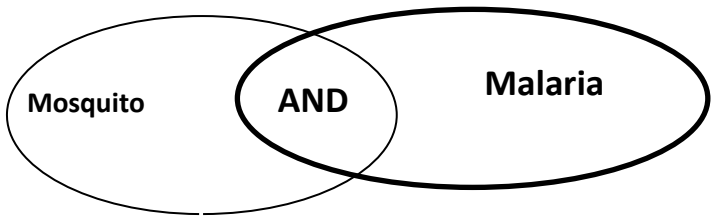
1.3.1 Boolean Searching

The means of specifying a combination of keywords that must be matched for successful retrieval is known as search logic. As a result, most systems search using Boolean search logic. Boolean searching is the search

technique used in querying the internet or a database. The logic is used to connect terms that describe the concepts in the search query. The technique involves combining keywords or phrases in a single search query to retrieve on relevant and desired search results. It can connect terms from both controlled and natural indexing languages, and both. In order to frame a search statement, search logic may link up to 30 or more search phrases together. According to Lasic-lazic, Seljan, and Stancic (2000), using knowledge representation language (KRL) to index Web content is one technique to improve information retrieval. RDF (Resource Description Framework) is one of them, as it is built on XML, which is more machine-readable than human-readable formats. Another option to make representation easier is to utilise a set of straightforward and combinable commands corresponding to first-order logic, such as Conceptual Graphs (CGs), which may be used to index any Web material.

The Boolean technique makes use of the Boolean logic or operators to refine or limit or widen search results. The Boolean logic operators are: AND, OR, and NOT. By applying these operators 'AND', 'OR' and 'NOT', the researcher is able to teach the system the information to include or leave out in a search results. This saves the researcher's time and allows him/her to focus on only the relevant information. The Boolean search technique is often applied when one need more than one word to describe a search problem.

The 'AND' operator is used when one needs to retrieve all information about two or more keywords. For instance, a search term containing the words 'Mosquito and Malaria' will bring results containing both Mosquito and Malaria. The use of the 'AND' operator allows the researcher to combine more than one search term in a single query, which in this case could be Malaria 'AND' Mosquito.



The 'AND' operator could be used if a researcher wants to retrieved search results related to celebrities with sport cars. The researcher could use the search term celebrities 'AND' sport cars. This will fetch out relevant results related to celebrities having dealings with sport cars. In the same vein, a researcher who wants to learn about cloning related to humans could use the 'AND' operator to retrieve only results related to human cloning and not cloning of other animals or object. Such researchers could use the search term: `Cloning 'AND' humans`. If the researcher wants to learn about the ethical issues associated with human

cloning, he or she can then extend the search term by adding the 'AND' operator. Then we will have: Cloning 'AND' humans 'AND' ethics. This will help refine the search results for the researcher.

The 'OR' operator is used to retrieve search results containing any of the search terms. This allows this researcher to still combine search terms in a single query and retrieve results on any of the search terms. In other words, the 'OR' operator allows the researcher to make a broader search as such search could yield more results. For instance, a researcher seeking to retrieve information on transmittable viruses could search for Corona virus or Human papilloma virus. Similarly, a researcher could also use the search term: bioengineering or cloning to learn more of cloning or genetic reformation. Also, a researcher could use the search term: ecology or pollution to retrieve results about ecology (without pollution) or pollution (without ecology) or documents containing either.

The 'NOT' operator is used to narrow the results for a single query. The use of 'Not' helps the researcher to eliminate or exclude (irrelevant) terms or records in a research result. For instance, a researcher interested in searching for information on the 2019 Corona virus could use a search term as "SARS-CoV-2 NOT MERS-CoV". The search will focus on only one type of Corona virus, which is SARS-CoV-2. A researcher interested in reading about Gombe as city in Gombe state could use the search term: `Gombe 'NOT' State` to narrow down his or her search results. This will eliminate results about Gombe as a state.

The search statements are evolved one at a time in online search, with feedback provided at each stage. The searcher enters a search statement, and the computer returns the number of relevant entries. Using this type of search capability, the search strategy may be fine-tuned to get a satisfying result. Logic operators can be used to improve queries (AND, OR, NOT). The basic benefit is that it may be quick owing to automatic indexing, which eliminates the need for human participation. The downside is that replies may be partial or that the amount of results may be excessive, rendering them useless.

1.3.2 Stemming

Stemming involves the practice of producing a morphological variant of a root word. It is generally a language dependent procedure involved in the removal of suffixes from words so that words with the same root match each other. For instance, words such as 'likes', 'likely', 'liked' and 'liking' could be stemmed to the root word 'like'.

Stemming allows the researcher to reduce words to its root without significantly affecting the meaning or purpose of the search. However, in Stemming, there are two main errors that could occur. They are over-

stemming and under-stemming. Over-stemming, otherwise known as false positives, occurs when the stemmer removes more characters or terminations than is necessary, thereby erasing part of the characters that form the morphological root of the word. On the other hand, under-stemming, also known as false negatives, occurs when the stemmer removes parts of the characters or terminations in a word or words and keeps some more complex suffixes in the root word that should ordinarily be stripped. When this occurs, we say the stemmer has under-stemmed.

1.3.3 Algorism

Algorism in information retrieval is concerned with the way data are filtered, indexed and retrieved with high level of precision in a single query. Consequently, information retrieval algorithm can be classified into three main classes. They are retrieval, indexing, and filtering algorithm. The retrieval algorithm, being the main class of algorithm in information retrieval, focus on how information are extracted from a word-based databank.

1.4 Precision and Recall

Precision and recall are two measuring technique in the sphere of information retrieval used to measure how well an information retrieval system retrieves documents that are relevant to the researcher. Typically, in an information system, the total data base and the information to be retrieved could be divided in four main categories. They include relevant and retrieved; relevant and not retrieved; non relevant and retrieved and non-relevant and not retrieved. While the relevant information helps the researcher to answer his/her questions, the non-relevant information does not contribute to the solving of the research problem. Consequently, in a single query, there is the possibility of retrieving relevant information and also the possibility of retrieving non relevant information. Therefore, precision is the total number of relevant retrieved documents (information useful to the researcher) over the total number of retrieved documents from a single query. Hence, precision encompasses the ability to retrieve top ranked relevant documents in a single search. For simplicity sake, the description of precision could be given as:

$$\text{Precision} = \frac{\text{Number of Retrieved Relevant documents}}{\text{Number of Total Retrieved documents}}$$

On the other hand, recall is the ratio of retrieved relevant document to the number of possible relevant documents in the database. Basically, recall measures the extent to which a retrieval system is able to retrieve all relevant documents in a database. The formula for recall is given as:

$$\text{Recall} = \frac{\text{Number of Retrieved Relevant documents}}{\text{Number of Possible Relevant documents}}$$

1.5 Functional Process of Information Retrieval

Information retrieval comprises some functional processes. While several opinions and thoughts have been raised in the past regarding the functional process of information retrieval, it is now widely accepted that the functional process of information retrieval is composed of four main parts namely; item normalisation, selective dissemination of information (i.e. mail), archival document database search, and index database search. These four parts describe the functional process of information storage and retrieval.

1.5.1 Item Normalisation

In the functional process of information retrieval, item normalisation is usually considered the first step in the process since it is at this stage that the format of an incoming item is being normalised in an integrated system. In this stage, the system is able to normalised all external item into a conformable system and also logically restructure all items in the system. The normalisation of data here provide room for the creation of searchable data structure as well as stemming and characterisation of tokens. This ensures standardisation of all items or records in the system as the system could have single formats for all items or allow multiple formats.

1.5.2 Selective Dissemination of Information

This is one of the functional processes of information retrieval characterised by search process, indication of interest by users and the user mail files. At this stage, one is able to match each user need or interest to given or up-coming information in the system and when the user needs is matched with the user, the information is then sent to the user via mail. The information system is able to track a user profile and understand their search interest through their search history and then provide future recommendations to the users in the light of the accumulated information. While this process is primarily used in a text-based environment, it can also be used in a multimedia environment.

1.5.3 Document Database Search

The document database search primarily comprises three segments namely; the document database that comprises all information entered and stored in the database, the search query or search term entered by the researcher, and the search process. At this stage, a query entered by the

user search through the database to retrieve relevant information based on the search query entered. This search is mainly considered retrospective and the user may benefit from the selective dissemination of information feature if he/she is online. Users may be able to filter their search at this stage depending on the interface and feature of the design database.

1.5.4 Index Database Search

This process involves searching through an index database or a collection index to retrieve an earlier saved or stored file for planning or decision making. In this search, information retrieved only direct the user to where he/she can acquire detailed information. The information in the database only points one to where detailed information could retrieve. The index database houses multiple index files. Index files are classified into two namely; public and private index files. The classifications of these files are usually set at the point of entering the data in the database.

Metadata-based search is the second option. Metadata-based search uses meta-descriptions from documents to do searches. Attributes should explain the document's properties and content. There are two types: basic type, which works as a library system and has a predefined set of attributes (e.g., author, title, ISBN), and intelligent type, which can extract information from a semantic network. One important issue is that there is no reliable way to find relevant documents without having to go through each one manually. As a result, the WWW Consortium has introduced RDF as a standard language for machine-readable descriptions of Web resources

Automatic metadata generation - Automatic metadata generation can be used to collect context sensitive metadata, which is subsequently represented using RDF, according to Dewey decimal classification (ref. 5). Automatic classifier is a Java-based object-oriented system that gets HTML documents from a URL, analyses the content, and applies a DDC classification class mark to the document. It compares terms discovered in the document to terms manually designated as DDC hierarchy nodes. As a result, relevant metadata including the document title, keywords, abstract, and word count are generated. Documents with similar subject matter will be grouped together under the same classification mark. RDF is viewed as a "Web of Trust," with each document being well-described and generally recognised. Metadata has been attempted in HTML texts in a variety of ways. The issue is that such a technique is not required.

1.6 Summary

In this unit, we discussed information retrieval techniques. At the beginning of this unit, we introduced the different information retrieval techniques such as Boolean searching, stemming and algorism. We identified the three different Boolean operators as ‘AND’, ‘OR’ and

'NOT'. During this unit, we explained how the different Boolean operators could be employed when searching for information in the internet of in any database.

In this unit, we equally introduced the information retrieval measuring concepts such as precision and recall. It was explained that while precision deals with the total number of relevant retrieved documents over the total number of retrieved documents from a single query, recall focus on the ratio of retrieved relevant document to the number of possible relevant documents in the database.

Self-Assessment Exercise(s)

1. Explain Boolean searching.
2. What is stemming?
3. Differentiate between Precision and recall in information retrieval.

1.7 References/Further Reading/Web Resources

Geegs (2020). Introduction to Stemming. Retrieved from https://www.geeksforgeeks.org/introduction-to-stemming/

Introduction to Information Retrieval Systems. (n.d.). The Information Retrieval Series, 1–25. doi:10.1007/0-306-47031-4_1

Lasic-Lazic, J; Seljan, S. & Stancic, H. (200). Information Retrieval Techniques. Retrieved from: https://www.researchgate.net/publication/242403883 Information Retrieval Techniques

Manning, C.D., Raghavan, P. & Schütze, H. (2008). Introduction to Information Retrieval. London. Cambridge University Press. Retrieved from https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Moral, C., de-Antonio, A., Imbert, R. & Ramírez, J. (2014). A survey of stemming algorithms in information retrieval. *Information Research*, **19**(1) paper 605. Retrieved from http://InformationR.net/ir/19-1/paper605.html

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies. 2 (1): 37–63.Retrieved from https://en.wikipedia.org/wiki/Precision_and_recall

Sammut, C., & Webb, G. I. (Eds.). (2010). Encyclopedia of Machine Learning. doi:10.1007/978-0-387-30164-8 Retrieved from https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_652

Swiftype (n.d). What is Stemming? Retrieved from https://swiftype.com/search-concepts/stemming

Zeugmann, T., Poupart, P., Kennedy, J., Jin, X., Han, J., Saitta, L., … Fürnkranz, J. (2011). Precision and Recall. Encyclopedia of Machine Learning, 781–781. doi:10.1007/978-0-387-30164-8_652

Unit 2: Information Retrieval Models

Unit Structure

2.1 Introduction

In the last unit, you learnt about the different information retrieval techniques, and you were introduced to Boolean search techniques, stemming and algorism. You also learnt about the information retrieval measuring concepts such as precision and recall. In this unit, you shall be introduced to the different information retrieval models. Information retrieval process involves the searching of information needed to meet one's need. While the current explosion of information has made it a cumbersome procedure to retrieve needed information, the knowledge of the different information search techniques and information retrieval models will enhance the easy retrieval of needed information. This unit therefore focuses on the information retrieval models as identified and explained by different information experts. The unit will examine the general models of information retrievals in addition to the major information retrieval models.

2.2 Intended Learning Outcomes

By the end of this unit, you will be able to:

- identify the major information retrieval models
- differentiate between vector space model and Boolean model
- explain Probabilistic Model of information retrieval.

2.3 General Models of Information Retrieval

Information retrieval is an important part of an information seeking process. An information retrieval system comprises some classical models that are based on a combination of uncertain inference and propositional logic. However, while the propositional logic is sufficient for text-based documents, it is normally insufficient for the retrieval of multimedia documents or contents, hence the introduction of models

based on predicate logic. This further implies that models in information retrieval could either be based on propositional logic or predicate logic.

2.3.1 Major Information Retrieval Models

According to Lancaster and Warner (1992), the major information retrieval models are divided into four namely; Boolean model; Vector space model, otherwise called statistical model; probabilistic model; linguistic and knowledge-based models. These models were developed in order to retrieve information seamlessly.

2.3.1.1 Boolean Model

The Boolean model is a model used for current large-scale retrieval systems and in on-line information services. The model, which is based on a set of theories, comprised Boolean algebra and the Boolean logic components (such as 'AND', 'OR' and 'NOT'. While the model is very useful in information retrieval, it has a major drawback in that it does not rank result list of retrieved information. This model is divided into different sub Boolean models such as the standard Boolean model, narrowing and broadening techniques, smart Boolean and extended Boolean models. Each of these sub Boolean models has its own advantages and disadvantages. For instance, the standard Boolean model is easy to implement and is computationally efficient but it is difficult to construct Boolean queries. Furthermore, while the model has a feature of expressiveness and clarity, it is difficult to control output. This therefore reinforces the need for the adoption of other forms of Boolean models like the smart and extended Boolean models.

2.3.1.2Vector Space Model (or Statistical Model)

The vector space model, sometimes referred to as the statistical model, is a type of information retrieval model that ranks documents based on the similarities between the search terms or query and the available documents. The model assigns high value to documents that have high number of the query terms in the document. So, a document can have a few of the query terms in the title but may be ranked higher than other documents that have high number of the query terms in the title. The vector space model builds index by perusing through the documents to identify important terms in the documents. It is through this process that the document identifies important or relevant documents. Consequently, the vector space model thrives on two main assumptions:

1) The more similar a document vector is to a query vector; the more likely it is that the document is relevant to that query.

2) The words used to define the dimensions of the space are orthogonal or independent.

2.3.1.3Probabilistic Model

According to Belkin and Croft (1992), the probabilistic model of information retrieval operates on the principle of ranking documents based on the probability of the documents being relevant to the researcher's query or search terms. The model ranks documents by comparing the documents to the user query(ies). This is denoted by binary vectors ~d and ~q. One of the merits of this model is that it has higher chances of providing users with a ranking of high number of relevant documents. This is beside the fact that queries are easy to formulate in this model as users send queries using natural language and not query language. However, the model does not support the Boolean relation as it has a limited expressive power.

2.3.1.4Linguistic and Knowledge-based Models

The linguistic and knowledge-based model is used to retrieve documents. The model does this by performing a morphological, syntactic and semantic analysis of available documents to determine the relevance of the documents to retrieve. This model is seen by many as the simplest form of document retrieval because the model uses search terms or keywords entered to search the documents keywords in the database. As a result, the model only retrieves documents by picking documents (in the database) with similar keywords entered by the user. This method of information retrieval has often been criticised since there is a good chance of the system missing some relevant documents. One major constraint facing the linguistic approach is it inability to resolve word ambiguities and/or generate relevant synonyms or quasi-synonyms based on the semantic relationships between words. However, the presence of advanced technologies such as experts' systems and other forms of artificial intelligence has helped in the retrieval of relevant documents based on the linguistic and knowledge-based model.

2.4 Summary

In this unit, we highlighted the different information retrieval models and went on to explain the different models. It was mentioned that the different information retrieval models include the Boolean model, vector space or statistical model, probabilistic model and linguistic and knowledge-based models. From the unit, we learnt that the vector space model assigns high value to documents that have high number of the query terms in the document while the linguistic and knowledge-based model use search terms or keywords entered (by the user) to search for document's keywords in the database. It is through this method that relevant documents are determined in the linguistic approach. Also, it was noted that the probabilistic model of information retrieval operates on the

principle of ranking a document based on the probability of the document being relevant to the researcher's query or search terms.

Self-Assessment Exercise(s)

1. List the major information retrieval models.
2. Differentiate between vector space model and Boolean model.
3. What is probabilistic Model of information retrieval?

2.5 References/Further Reading/Web Resources

Information Retrieval Models (n.d). General Model of Information Retrieval Retrieved from https://aspoerri.comminfo.rutgers.edu/InfoCrystal/Ch_2.html

Lashkari, A., Mahdavi, F., & Ghomi, V. (2009). A Boolean Model in Information Retrieval for Search Engines. IEEE International Conference on Information Management and Engineering, 385-389.

Saini, B., Singh, V., & Kumar, S. (2014). Information Retrieval Models and Searching Methodologies: Survey. International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE), 1 (2), 57 – 62. Retrieved from https://www.researchgate.net/publication/274837522_Information_Retrieval_Models_and_Searching_Methodologies_Survey

Module 4 Multimedia Information Retrieval Systems, Retrieval on the World Wide Web and Digital Libraries

This module introduces students to the concepts of:

i. Multimedia,
ii. The World Wide Web and
iii. Digital Libraries.

The module extensively discusses the need for multimedia information retrieval systems, the basic multimedia search technologies and multimedia information retrieval systems, with specific reference to the context-based, content-based and concept-based multimedia information retrieval systems as well as the imperatives of Digital Libraries. Furthermore, the module addressed issues relating to the retrieval of information on the World Wide Web, like the search engines, indexing of web documents, taxonomy for web search tools and web services amongst others.

Unit 1 Multimedia Information Retrieval Systems
Unit 2 Information Retrieval on the World Wide Web
Unit 3 Digital Libraries and Retrieval Imperatives
Unit 4 Evaluation of Information Retrieval Techniques and Processes

Unit 1: Multimedia Information Retrieval Systems

Unit Structure

1.1 Introduction
1.2 Intended Learning Outcomes
1.3 Concept of Multimedia
  1.3.1 Multimedia Database Management System Architecture
1.4 What is Multimedia Information Retrieval System?
1.5 Need for Multimedia Information Retrieval Systems
1.6 Multimedia Information Retrieval System
  1.6.1 Content-based Multimedia Information Retrieval
  1.6.2 Context-based Multimedia Information Retrieval
  1.6.3 Concept-based Multimedia Information Retrieval
1.7 Basic Multimedia Search Technologies
  1.7.1 Metadata Driven Search
  1.7.2 Piggy-Back Text
  1.7.3 Automated Image Annotation
  1.7.4 Fingerprinting
1.8 Summary
1.9 References/Further Reading/Web Resources

1.1 Introduction

Advancement in technologies has resulted in the use of different media for the processing and storage of information in different format. This development has allowed for the use of information in the form of images, videos, music amongst others, for planning and decision making. This is what gave rise to the concept of multimedia, which means two or more media or modes of consuming information which could be in form of reading, watching and listening. To this end, this unit will seek to give light to the concept of multimedia, describe the basic multimedia search technologies, elucidate on image indexing and retrieval, image representation and content based retrieval. Overall, this unit will attempt to describe multimedia information retrieval systems.

1.2 Intended Learning Outcomes

By the end of this unit, you will be able to:

- defined the concept of multimedia
- identify and explain the basic multimedia search technologies
- describe the multimedia content-based information retrieval
- identify the approaches to Image Indexing and Retrieval.

1.3 Concept of Multimedia

The ability to gather, store, edit, integrate, and query information given in several formats, such as text, graphics, audio, video, and photographs, is referred to as multimedia. Multimedia is not a technology in the traditional sense. It's more of a notion that represents a collection of technologies that work together to benefit the end user. Multimedia databases are becoming increasingly common in today's computerised world because they allow users to effortlessly manage various sorts of complicated data modeled after the actual world. As a result, users can always find new data kinds in a multimedia database, such as:

Image Data: This is a type of data that is typically found in multimedia databases, and its uses include simple figures, icons, and medical images such as X-rays.)

Video Data: These are video files that have grown increasingly significant as new technologies such as video distribution have emerged. The ability to store a home video on a computer is now more convenient than ever.

Audio Data is used to both store and deliver music, noises, and speech.

Document Data: These are traditional text files that contain information in the form of text. These files are still in use however, their storage capacities have increased. Multimedia objects differ from standard text or numerical documents in that they typically need a significant capacity of memory and disk storage. In addition, the processes performed on multimedia files differ (e.g., displaying a picture or playing a video clip is different from displaying a text paragraph). A multimedia database management system should be able to provide an environment that is conducive to the use and management of multimedia items. A multimedia database management system must be able to support the following basic operations in addition to the standard tasks of a database management system:

- Handles image, voice, graphics and other multimedia data types
- Handles a large number of multimedia objects
- Provides a high-performance and cost-effective storage management scheme
- Provides efficient storage and retrieval of multimedia objects and a multimedia database management system with a dynamic architecture

Multimedia is the field of computing which deals with the integration of text, graphics, videos, animation, images and any other type of media files for processing, transmission and storage (Ftsm, 2018). Basically, multimedia connotes that computer information can be represented through different media such as text, images, videos etc. It uses multiple forms of information content and information processing. The basic elements of multimedia include videos, images, text, audio, animation and graphics. Examples of multimedia include audio slideshow and video podcast. Contrary to traditional media like printed materials, multimedia combines information in diverse kinds of formats. Multimedia can be used to store information and its contents viewed through technologies like laptop, smart phones, and smart television.

Moreover, multimedia can be broadly categorised into two broad categories, namely:

1. Linear category: In this category, there is smooth progression of content and it does not require any form navigation control from the viewer to make progress. An example of this is when you watch a blockbuster movie.

2. Non Linear category: This category makes progression through the interactivity of the participants involved. For instance, the use of video games for entertainment requires the direct involvement of the actors.

Due to the increasing importance of multimedia information and technology in the today's society, multimedia is now being applied in different industries and places such as engineering, mathematical and scientific research, education, entertainment, social work, creative industries, journalism, and medicine, amongst others.

1.3.1 Multimedia Database Management System Architecture

According to Volovici, Breazu, Mitea, and Morariu (2010), the architecture of a multimedia database usually consists of three layers. These are the following:

The user interface layer, the conceptual layer, which includes the object and composition layers, and the internal layer, which eventually becomes the storage layer.

Tasks to be handled at the interface level include object browsing, query processing, and object composition interaction. The user can use object browsing to identify multimedia resource entities that can be reused. The user specifies information through text-based or visual queries a set of conditions to the resource's properties and returns a list of candidate objects the appropriate objects are then reused. A text-based query language cannot successfully locate multimedia resources, unlike text or numerical data. The method that allows the user to effectively find reusable multimedia elements, such as photos, sound, video, and other forms, is the focus of content-based information retrieval research. The database interface should assist the user in composing or decomposing multimedia documents after a successful retrieval. To handle objects, the second layer collaborates with the interface layer. To express distinct relations among objects, object composition often involves a number of links, such as association links, similarity links, and inheritance links of an object-oriented system. The database graphical user interface or a number of application program interface (API) functions are used to specify these linkages.

Clustering and indexing are two performance-related challenges addressed by the storage management layer. Clustering is the process of physically organising multimedia material on a hard disk such that when it is retrieved, the system can efficiently read huge binary data (Sifaoui, Abdelkrim, Alouane, & Benrejeb, 2009). Typically, retrieval performance must ensure some level of service quality as well as multimedia synchronisation. To find the physical address of a multimedia object, indexing requires a rapid finding technique. A complicated data or file structure is sometimes used in the scheme.

1.4 What is Multimedia Information Retrieval Systems?

This refers to the technology used in the searching for and finding of multimedia documents (usually in the form of text, video, images, music etc.) stored in a database or in the internet. According to Kambau and Hasibuan (2017), multimedia information retrieval system is an information retrieval system which seeks to extract information from multimedia data sources. This retrieval system allows one to retrieve and utilise text, images, audio and video after a search process. It is imperative to add that while multimedia information retrieval system refers to the technology or software used in the extraction of multimedia contents, multimedia information retrieval refers to the process of searching and retrieving multimedia documents. The retrieval of multimedia contents typically begins with a query (which could be in the form of text or an image) for the purpose of retrieving similar contents. An overview of multimedia information retrieval systems is presented in figure 1.1.

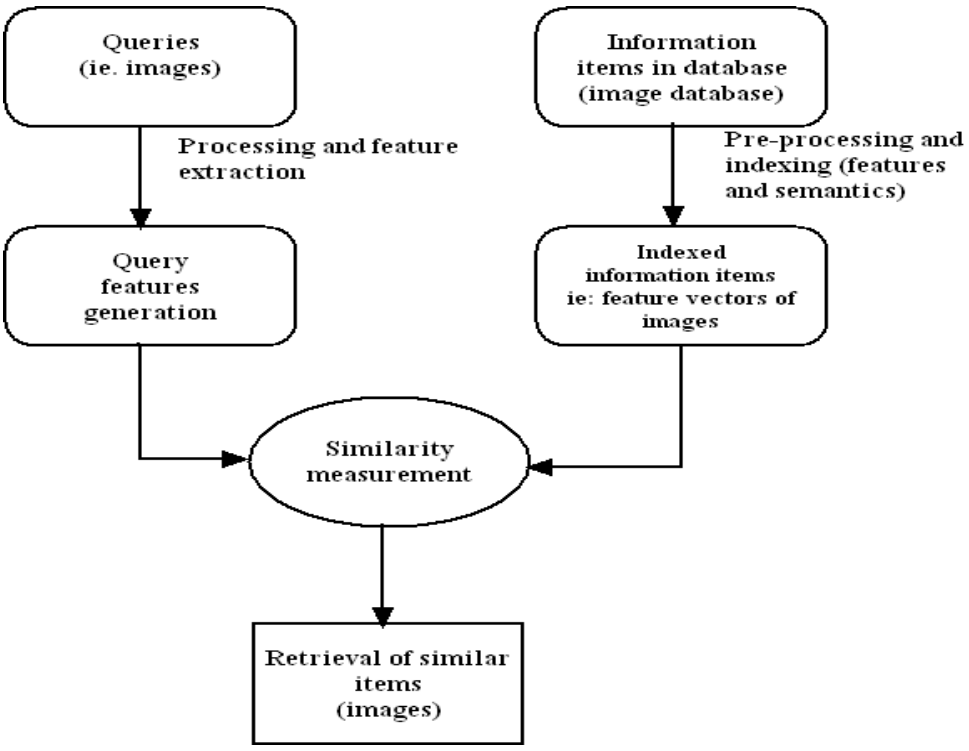


Figure 1.1: An Overview of MIRS, Adopted from COMP9314 Advanced Database Systems – Lecture 5 – Slide 16 – J Zhang

1.5 Need for Multimedia Information Retrieval Systems

The rapid growth in multimedia related contents in different online and offline platforms, coupled with the complexities associated with many multimedia contents has reawakened the need for effective multimedia

information retrieval systems. This need is further reinforced by the diversity of multimedia related content types which can be found in variety of information environments. As a result, computer and information experts have attempted to develop different multimedia information retrieval systems. It should however be noted that there is no single ideal solution for information retrieval as there are different multimedia content types available under different environment. Consequently, some of the reasons of developing multimedia information retrieval systems include:

1. **Multimedia Data Types:** One major issue surrounding multimedia information is that they come in a variety of data types such as text, audio, video, animation, graphics etc. this therefore justify the need to have an effective retrieval system for the purpose of retrieving multimedia information.
2. **Alpha-numeric Nature of Multimedia Data:** The regular database management systems may not effectively handle multimedia data because its characteristics and requirements are significantly different from regular alpha-numeric data; hence, information managers result to developing a separate retrieval system for multimedia data.
3. **Vast Multimedia Data:** The development of a multimedia information retrieval system allows for the capturing, storage and retrieval of relevant vast multimedia data.
4. **New Media:** In recent years, the world has experienced the introduction of new media in addition to the existing media. This has led to a shift from content based image retrieval to the retrieval of other forms of multimedia information.

1.6 Multimedia Information Retrieval System

Multimedia information retrieval system focuses on the techniques in retrieving multimedia related content. The system primarily deals with the bridging the gaps and easily translating multimedia content from their low-level to high level concepts. Consequently, the retrieval of multimedia information could be divided into three parts namely; content-based multimedia retrieval; context-based multimedia information retrieval and concept-based multimedia information retrieval.

1.6.1 Content-based Multimedia Information Retrieval

Content-based multimedia information retrieval is used to retrieve image, audio and video content. Hence, it is divided into three categories namely; content-based image retrieval, content-based audio retrieval and content-based video retrieval. The content-based image retrieval makes use of elements such as colour, shapes and texture to retrieve image content from

the database while the content-based video retrieval is an extension of the image retrieval but applies motion features to the moving image. The video retrieval system uses video parsing, content analysis features and abstraction to retrieve relevant videos from the database. The audio retrieval system makes use of audio signals with acoustic or semantic features. In this case, acoustic feature could include the pitch, cepstrum and loudness while semantic feature could include rhythm, timbre and events.

1.6.2 Context-based Multimedia Information Retrieval

This type of retrieval combines the user's context and knowledge of information retrieval with multimedia search technologies to retrieve needed contents. The implication of this is that the user search behaviour and characteristics will influence the search outcome. Therefore, context-based multimedia information retrieval may include several dimensions such as time, location, user, current task.

1.6.3 Concept-based Multimedia Information Retrieval

This category of multimedia information retrieval focuses on using manually built thesauri or by extraction of latent word relationship and concept from the corpus to retrieve relevant multimedia contents. It uses a classifier to select multimedia data from a pool of data or from a database when a query is made. It became necessary to develop the concept-based multimedia information retrieval due the inability of the content-based and context-based multimedia information retrieval to describe semantic visual features or semantic audio features.

1.7 Basic Multimedia Search Technologies

Among the basic multimedia search technologies or retrieval method are the metadata driven search retrieval, piggy-back text search, automated image annotation, fingerprinting amongst others.

1.7.1 Metadata Driven Search

The metadata provide information that is used to describe, manage, store, create and retrieve multimedia information. Metadata could be used in multimedia search queries and it comes in layers which enable one to submit a piece of information about an image, video, audio or text in a database. Multimedia metadata also play a major role in the indexing, classification and location of multimedia contents through the development of appropriate schemas. For example, the World Wide Web consortium uses the XML schema language to document and retrieve contents. Furthermore, several standards for bibliographic metadata have

been developed for easy management of stored documents. For instance, the Machine Readable Catalogue 21 (MARC 21), MPEG-7 and the Dublin Core are among the most used metadata standards.

Fuzzification: It is a well-known truth that retrieving necessary information from the Internet can be a time-consuming operation. The fundamental reason is that the content on the Internet is poorly classified. Of course, many Web search engines employ specific robots in their search for new Web sites, and when a page is discovered, it is assigned to the appropriate categorisation category based on the classification method employed by the web search engine. Metadata classification, on the other hand, can be added to web objects. This means that the classification duty is delegated to people who design and maintain web elements in part.

M. Marchiori (ref. 6) suggests the fuzzification method as an advanced solution for Web classification. According to him, existing Web metadata sets have properties allocated to objects, but they are either present or absent. Instead, he proposes that attributes be fuzzified, meaning that each attribute be assigned a "fuzzy estimate of its relevance for the Web object, specifically a value ranging from 0 to 1." This indicates that an attribute with a value of 0 is irrelevant to the Web object to which it belongs. The attribute's relevance to the Web object is 40% if the value is 0.4. Because classification is only a rough approximation, fuzzification provides for more freedom within a predefined 9 classification system, allowing for more detailed ranking while keeping the core collection of ideas limited.

Back-propagation so far, we've seen how to determine whether a Web object is relevant to a search query by looking at the information it contains. Because the Web is a dynamic medium intertwined with hyperlinks, it is a typical occurrence for one Web object to point to another or even multiple Web objects. This leads us to the challenge of determining the significance of an object that is directed to by another object, as explained in M. Marchiori's paradigm (ref. 6). Assume that a Web object O has the metadata A: v, indicating that the attribute A has the fuzzy value v. We can "back-propagate" this metadata information from O to O' if there is another Web object O' with a hyperlink to O.

Because we can simply activate the hyperlink, we have the impression that the information contained in O (classified as A: v) is also accessible from O'. However, it is not as if we are already in O because in order to access the information, we must first activate the hyperlink and then wait for O to be retrieved. As a result, O"s relevance to the property A is not the same as O'(v), but it has faded in some ways, because the information in O is only indirectly accessible from O', not directly present within. To solve this problem, multiply the attribute's value v by a "fading factor" f (with 0<f) and fade it.

1.7.2 Piggy-Back Text Search

Simply put, this kind of multimedia search allows for automated processes to create text surrogates for multimedia. This category of multimedia search focus on the retrieval of multimedia contents, especially videos by extracting its search strings from Teledex, closed caption and subtitles. While the speech recognition and optical character recognition is used for audio and text retrieval respectively. However, information about some audio sounds like music could be converted into text like the MIDI files that also has some note representation. The piggy-back text search employs a variety of text matching techniques to search, edit and manage multimedia contents.

1.7.3 Automated Image Annotation

This is an image retrieval system where metadata are automatically assigned to digital image by computer systems using keywords or capturing to locate, organise and retrieve relevant images from a database. This process is sometimes referred to as automatic image tagging and can be likened to a multi-class image classification with a very large number of classes. The automated image annotation extract requested images by conducting an image analysis through machine learning technology. This retrieval system holds more advantages than the regular content-based image retrieval as queries can be more naturally specified by the user. Queries do not have to rely heavily on colour and texture as it is with the content-based image search.

1.7.4 Fingerprinting

This is an important retrieval method in multimedia information retrieval. This retrieval method uses a prescribed algorithm to identify and retrieve a specific multimedia object based solely on its content. Fingerprints can be applied to both audio and image. In fact, both set of media has the same fingerprints requirements, i.e., they are small, reliable, fast, and have unique location of different records in the database even under degradation or little change. However, the primary distinction between audio and image fingerprinting is that while the audio is a one-dimensional air pressure function of time, the image fingerprint contains static two-dimensional colour distributions (Rüger, 2010). The fingerprint retrieval system is an ideal system against degradation or deliberate modification. Some of the techniques involved in multimedia fingerprinting include computing of salient points, extracting features from vicinity, making invariant under rotation, quantising, indexing or text search engines, enforcing of spatial constraints after retrieval etc.

1.8 Summary

In this unit, we examined the concept of multimedia and multimedia information retrieval system. We explained the linear and nonlinear category of multimedia and it was noted that the use of multimedia connotes that computer information can be represented through different media such as text, images, audio, videos etc. If you recall, we defined multimedia information retrieval system as an information retrieval system which seeks to extract information from multimedia data sources. We then went further to explain the need for a multimedia information retrieval system. Also in this unit, we identified and explained the three parts of multimedia information retrieval system which include content-based multimedia retrieval; context-based multimedia information retrieval and concept-based multimedia information retrieval. Finally, the different multimedia retrieval techniques like the metadata driven search, piggy-back text search, automated image annotation and fingerprinting were discussed to underline the method of retrieval of multimedia information.

Self-Assessment Exercise(s)

1. What is multimedia information retrieval system?
2. Why do we need a multimedia information retrieval system?
3. Explain content-based multimedia information retrieval system.
4. Discuss basic multimedia search technologies.

1.9 References/Further Reading/Web Resources

Nordbotten, J.C. (2008). Multimedia Information Retrieval Systems. Retrieved from http://nordbotten.com/ADM/ADM_book/MIRS-frame.htm

Rüger, S. (2009). Multimedia Information Retrieval. Synthesis Lectures on Information Concepts, Retrieval, and Services, 1(1), 1–171. doi:10.2200/s00244ed1v01y200912icr010

MMGD0101 (n.d) Introduction to Multimedia. Retrieved from https://www.ftms.edu.my/images/Document/MMGD0101%20-%20Introduction%20to%20Multimedia/MMGD0101%20chapter%201.pdf

Marshall, D. (2001). What is Multimedia? Retrieved from https://users.cs.cf.ac.uk/Dave.Marshall/Multimedia/node10.html

Martin, P., Eklund, P., Embedding knowledge in Web documents, Griffith University, Australia, http://decweb.ethz.ch/WWW8/data/2145/html/bindex.htm, April 24, 2000

Sifaoui, A. A.; Abdelkrim, S.; Alouane, M.; & Benrejeb, (2009). On New RBF Neural Network Construction Algorithm for Classification, *Journal of Studies in Informatics and Control*, Vol. 18, No. 2,

Volovici, D.; Breazu, M.; Adi-Cristina Mitea, A. C.; &D.L. (2010). Digital Information Retrieval Studies in Informatics and Control · June 2010, DOI: 10.24846/v19i2y201009

Unit 2: Information Retrieval on the World Wide Web

Unit Structure

2.1 Introduction

The Web or the World Wide Web is one of the remarkable inventions in modern era. Sometimes called the Internet, the web holds together a pool of information from different sources (web addresses). In fact, the existence of the World Wide Web is one of the reasons the World is currently experiencing an influx of information (information explosion). Due to the large amount of information in the World Wide Web, information seekers are often faced with the challenge sieving through a pool of information to retrieve the relevant information from the pool. For instance, a single query in the Web could generate thousands of results comprising of both relevant and irrelevant information. As a result, individuals, especially information experts, must be skilled in the process of retrieving information from the World Wide Web. Consequently, we will be looking at information retrieval on the World Wide Web in this unit. This unit will introduce you to the World Wide Web and its related technologies such as search engines and other web search tools. In this unit, we will also discuss the indexing of web documents and web search services

2.2 Intended Learning Outcomes

By the end of this unit, you will be able to:

- identify the different search engines in the World Wide Web
- understand the indexing of documents in the World Wide Web
- explain the different web search services
- discuss the taxonomy of web search tools.

2.3 Introduction to the World Wide Web

The World Wide Web is an interconnected information system which gives universal access to wide range of documents with a uniform resource identifier. The Web contains web resources which are hyperlinked by means of hypertext or hypermedia and the linked documents connect related information which allows users to access them directly by clicking on the linked word or phrase. Being an information system, the web contains information resources in different media which include images, audio, video, animations, text and other forms of information usually with hypermedia feature. Invented in 1989 by Sir Tim Berners-Lee, the web operates with some software applications and standard protocols like the Hyper-Text Transfer Protocol (HTTP), Hypertext Mark-up Language (HTML) and Web browser. The hypertext mark-up language for instance, is used to format Web Pages which usually contains resources with common theme and domain name. The HTTP on the other hand, is an internet protocol that allows the retrieval and transfer of linked resources across the web. These web technologies aid the retrieval and transmission of information over the internet.

The World Wide Web first gained the acceptance of the public in September 1993 after the launching of Mosaic, a web browser developed in the United States of America by Marc Andreessen. Since then, several other web browsers have been developed such as the Internet Explorer developed by Microsoft Corporation; Safari, developed by Apple; Firefox, developed by Mozilla; Chrome, developed by Google; and more recently, Edge, developed by Microsoft.

2.4 Web Terminologies and Technologies

Users can now access a wide range of textual information resources. OPAC, Gateways, Portals, Subject Portals, Electronic Journals, Online Databases, Subject Directories, and Search Engines are examples of web-based reference resources and services for accessing information from libraries (Borgman, 2000, Bhatnagar, 2005) The kind of information covered by these resources are quite similar. It can be difficult to tell the difference between some of them at times. Because digital libraries are expected to have a good collection of these materials, it is necessary to describe a few of them in this study for a better understanding.

Library OPAC - On Line Public Access Catalogues, is an essential part of many digital library collections and retrieval. It enables users to search for and retrieve bibliographic entries from a library's collections. In addition to traditional bibliographic records, several OPACs now allow access to electronic resources and databases.

Library Gateways - Library Gateways - A gateway is defined as a facility that enables easy accessibility to network based resources in a given subject area. Through a resource database and indexes that can be searched through a web-based interface, gateways give a simple search function as well as a much-improved service. The information provided by gateways is manually cataloged. A wide range of topics are covered by gateways. The following are some well-known gateways:

- Internet Public Library (IPL),
- Bulletin Board for Libraries (BUBL),
- National Information Services and Systems (NISS),
- Etc. **(Bhatnagar, 2005)**

Library Portals - In the library community, portals may be defined as gateway to the services library provides to the users where the process is achieved through seamless integration of existing services by using binding agents such as customisation and authentication services, search protocols such as loan protocols, ISO10161, and e-commerce (Bhatnagar, 2005). As a result, consumers can get a customised solution that allows them to access content from both print and electronic platforms

Subject Portals - Subject Portals - Information professionals are increasingly using their abilities to help organise the vast amount of information available on the Internet. The emergence of subject-specific web search engines known as subject portals, where appraisal of information covered is a primary priority, is an excellent example of their influence.

Electronic Journals Electronic journals make up a significant portion of academic libraries' collections. Many journals are now available electronically; some are full text, while others merely contain bibliographic data and an abstract. Electronic journals have the advantage of being regularly updated and quick to access. The negative is that it is relatively easy to break copyright laws. They're available in a variety of formats, including bitmaps, PostScript, PDF, ASCII, SGML, and HTML. Users may receive library services via CD-ROM, via email, or via the web.

Online Databases - Online Databases - These are enormous collections of machine-readable data kept by commercial organisations and accessed via communication lines. They are subscribed to by many libraries for convenient access to and use of current information. The downside is that only bibliographic information is provided, rather than the entire text. When the system is offline for whatever reason, the information cannot be accessible'.

Search Engines - Search Engines - Search Engines are massive databases of web page files that are automatically compiled by machines, whereas subject directories are compiled and maintained by humans. Every page of a website is indexed by a search engine, and subject directories only link to homepages. An information retrieval (IR) system is commonly referred to as a search engine. A search engine is a piece of software that searches a database of electronic documents for citations, documents, or information that matches or responds to a user's query. Text documents, extracted information from text, photos, and audio are all examples of materials that can be retrieved.)

Subject Directories - Subject directories are different from search engines in that search engines are filled by robots that find and index sites, whereas subject directories are populated by humans who make editorial choices. General, academic, commercial, and portal subject directories are essentially index home pages of websites. Relevance, effectiveness, and a high level of content quality are among the advantages. The lack of depth in their coverage of the issues is one of their flaws.

E-Book - The term "electronic book" or "e-book," invented by Van Dam of Brown University in the 1960s, is now widely used. An e-Book, according to the Harrod Librarians Glossary, is a broad term for electronic products and multimedia publications that may be accessed directly from the Web or in physical form via optical disk.

Online Database - There are different types of online databases which includes, **c**omplete indexing and abstraction databases, text databases, reference databases, and statistics databases. *Indexing and abstract databases* provide bibliographic information about journals included summary of articles. For example, SCOPUS, LISA, LIST, etc., are full *text databases* containing an organised collection of information on a particular multidisciplinary topic or theme, example Science Direct. *References databases* are the many Dictionaries, Almanacs and Encyclopedias, that are available on the Internet in electronic format. *Statistical databases* contain useful numerical data for the mass.

Use-Net - The Use-net is a global electronic bulletin board where millions of individuals exchange public knowledge on every subject imaginable. It's also known as "Net-news," and it's made up of thousands of newsgroups covering a wide range of topics. Use-net/news-feed information can be read using a variety of news-reader software programs. Unlike e-mail communications, Use-net news-group messages are not saved on (users') computers unless they are explicitly and consciously saved there (Martin, Murugiah, & Nandhini, 2016, pp. 82-84).

UnCover – In libraries in developed countries, UnCover is an online periodical article distribution service and a current awareness alerting tool for libraries in developed nations. Its database has about 18000 English language publications and is still increasing. A simple online order system makes over eight million articles available. Every day, 5,000 new citations are added. Articles emerge in UnCover at the same time as the issue of the periodical is delivered to the library or local newsstand, making it the most up-to-date index available. It is extremely useful for librarians, presenters, students, and the general public who require current information given rapidly. It is completely free to search the UnCover database. Users must only pay for the items that they have ordered (Martin, Murugiah, & Nandhini, 2016). It is impossible to discuss the retrieval of information in the World Wide Web without first identifying and discussing some of the web terminologies and technologies. Being an information system, the World Wide Web does not function in isolation. The web brings together several related technologies in order to execute tasks and to function effectively. For the purpose of this study, we refer to web technologies as tools and techniques available over the internet which allows for communication among different devices. These technologies and techniques include standard protocols like Gopher, the HTML and HTTP, as well as technologies like web servers, web browsers, JavaScript and CSS.

a. **Gopher:** Gopher is a distributed document retrieval system that searches the internet for content using menu-driven software. The gopher system organises information into a hierarchy of subject-oriented menus that link to text files, other menus, binary files, file transfer protocol (FTP) (a client–server protocol that allows users to transfer data from one computer to another.) and telnet sites.

b. **Hypertext Mark-up Language (HTML):** This is one of the standard protocols used in creating documents in the web. It is a document format established by Tim Berners-Lee of the European Particle Physics Laboratory that employs tags to indicate the various portions or aspects of a text, according to Fecko (1997). The HTML provides a standard format for labeling web document's structure for transmission through hyperlinks. The hypertext in the standard allows for easy navigation across different pages in the web.

c. **Hypertext Transfer Protocol (HTTP)/ Hypertext Transfer Protocol Secure (HTTPS):** This is a standard protocol used for the exchange or transfer of data from one website to another. The protocol allows for exchange of information between one's browser and the website. The secure version (HTTPS) of this protocol ensures a safe connection to a website.

d. **JavaScript and CSS:** The CSS stands for cascading style sheet and it is used to add style to a website with regards to color, background, shadows, size, spacing etc. Some of the CSS frameworks used include Bootstrap, Bulma and Tailwind CSS. On the other hand, the JavaScript is one of the core components of the World Wide Web alongside HTML and CSS. It is a language that allows one to add interactive behavior to web pages, build web and mobile apps, build web servers and interact with Web Pages.

e. **Web Servers:** This is a computer program that runs websites and distributes Web Pages as they are requested. The web server primarily store, process and deliver Web Pages to users via the web browser. The program uses the HTTP to respond to users' requests. While some web servers could come in form of hardware, others are in cloud form. Many websites often use the cloud servers because of security and cost effectiveness.

f. **Web Browser:** This is a computer program or application used to interact or access information on the web. The web browser sometimes referred to as 'browser', helps to locate Web Pages. Example of web browser includes Chrome, Firefox, Edge, Safari and Opera.

g. **Uniform Resource Locator (URL):** This refers to a unique global address given to a document on the web. The URL often point to a resource on the web. The resource could include a document, webpage, image, video and more. More often than not, a URL gives information about the type protocol (whether secured or not), domain and sub domain names of the website or item.

2.5 Search Engines

The search engine is a web-based tool or software system that is used to locate and retrieve information on the web. They are primarily used to conduct web searches in a systematic way using the supplied web search query. Examples of search engines include Google, Microsoft Bing, Yahoo, Ask.com, Duck Duck Go, Alta Vista, Lycos etc. However, it is important to note that all of the above examples of search engines falls under the general purpose search engines while search engines like mamma.com, dogpile.com and Meta Crawler can be categorised as Meta search engines. Meta search engines collect results or documents from several search engines in order to provide relevant documents. In summary, Meta search engines ranks results from other search engines.

When a query is made on the search engine, the search engine searches the internet for content by looking over the code/content for each URL.

They then store and organise the content found during the crawling process. After which, they rank the document according to their relevance to the research query. Consequently, it can be said that the search engine performs three main tasks when a query is entered. The tasks include crawling, indexing and ranking of results/documents. Each of these tasks is essential in web information retrieval process on the World Wide Web.

2.5.1 Web Crawling/Resource Discovery

The web crawlers, otherwise referred to as spiders, are internet robots or bots that searches through the World Wide Web for documents or links based on the search query entered. The crawler works together with the search engines to retrieve and index information from the web based on their relevance to the query. In recent time different search engines have come up with their own search engine crawlers to enhance the retrieval of highly relevant and increased results. While there are commercial web crawlers like Swiftbot, there also exist, numerous open source crawlers like Frontera, GNU Wget, GRUB etc. The advancement in technology in recent time has increased the overall functionality of web crawlers as they are able to validate hyperlinks and HTML code. This is beside their web scraping and data-driven programming capabilities. Figure 2.1 shows the architecture of a web crawler.

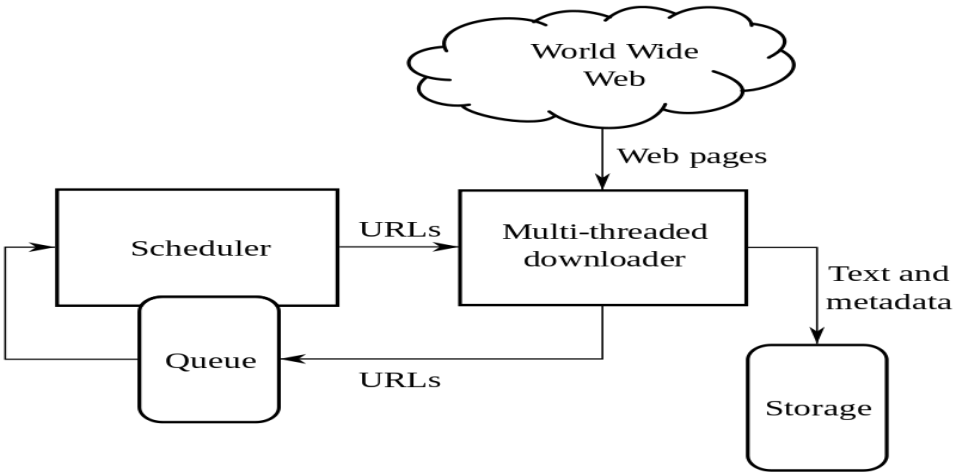


Figure 2.1: Architecture of Web Crawler (developed by Vector Version)

2.5.2 Web Search Engine Indexing

The web search engine indexing comes up after the web crawler must have executed its task. Search engine indexing use the stored information about Web Pages to return relevant and high-quality result. The indexing process occurs when the search engine organise information before presenting the result to the interface. Unlike traditional indexing were relevant documents are identified and retrieved, search engines use the inverted or reverse indexing to rank relevant documents. The reverse or inverted indexing is when text elements are compiled along with pointers

to the documents which contain those elements. The overall goal of search engine indexing is to facilitate fast and accurate information retrieval.

2.5.3 Search Result Ranking

Simply put, search result ranking is the process of sorting search results according to their relevance to the query. The sorting in this case is usually done from the most relevant to the least relevant result.

2.6 Classification of Web Search Tools

Web search tools aid the retrieval of information from the World Wide Web. These tools use robots to index documents in the web before presenting a result via an interface. The search tool house different components and could be classified into type 1 search tool and type 2 search tools (Gudivada, Raghavan, Grosky & Kasanagottu, 1999). Gudivada *et al* (1999) distinguished between the type 1 and type 2 search tools using different dimensions such as the indexing techniques, strategies for query-document matching, methods of web navigation, query language or specification scheme amongst others.

One major distinction between the type 1 and type 2 search tools is that while the type 1 search tools completely hide the organisation and content of the index from the user, types 2 do not. In fact, a well-known characteristics of type 2 search tools is that its hierarchically organise subject catalogue or directory of the web and makes it visible to end-users as they search (Gudivada, *et al*; 1999). Examples of the type 1 search tools include Alta Vista, HotBot, Excite and Lycos while examples of the type 2 search tools include Magellan, Yahoo, and WWW Virtual Library.

2.7 Web Search Services

The web search services include activities by computer programmes to extend users' query from one search engine to others. The search services send out queries to multiple search engines and information sources simultaneously in order to retrieve comprehensive results. For instance, the MetaCrawler is an example of search services. When in action, the MetaCrawler broadcast user's queries to several separate search engines. Some of the search engines often include Infoseek, WebCrawler, OpenText, Lycos, Excite, Yahoo, AltaVista and Galaxy. Furthermore, aside broadcasting queries to other search engines, search services also merge results retrieved, exclude redundant or duplicate information and present the final output as hypertext markup language page with clickable universal identifier. The search services primarily help to simplify web search by providing layer of abstraction to the user over several search tools.

2.8 Web Search Strategies

In searching for information on the web, there are two main classes namely; the known item search and the unknown item search. While the known item search is conducted when the information seeker knows the item s/he is looking for, the unknown item search is carried out when the information seeker is unaware about the availability of the information he/she seek. For a known item search, the user may have some information such as the title of the work, the name of the author, the International Standard Book Number (ISBN) etc. However, in web searching, there are different approaches or search strategies to ensure information are easily retrieved from the web. Among the search strategies are:

Keyword and Phrase Search: The keyword search strategy is the simplest and oldest form of web search strategy. This search method is done by entering a keyword or phrase in the system and the system then conduct an inverted file (index search) for each keyword. The keywords or phrase used in this search strategy is often retrieved from the subject heading list. The phrase search matches the phrase entered to other keywords in the system to retrieve relevant results. If the researcher has need to combine multiple keywords or phrases, then the researcher can introduce the Boolean operators or logic (AND, OR, NOT). To conduct a keyword or phrase search, the researcher must enter the keywords or phrases in the system using the keyboard or select the keywords from an index file or vocabulary control tools like subject headings list.

Boolean Search: This is the search technique that combines search terms or phrases using certain logic or operators known widely as Boolean logic. The Boolean logic helps to narrow or expand a search by introducing some basic operators such as AND, OR, NOT. While the AND operator allows the research to combine search terms of keywords and retrieve wider results, the NOT operator helps to narrow down the search results by eliminating related terms from the result. The `NOT` operator seek to retrieve only needed information while irrelevant information and `noise` are excluded from the search results.

Truncation: In truncation, a researcher is able to search for different forms of words by searching using the lexeme or root word. In this case, the researcher identifies a word having the same common root words and use the root words to search for words with similar root word. For instance, the word `astro` (relating to star or celestial objects) could be used to search for words such as astronaut, astronomy, and astrophysics. Similarly, the `acri` (relating to bitter) could be used to search for meaning of words such as acrid, acrimony, acridity etc. However, it should be noted that different forms of truncation exist. For instance, we have the

right truncation, left truncation and making letters in the middle. The earlier examples of `astro` and `acri` are examples of right truncation. The left-turn truncation is used when one need to retrieve all words having the same characters at the right. Example includes words ending with `HYL` such as Methyl, Ethyl etc. Middle truncation retrieves words having the same characters at both left and right hand side. Irrespective of the type of truncation, it must be stated that poor use of truncation in the search and retrieval of information from the web could lead to truncation error.

Proximity Search: This type of search strategy is commonly used when searching for contents in CD ROM and in online databases. In proximity search, the user or information seeker is able to determine if two or more search terms should occur adjacent to one another; or if one or more words occur in between the search terms. In addition, the proximity search also enables the user to determine if the search term should occur in the same paragraph regardless of the intervening words. The user is at liberty to use different operators in proximity search and such operators used may differ from one system to another.

Field-specific Search: The field-specific search is primarily used for electronic databases where an information seeker can search for all fields in a database or may limit the search to a particular field or fields. A field in this context refers to a single piece of information from a record. It could also be seen as a data structure for a single piece of data. For example, in a table called students record, matriculation number, name of students, level of study etc. will each be a field. As such, field information will vary from one database to another.

Limiting Search: Limiting search occurs when a researcher narrows a search by introducing certain criteria such as the type of information source, the language, date, title or the year of publication. By using these parameters, the information seeker is able to limit or narrow down the search results. However, the choice of which parameter to use to limit the search result is heavily dependent system design or database concerned.
Range Search: The range search is primarily used to select record or data within a certain range. This type of search strategy is mainly used when retrieving numerical information. Some of the symbols used include the less than and greater than operators. Other operators are Greater than (>); Less than (<>); Greater than or equal to (>=); Less than or equal to (<=).

2.9 Summary

In this unit, we gave a simplistic introduction to the World Wide Web. We explained that, the World Wide Web sometimes referred to as the ‘Web’, holds together a pool of information from different sources (web addresses). We also noted that the information in the Web is in different

formats such as text documents, images, videos, animations, audios amongst others. If you recall, you will remember we also identified and discussed some of the Web associated technologies and terminologies like HTTP/HTTPS, HTML, web servers, web browsers and URL. In this unit, we equally discussed the web search engine as integral software for the retrieval of information from the World Wide Web. We noted that the search engine performs three main tasks which include web crawling, indexing and ranking of web results after a query has been entered by the user. We then went further to classify the web search tools into two categories of type 1 search tool and type 2 search tool. Finally, we explained the web search services within the context of information retrieval on the World Wide Web.

Self-Assessment Exercis(s)

1. Discuss the function of the search engines in information retrieval.
2. Identify five examples of General search engines and Meta search engines.
3. Describe the role of HTTP, HTML and CSS in the web.
4. What do you understand by web search services?

2.10 References/Further Reading/Web Resources

Aggarwal, C. C. (2018). *Information Retrieval and Search Engines. Machine Learning for Text, 259– 304.* doi:10.1007/978-3-319-73531-3_9

Britannica (2021). World Wide Web. Retrieved from https://www.britannica.com/topic/World-Wide-Web

Couldry, N. (2012). Media, Society, World: Social Theory and Digital Media Practice. London: Polity Press.

Fecko, M. B. (1997). Electronic Resources: Access and Issues. London: Bowker Saur.

Martin, S.; Murugiah, P.; & Nandhini, K. (2016). Web Based Library Services: An Overview. *International Journal of Library Science and Information Management (IJLSIM*) Vol.2 (3) July-September, PP.19-85.

Web Crawler (2008). Architecture of a Web crawler. Retrieved from https://en.wikipedia.org/wiki/Web_crawler#/media/File:Web CrawlerArc hitecture.svg

Unit 3: Digital Libraries and Retrieval Imperatives

Unit Structure

3.1 Introduction

The retrieval of information is one that has received several attentions in recent times due to the emergence of new technologies and their associated complexities. Among such technological innovations are digital libraries. Digital libraries, loosely refers to as an organised collection of information resources in digital or electronic form, has its own unique ways of operations which differs from other technologies and traditional libraries. Due to the structure and organisation of information in digital libraries, there is the need to discuss information retrieval in digital libraries. The fusion of digital libraries and information retrieval could help to meet the information needs of library users while eliminating time wastages. This unit will therefore focus on the digital libraries and information retrieval imperatives. Effort will be made to discuss the text search for bibliographic databases as well as the document or full text search in a digital library.

3.2 Intended Learning Outcomes

By the end of this unit, you will be able to:

- explain the concept of Digital Libraries
- explain the concept of bibliographic search and full text search
- explain information retrieval imperatives and digital libraries
- Mention the challenges of digital libraries.

3.3 Concept of Digital Libraries

According to Solvberg (2020), the primary goal of any Digital Library (DL) is to meet the demands of its users. The ability to access digital assets through a computer network is a must. The fact that Digital Libraries' information is maintained, permanent, and reliable is critical. Digital collections, a working environment, and technology and services make up a Digital Library. Information Discovery, or how to find information in the Internet environment, is a common issue. Digital Libraries are collections of digital items that include text, sound, maps, movies, photographs, and other media. It will be discussed how metadata may be used to describe digital items. Metadata is frequently categorised as follows: descriptive, structural, and administrative, and as a result, it supports more labor duties than only information retrieval (IR). The IFLA paradigm and the Resource Description Framework (RDF) are described, as well as other metadata formats (MARC, Dublin Core, and others) (Solvberg, 2020).

The term digital libraries have been used interchangeably with other related terms such as online library, virtual library and digital repository. As a result, several scholars have given different explanation of what the term means. According to the Digital Library Federation (1998), the term digital library refers to "organisations that provide the resources, including the specialised staff to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities". However, a more generally accepted definition of digital libraries was given by David (2012). David defined digital libraries as a collection of documents in organised electronic form, available in digital format and which requires the use of digital technologies to appreciate its contents. The content in a digital library could range from digital objects such as text, videos, audio, images, magazines, books etc. Beyond aiding the storage of information resources, digital libraries are used to organise, search and retrieve information resources contained in the database.

Digital libraries may be accessed online or offline, they may also vary in sizes depending on the storage medium used and the capacity of the medium. The scope of digital libraries could also vary from one library to another and it is mainly dependent on the category of users the library was established to serve. The information bearing materials in a digital library could include all disciplines and all areas of human knowledge. As opposed to the traditional libraries, digital libraries are not limited by geographical location (especially when resources can be accessed online) and time.

3.3.1 Features of Digital Libraries

Digital libraries have some distinct features which make it different from traditional libraries and give it an edge of the physical library. Aside the issue of requiring digital technologies to access the content in a digital library, several features of digital libraries gives advantage to it when compared with the conventional library.

For instance, the issue of adequate space in conventional library does not come up with digital libraries as a large amount of information resources could be stored in a very little space in digital libraries. More importantly, digital library increases accessibility to information resources as the content can be access at any time and in any location thus, eliminating the restrictions faced in conventional libraries.

Similarly, the digital library increases the availability of information resources to users who may not otherwise be patrons of traditional libraries due to geographical barrier or institutional affiliation. Cleveland (1998) identified the following as features of digital libraries:

1. **No Physical Boundary (Libraries without walls):** One prominent feature of the digital library is that it is not confined by physical boundary. The content in the library could be accessed, updated and retrieved within any given location since the library is often designed within a given network such as the Local Area Network (LAN) Municipal/Metropolitan Area Network (MAN) and Wide Area Network (Internet) (WAN).

2. **Require Digital Technology to Access its content:** A unique feature of digital library is that it requires the users of the library to use digital technologies such as laptop or desktop computer, computer networks, Tablets, smart phones, etc. to access its contents. This is because resources in digital libraries are usually in digital form.

3. **Multiple and Simultaneous Access:** As opposed to the conventional library where only one user can use an information resource at a given time, the digital library allows for multiple or simultaneous access to information resources.

4. **Space Requirement:** The digital library is able to store large amount of information resources in a small space without taking too much physical space. This makes it possible for many traditional libraries to switch to hybrid libraries.

5. **Preservation and Conservation:** Information resources stored in digital libraries offers long term preservation to materials that would ordinarily suffer from degradation as a result of repeated use.

6. **24/7 or Round the Clock Access:** One notable feature and advantage of the digital library is that it provides round the clock access to it resources. The implication of this is that library users are able to access the content of digital library at any given time. This gives it advantage over contents in traditional libraries that can only be accessed during library opening hours.

7. **Easy Retrieval of Information Resources:** A key feature of the digital library is that its contents can be easily retrieved. Many digital libraries provide friendly user interface which allows users to search for documents using search terms such as title, words/phrase, name of authors, subject etc., to search the entire collection.

3.4 Digital Libraries and Information Retrieval

The idea of digital libraries and information retrieval shares common objective of helping to meet the information needs of users through easy access to information. However, digital libraries, more often than not, houses diverse kinds of content such as bibliographic records, full text content, aggregated contents amongst others. Therefore, to understand information retrieval in digital libraries, it is useful to group the different types of knowledge-based information so as to comprehend issues of information retrieval in digital libraries.

3.4.1 Bibliographic Content/Databases

The bibliographic content is often seen as the first category of electronic content in a digital library. This content primarily constitutes the bibliographic citation of information resources such as books, article, videos, images and other kinds of information materials. The citation could include the name of the author, title of the content, publisher's information, and keywords amongst others.

In the case of a journal article, the citation could include the volume, issue and page number. The bibliographic content, commonly found in bibliographic databases of a digital library, has been the mainstay of information retrieval systems for a long time now.

The content in this case does not contain the actual information being sought by a user but is a pointer to where the information can be accessed.

The bibliographic database typically contains bibliographic references on a particular subject. For example, one notable and widely used bibliographic database is MEDLINE (Medical Literature Analysis and Retrieval System Online), produced by the National Library of Medicine in the United States of America. Other examples include the CINAHL (Cumulative Index to Nursing and Allied Health Literature), GALILEO database Social Science Abstracts, and the Internet Movie Database. Web catalog is one of the latest examples of bibliographic database. The catalog provides information consisting mainly of Web pages which contain links to other Web pages and sites. Health-web, Health-Finder, Yahoo and Open Directory all falls with the examples of a web catalog. In the context of information retrieval, the bibliographic databases in a digital library provide descriptive information of an item or information resource.

However, such item or resource is not in itself provided in the database; only information about the items such as author, subject, title, publishers' information (or imprint), etc. In some cases, a short abstract is provided.

3.4.2 Full Text Content/Database

As the name implies, this type of content or database contains full text of a publication. The full text search in a database is widely known as document search and it is a common search conducted in digital libraries. This type of database started to gain prominence from the 1990s when advancement in the technology made it economically and technologically possible for full content to be stored using technology. The content in this category consist primarily of E-books, journal articles, reports and other kinds of document. The full text database is considered an improvement to the bibliographic database as it gives users the actual content being sought for. Information retrieved from the bibliographic databases could be used to retrieve content from the full text databases. For instance, Research Library in GALILEO provides information on the citation and the entire content (text). It should be noted that the full text databases could also provide non-textual content such as numeric information, image, audio (such as MP3 format) and video files. The retrieval of these content only require the information seeker to enter the keywords or citation of the item in the search box of the database and click the enter key. This makes it easy for users to retrieve and use the content in a full text database when compared to searching and retrieving the information resources in the conventional library. Examples of full text databases include EBSCO Host, JSTOR, Lexis-Nexis, ProQuest Central etc.

3.4.3 Aggregations

This is another category of knowledge-based information in a digital library in connection to information retrieval of content. In the context of

digital library, aggregation loosely refers to the process of grouping or combining of two or more entities to more a new and more meaningful entity. Aggregation of content combines the different kinds of content such as citations and full text content to make a whole meaningful new content. This means that aggregations have a wide variety of different types of information serving diverse needs of information seekers or users. Aggregated contents are developed for all kinds of users, such as students, practicing professionals and researchers. Example includes the MEDLINEplus which has both full text content and bibliographic content (citations) aggregated for ease of access to a particular topic. Other examples of aggregators for academic databases are EBSCO, ProQuest, Dialog etc. In relation to digital libraries, aggregators can be divided into three kinds, namely; hosting aggregators (such as Highwire Press, Ovid and Dialog), Gateways aggregators (such as CrossRef, ISI, BIOSIS and SFX from ExLibris) and Full text aggregators (such as EBSCO and ProQuest). These aggregators are useful tools for researchers and information seekers as they facilitate searching and access to multiple related databases.

3.5 Indexing in Digital Libraries

Digital libraries, just like the Internet and other repositories indexes its contents to make it more easily discoverable and accessible to users. Indexing in digital libraries makes search and retrieval of information resources faster than in conventional libraries. The indexing in this case could be done manually using controlled vocabulary or it could be in automated form. The manual indexing is usually done by human indexers who use standard terms, keywords and synonyms to assign indexes to an item or a document. The indexers follow a set of protocol which ensures that only terms that are the canonical representations of the content are used to generate the index. The automated indexing, on the other hand uses computer programs and technologies to generate indexes for content by using notable words in the document as index terms. This method of indexing is what is currently being applied to most content in digital libraries has proliferation of information resources (information explosion) has made it impracticable to manual indexing to be carried out on all published document.

3.6 Challenges of Digital Libraries

Digital libraries face a lot of challenges which are highlighted below.

Infrastructure constraints – The main roadblock here is a lack of high-capacity bandwidth for network and internet access, as well as a weak computer infrastructure in libraries

Lack of professional expertise – In today's changing climate, where technological developments push librarians to embrace and adapt to the changes or risk being bypassed on the information highway, there is a lack of professional expertise.

Absence of high-quality contents — Nigeria has a rich heritage of arts, folklore, spirituality, traditional knowledge, and so on, all of which are unused and may have been destroyed as a result of disasters, terrorism, and conflict. As a result, because information is largely in print, it is not easily transmitted or retrieved. In order to tackle the problem, they must be digitised in order to have access to or retrieve the extraordinary information.

Lack of strong digital policy - In Nigeria and other developing countries, most libraries lack ICT planning and strategy plans to address the issues brought by technological advancements, information overload, and user demand for effective retrieval. Legislators who are unaware of or unfamiliar with the requirements of digital preservation are more likely to pass laws that either neglect or inadequately address digital preservation challenges. As a result, digital resource preservation is not taken seriously.

Technological Obsolescence - Markets are full of a variety of digital formats that continually change from time to time with some formats getting obsolete (Caplan, **2004**). Technological obsolescence comes as a result of continuous upgrade of operating systems, programming languages, applications and storage media. Such changes or updates make preservation of digital materials meaningless.

For example, a Ph.D. dissertation submitted in the School of Postgraduate Studies was backed-up in a floppy disk before submission five years ago. The researcher after the five years, wanted to access the thesis so as to show it to his colleague working on related topic. Unfortunately, he was unable because the generation of computers currently available was never and could not open the disc due to technology updates or decay. The available hardware does not have the right drive for the researcher to access the information and therefore rendered the dissertation inaccessible (**Asogwa, Ilo, Asadu, Igbo, and Asogwa, 2021**).

Copyright Legislation - Most national legislations do not clearly express the copyright of software required to access digital material, as well as the right to copy for preservation. For example, due to copyright rules, a subscriber to an internet-based information service is obliged to renew the access license on a regular basis, even for items that have been paid for for a long time, in order to continue viewing the same information. Another issue is that digital progress has been far too fast and expensive

for governments and organisations to adopt timely and well-informed preservation policies.

Lack of Collaboration and Partnership - Another important issue with digital information is the absence of coordination and communication among libraries and stakeholders, as well as the lack of explicitly defined responsibility for the long-term preservation of digital assets. Governments, curators, publishers, relevant companies, and heritage organisations are all lacking in collaboration and partnership.

Lack of Disaster Preparedness - Another issue with digital preservation is the risk of digital media being lost in the event of a calamity such a fire, flood, device failure, or virus assault. In the absence of crisis preparedness, planning, and migration actions, precious information resources in libraries are frequently lost forever.

3.7 Summary

At the beginning of this unit, we introduced the concept of digital libraries and we also explained some of the features of digital library. You will recall that we defined digital library as a collection of documents in organised electronic form, available in digital format and which requires the use of digital technologies to appreciate its contents. We also identified some of its features as allowing for multiple and simultaneous access to its collection, allowing for 24/7 or round the clock access, no physical boundary, requires digital technology to use its contents amongst others.

In this unit, we also discussed information retrieval within the context of digital library. The different knowledge-based information such as bibliographic content, full text content and aggregators were discussed. Finally, the method of indexing in digital libraries which allows for easy retrieval of content was also discussed.

Self-Assessment Exercise(s)

1. Discuss the concepts of Digital Libraries.
2. Differentiate between Bibliographic databases and Full text databases, citing examples.
3. Explain the two types of indexing in digital libraries.

3.8 References/Further Reading/Web Resources

Asogwa, B.E; Ilo, P.I; Asadu, B.U; Igbo, H.U & Asogwa, M.N. (2021). Preservation of Digital Collections in Academic and Research Libraries: Issues, Strategies and Challenges. Proceedings of the 7th

national conference of academic and research libraries section held September 27 – October 1, 2021 at Nnamdi Azikiwe Library University of Nigeria, Nsukka.

Akst, D. (2003). The Digital Library: Its Future Has Arrived. *Carnegie Reporter*, 2(3), 4-8.

Borgman, C. (1999). "What are Digital Libraries? Competing Visions," Information Processing and Management, 35: 227-244.

David, B. (2006). The Place of Universities in the Information Value Chain. Retrieved from: http://www.iatul.org/conference/proceedings/vol11/papers/Ball/Ball.htm

Lanagan, J., Smeaton, A.F. (2012). Video digital libraries: contributive and decentralized. *International Journal on Digital Libraries*. 12 (4): 159–178. doi:10.1007/s00799-012-0078-z.

Miller, N., & Lacroix, E. (2010). MEDLINEplus: Building and Maintaining the National Library of Medicine's Consumer Health Web Service. *Bulletin of the Medical Library Association*, 88: 1 1-17.

Salton, G. (1968). Automatic information organization and retrieval, New York NY: McGraw-Hill.

Solvberg, I. (2020). Digital Libraries and Information Retrieval. Proceedings of the 3rd European summer school on lectures on information retrieval, pp. 139-156.

Unit 4: Evaluation of Information Retrieval Techniques and Processes

Unit Structure

4.1 Introduction

The choice to use or adopt an information retrieval system is dependent on its effectiveness. Such effectiveness is determined by several factors and functionalities which includes recall, precision, relevance, exhaustively and specificity amongst others. The importance of an effective retrieval system cannot be overemphasised as it has the capability to promote user happiness. When a retrieval system retrieves relevant information in timely fashion, users of such system is bound to be happy. Thus, making relevance of the result as an important factor in evaluating a retrieval system is crucial. This unit therefore focuses on the different evaluation criteria for information retrieval systems, online systems, Online Public Access Catalogue (OPAC), and Internet System.

4.2 Intended Learning Outcomes

By the end of this unit, you will be able to:

- explain the evaluation criteria for online systems
- identify the general evaluation criteria for information retrieval systems
- justify the need for the evaluation of information retrieval systems.

4.3 General Evaluation Criteria for Information Retrieval Systems

Over the years, a considerable deal of study has been done on the evaluation of IR systems. There is debate about whether ways to evaluate produce the best results for assessing searching in the systems they are utilising, as there is in many other fields of research. To put the findings into context, a variety of frameworks have been proposed. One of such frameworks (Hersh, Starvri, and Detmer, 1998) framed evaluation around six questions that someone promoting the adoption of IR systems might ask:

1. Was the system used?
2. For what was the system used?
3. Were the users satisfied?
4. How well did they use the system?
5. What factors were associated with successful or unsuccessful use of the system?
6. Did the system have an impact?

According to Chimah, Unagha and Nwokocha (2010), information retrieval system is a system for recovery and extraction of information from a collection or database in response to an information problem".

Due to advances in computer systems, several information retrieval systems with diverse range of functionalities have been developed over the years, thus, rendering it impracticable or difficult for libraries or organisations to adopt a particular information retrieval system at any given time due to the number of options to choose from. In order to address this, Bar-llan (n.d) identified different criteria for evaluating and choosing an information retrieval system for organisations such as libraries.

However, a simpler method of arranging evaluation results divides methodologies and studies into two categories: system-oriented (i.e., the evaluation focuses on the IR system) and user-oriented (i.e., the evaluation focuses on the user) (i.e., the focus is on the user.

Among the criteria are relevance of the retrieved results, response time, user friendliness and accuracy of retrieved results. Beyond the above listed criteria, effectiveness and efficiency are also regarded as two primary parameters for determining the suitability of information retrieval systems. While effectiveness in this context refers to the level up to which the given system attained its LEARNING OUTCOME, efficiency refer to how economical the system is in achieving its LEARNING OUTCOME. As such, cost factors such as response time, time taken by the system to

provide an answer, user effort, the amount of time and effort needed by a user to interact with the system and analysed the output retrieved in order to get the correct information are considered general evaluation criteria for information retrieval systems.

Other specific criteria are relevance, recall and precision, exhaustivity and specificity, amongst others.

4.3.1 Relevance

The measurement of the effectiveness of an information system is determined by the relevance of the information retrieved, especially in terms of precision and recall. Lesk and Salton (1968) affirmed that relevance assessment is crucial to ascertaining information retrieval effectiveness of a retrieval system. It is believed that an information retrieval is very effective if it retrieves large content of precise and relevant information in a single search. In the context of information retrieval systems, relevance denotes how well information retrieved from a retrieval system meets the users` information need. When measuring relevance, factors such as currency/ novelty of the result, authority and timeliness are considered.

4.3.2 Recall and Precision

The relevance-based measures of recall and precision are the most extensively used methods for evaluating the performance of IR systems. These metrics quantify how many relevant documents the user found in the database and throughout his or her search. They use the number of relevant documents (Rel), the number of recovered documents (Ret), and the number of retrieved documents that are also relevant (Ret) (Retrel. The percentage of relevant documents retrieved from the database is known as recall.

Recall and Precision are two primary indicators used to determine the effectiveness of an information retrieval system and how well the system satisfies the user requirements. This is so because recall and precision relate directly with the performance of an information system. To determine the effectiveness of an information system, a comparison is made of the system performance and functionality so as to achieve user satisfaction. In information retrieval, recall is defined as ratio of the number of retrieved and relevant documents (the number of items retrieved that are relevant to the user and match his needs) to the number of possible relevant documents (number of relevant documents in the database). This definition implies that, in a retrieval system, there is the possibility of retrieving irrelevant information or document. However, an effective system will always return a high volume of relevant information that meet the information need of the information seeker from a pool of relevant database. The usefulness of recall is seen in its

ability to assess the extent to which a retrieval system processes a query and retrieve relevant information from a list of related information.

On the other hand, McCafrey (2016) defined as the fraction or percentage of retrieved relevant documents. Precision identifies with information or documents that are relevant to the information seeker. Both precision and recall focus on system performance by evaluating the collection of relevant document retrieved from an information system with a specific period of time. The result of recall and precision, to a large extent, determine the overall relevance of the information retrieval system. To create a recall-precision table for a single query, determine the recall intervals that will be utilised first.

Precision ($\overset{P}{\_\_}$) is the fraction or proportion of retrieved documents that are relevant in the search. A typical approach is to use intervals. The formula for both precision and recall is given below:

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved}) \quad (36)$$

This metric solves the question of how many of the documents returned in a search are relevant. When comparing ranking systems to non-ranking systems, one issue is that non-ranking systems, which often use Boolean searching, tend to retrieve a predetermined set of documents and, as a result, have fixed recall and precision points. Relevance ranking systems, on the other hand, have varying recall and precision levels depending on the size of the retrieval set the system (or the user) has chosen to display. As a result, many evaluators of relevance ranking systems would generate a recall precision table that shows precision at various recall levels. The table is constructed by determining the highest level of overall precision for a given recall interval at any point in the output. For the recall interval 0.0, the maximum degree of precision at which the recall is greater than, equal to, or less than zero would be used.

Recall ($\overset{R}{\_\_}$) is the fraction of relevant documents that are retrieved

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant}) \quad (37)$$

In other words, recall answers the query, "What fraction of all relevant documents have been received from the database for a given search?"

The denominator of Equation (37) implies that the total number of relevant documents for a query is known, which is a flaw. However, with

the exception of the tiniest of databases, locating all essential documents in a database is unlikely, if not impossible. As a result, most studies employ the relative recall metric, with the denominator renamed to indicate the number of relevant documents found through several searches on the query topic.

These notions can be made clear by examining the following contingency table:

(G)

	Relevant	Nonrelevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

Then:

$$\frac{P}{R} = \frac{tp/(tp+fp)}{tp/(tp+fn)} \quad (38)$$

4.3.3 Exhaustivity and Specificity

The concept of exhaustivity and specificity is usually used within the context of subject indexing in information retrieval as it helps to indicate what the document in a database is about, to summarise its content or to increase its findability. Through subject index, bibliographic indexes can be created to aid the retrieval of documents on a particular subject. Exhaustivity here list all possible index terms so as to increase the chances of retrieving relevant articles or having a higher recall. Simply put, exhaustivity describes the extent to which all the things and topics mentioned in the document is described in the index. Specificity on the other hand, explain how closely the index terms match the topics they represent or how narrow or wide the search terms are. Specificity in information retrieval is important because it increases the chances of retrieving relevant results. For instance, in specificity, narrow terms return fewer documents, but most of them are probably relevant.

4.4 Evaluation Criteria for Online Systems

The evaluation criteria for other information retrieval systems are not completely the same with online systems given the fact that the evaluation criteria for traditional information retrieval systems such as recall and precision does not comprehensively address online systems which operate in a highly dynamic and distributed environment. Furthermore, the one uniqueness of online systems is its ability to direct users to physical as

well as virtual resources, thus making them hybrid as they usually contain metadata (i.e. the data used to describe the information resources).

In addressing the evaluation criteria for online systems, it is important to note that online systems in this context include online databases and Web search engines. While the former refers to a collection of information resources, the latter includes directories, search engines, meta-search engines, and specialised search engines. Online evaluation of information retrieval system differs from traditional retrieval system in that it relies on users` experience as a yardstick for measuring success. Users' satisfaction is a paramount indicator of determining a good online retrieval system. In this method, real users are allowed to engage the information retrieval system and their interactions are observed in-situ while they interact with the system. According to Hofmann, Li and Radlinski (2016), there are several metrics for evaluation of online systems. Among the metrics are the absolute document level metrics, relative document level metrics, absolute ranking-level metrics, relative ranking-level metrics, and the session-level metrics.

4.5 Evaluation Criteria for Online Public Access Catalogue (OPAC)

The online public access catalogue, often simply referred to as OPAC or library catalogue is a database of library holdings which could either be accessed online or offline. The OPAC is a retrieval tool which allows library users to easily retrieve materials from the library. With OPAC, library users are able to retrieve information resources using the name of the author or title of the work or the subject. Unlike other forms of library catalogue, the OPAC require the use of technological tools (hardware and software) to appreciate its content. However, the continuous advancement of technologies has given rise to several OPAC technologies which are either proprietary or open source. Thus, there is the need for critical evaluation before adopting a particular OPAC system in the library. Among the popular evaluation criteria for OPAC system is the question of user friendliness. How simple or complex is the system to use? How much training is needed to use the system? Can users interact with the system themselves or will they always require the support of the librarians? Closely followed to the question of user friendliness is the question of cost. In selecting or choosing an OPAC system, the librarian must also put the overall cost of acquiring the system into perspective. A comparative cost study must be carried out to ascertain the most suitable considering the situation. Aside cost, the library must also understand the access restrictions associated with the OPAC system to be adopted. Questions must be asked if librarians can charge out library materials using the system. Is access to the information materials online only or remote site location, etc.?

4.6 Evaluation Criteria for Internet Retrieval Systems

The Internet is a system architecture with wider network capabilities which allows for global interconnectedness of computer networks and systems through Internet protocol suits such as the transmission control protocol and Internet Protocol. The Internet is widely seen as a global network of networks which carries a vast range of information resources and services, such as the inter-linked hypertext documents and applications of the World Wide Web (WWW), and electronic mail. In evaluation of Internet system, Lancaster and Warner (1993) identified three main criteria for evaluation of information systems. These criteria include cost, time and quality. While the first two criteria are often seen in the evaluation criteria for other information retrieval systems, the quality consideration deals with other factors such as coverage of the database, relevance of the search result (precision), completeness of the results retrieved (recall) and the newness of the result. Also included in the quality consideration is the issue of completeness and accuracy of data. Since the Internet houses other interrelated technologies such as search engines, Okeowo (2019) explained that factors such as relevance of ranking, hyperlink, as well as interface must be consider when evaluating the internet and its interrelated information systems. In 2016, Rendall used several criteria such as speed of search, ease of use, number of result returned in a search, number of quality return results, and relevance of search results. Furthermore, Backfiles (2007) identified ways of measuring the internet retrieval system. Among the ways identified by Backfiles are precision, output option, interface, response time, database coverage, relevance ranking (output ranking or scoring) amongst others.

In addition, Igbinovia (2017) affirmed that the time period between when a user submits a request and the time the user receive the search result (otherwise known as response time), is a good measuring tool for effectiveness of the Internet retrieval system. Likewise, the validity of links directly related to the database quality and the search capability of the internet to handle search queries are also important criteria for evaluating the Internet system.

4.7 Justification for the Evaluation of Information Retrieval Systems

The advances in information and communication technology have brought about multiple options of information retrieval systems to be adopted or used by an individual or by an organisation. Due to factors such as costs, it is difficult for organisations to adopt multiple information retrieval systems, hence, the need to critically examine and evaluate information retrieval systems in order to adopt the most suitable. According to Rasak (2017), one important justification for the evaluation of information retrieval systems is so as to measure information retrieval

effectiveness in a standard way. Furthermore, the evaluation of information retrieval systems helps the users to understand the document retrieval system of an information system. Through evaluation, users can conduct a test suite of information needs, expressible as queries. In addition, the evaluation provides an assessment of either relevant or non-relevant judgment for each query-document pair. Swanson (2018) explained that the evaluation of information retrieval systems helps users to determine whether and how well goals or performance expectation are being fulfilled. The evaluation of information retrieval systems helps users to uncover some basic information of a successful programme; the reasons for success or failure, understand how to increase the effectiveness of the programme amongst others. However, Martin (2019) added that a good justification for the evaluation of information retrieval systems is to establish a foundation of further research on the reason for the relative success of alternative technique.

4.8 Summary

At the beginning of this unit, we discussed the evaluation criteria for information retrieval systems. Evaluation criteria such as relevance, recall and precision, exhaustivity and specificity were comprehensively discussed. Also discussed in this unit were the evaluation criteria for online systems, online public access catalogues (OPAC), and Internet retrieval systems.

In this unit, we equally introduced the information retrieval measuring concepts such as precision and recall. It was explained that while precision deals with the total number of relevant retrieved documents over the total number of retrieved documents from a single query, recall focus on the ratio of retrieved relevant document to the number of possible relevant documents in the database. Justification or reasons for evaluating the different information retrieval systems was also discussed.

Self-Assessment Exercis(s)

1. Discuss the concepts of exhaustivity and specificity within the information retrieval context.
2. Differentiate between recall and precision.
3. Identify five criteria for evaluating online public access catalogue (OPAC).

4.9 References/Further Reading/Web Resources

Berenci, E., Carpineto, C., Giannini, V., & Mizzaro, S. (1999). Effectiveness of Keyword-based display and selection of retrieval results for interactive searches. In Lecture Notes in Computer Science, 1696: 106-125.

Bharat, K. & Broder, A. (1998). A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines. In Proceedings of the 7th International World Wide Web Conference. Retrieved fromhttp://decweb.ethz.ch/WWW7/1937/com1937.htm

Brewington, B. E., & Cybenko, G. (2000). How Dynamic is the Web? In Proceedings of the 9th International World Wide Web Conference, May 2000. Retrieved from http://www9.org/w9cdrom/264/264.html

Chakrabarti, S., Dom B., Kumar, R. S., Raghavan, P., Rajagopalan, S., Tomkins, A., Kleinberg, J. M., & Gibson, D. (1999). Hyper searching the Web. Scientific American, 280(6): 54-60. Retrieved from http://www.sciam.com/1999/0699issue/0699raghavan.html.

Cho, J., & Garcia-Molina, H. (2000). Synchronizing a Database to Improve Freshness. SIGMOD RECORD, 29(2): 117-128.

Chu, H. & Rosenthal, M. (1996). Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology. ASIS96. Retrieved from http://www.asis.org/annual-96/Electronic-Proceedings/chu.htm

Clarke, S.J., & Willett, P. (1997). Estimating the Recall Performance of Web Search Engines. Aslib Proceedings, 49(7), 184-189.

Douglis, F.; Feldmann, A.; Krishnamurthy, B.; & Mogul, J. (1997). Rate of Change and Other Metrics: A Live Study of the World Wide Web. In Proceedings of the Symposium on Internet Technologies and Systems, Monterey, California, December 8-11, 1997. Retrieved from http://www.usenix.org/publications/library/proceedings/usits97/full_papers/doglis_rate

Hersh, W.R; Starvri, P.Z; and Detmer, W.M. (n.d). Information Retrieval and Digital Libraries, pp.660-697.

Holscher, C., & Strube, G. (2000). Web Search Behavior of Internet Experts and Newbies. In Proceedings of the 9th International World Wide Web Conference, May 2000. Retrieved from http://www9.org/w9cdrom/81/81.html

Gordon, M., & Pathak, P. (1999). Finding Information on the World Wide Web: The Retrieval Effectiveness of Search Engines. Information Processing andManagement, 35, 141-18

Module 5 Information Retrieval in Nigerian Libraries and Information Centres

Unit 1 Users and Information Retrieval
Unit 2 Evaluation of Information Sources
Unit 3 Problems and challenges of information representation and retrieval in libraries and information centers in Nigeria.

Unit 1: Users and Information Retrieval

Unit Structure

1.1 Introduction

Information retrieval systems have diverse users with heterogeneous characteristics such search experience, information literacy skill, self-efficacy, available time, knowledgeability, and efficiency. Understanding the various users and their information seeking behavior or pattern would stimulate actions towards effective and efficient retrieval systems. This unit examines four types of users of information retrieval systems in the Nigerian context.

1.2 Intended Learning Outcomes

By the end of this unit, you will be able to:

- explain the different categories of users of information retrieval system
- distinguish between the Digital natives and immigrate users
- Identify professional users, information specialist users, novice users.

1.3 Users and Information Retrieval

Users have diverse characteristics and varied information needs. The information retrieval system must anticipate, provide for meet and exceed expectations of the various users. Mutshewa (2008) citing Meadow (1992) identified three categories of users of information retrieval systems as (a) information specialist or reference librarians (b) professional end – users (c) novice or others.

1.3.1 Information specialists or reference librarians

These categories of users are constantly engaged in providing information service to diverse clientele as a business. They interact with the retrieval system to satisfy the information need of others. Meadow (1992) opined that they must have good knowledge of the following to be successful:

- The types of information available in a discipline or specialised areas that they serve;
- The ways of using information systems relevant to the subject or institution; This includes basic information technology skills such as installing software in a computer system and using the relevant software to access and search both local and remote databases;
- Basic terminology in the area of specialty, including the use of search tools such as thesaurus and subject heading lists;
- Conducting basic studies that would enable them to determine the information needs of the client; and
- Conducting reference interviews.

1.3.2 Professional end-users

This constitutes of experts from all walks of life and various disciplines. These groups of users do not need the intermediary role of the reference librarian to retrieve needed information but engages the information retrieval system directly. This category of users are subject experts in the area of required information and can refine their search terms more specifically to obtain greater results of relevant information (Barsky and Bar-Ilan, 2005). They require appropriate training in information retrieval skills to complement their subject knowledge for greater effectiveness.

1.3.3 Novice Users

This group is not conversant with the diverse information sources and search strategies. They rely preponderantly on the specialist group to meet their information needs. Some of them cannot articulate accurately their information needs and often lack the capacity to utilise effectively information retrieved for them. The design of information retrieval

system should ensure that novice users, who are unfamiliar with the intricacies of the system, can navigate through without external help or formal training but aided by embedded user-friendly features for effectiveness. Novice users gain dexterity in their retrieval capabilities with repeated activities and engagement with the information retrieval systems.

1.3.4 Digital natives and digital immigrants

Digital natives refer to users born into the digital age also known as millennia. They are more comfortable with the digital environment and have preference for online retrieval system, show sophistication and versatility in retrieving information from the web or databases. Whereas the digital immigrants show aversion and apprehension to interacting with digital retrieval systems, some of them use online retrieval system minimally and clamour for a hybrid system that makes for parallel provision for manual and digital system. The overall structure of the information retrieval system should be flexible and user friendly that the multifaceted needs and characteristics of various users are catered for in the search or query structure.

1.4 Summary

This unit explored the diversity of users of the information retrieval systems in Nigeria. The various users include: information specialists or reference librarians, professional end-users, novice users, digital natives and digital immigrants.

Self-Assessment Exercise(s)

1. Discuss four categories of users of information retrieval systems in Nigeria libraries and Information centres.
2. What do you understand by digital natives and digital immigrants?
3. Differentiate between Information specialists as users and professional end-users?

1.5 References/Further Reading/Web Resources

Mutshewa, A. (2008). Information retrieval systems, strategies and challenges for African information searchers. In: Aina, Mutula and Tiamiyu (edited). *Information and knowledge management in the digital age: concepts, technologies and African perspectives*. Third world

Meadow, C. T. (1992). Text information retrieval systems, San Diego, USA; Academic Press

Barsky, E. and Bar-Ilan, J. (2005). *From the search problem through query formulation to results on the web*. Online Information Review, 29 (1) pp. 75-89.

Unit 2: Evaluation of Information Sources

Unit Structure

2.1 Introduction

In the process of information retrieval library patrons would come across varied sources of information such as books, journals, articles, encyclopedia, websites, databases, manuscripts, government publications, monographs, microforms. It is incumbent on the user to decipher between facts and opinions as well as interrogate the contents to determine their suitability for the purpose of meeting the information need. An information user should be able to critically evaluate the appropriateness of the information sources before relying on the information. This unit provides general criteria for evaluating information sources to ensure appropriateness and relevance.

2.2 Intended Learning Outcomes

By the end of this unit, you will be able to:

- define Information sources
- differentiate between primary, secondary and tertiary sources of information with appropriate examples
- discuss the major criteria for evaluating information sources.

2.3 Concept and Types of Information Sources

People have diverse information needs and the usual critical concern is where to find the relevant information for the use of library patrons. Information sources simply refer to where to find requisite information. Information source can be in print or electronic format. It includes electronic files, Internet, people and printed materials. Information sources are usually classified into primary, secondary and tertiary depending on the immediacy of the author or creator to the content/event.

Primary sources present information in its original state devoid of interpretations or critical evaluations or summarisations. They are first-hand documents that provide direct evidence on the issue under reference. Examples include diaries, memoirs, autobiographies, speeches, journals,

manual script; Government documents technical reports, interviews, surveys, letters and correspondences, emails, etc.

Secondary sources constitute works which describe, interpret, analyse, synthesise or assess the primary sources. They are commentaries or value-added information to already existing first-hand accounts. They include bibliographies, biographies indexes, treaties, reviews, works of criticism and interpretation.

Tertiary sources present summaries or condensed versions of materials usually with references to the primary and or secondary sources. Tertiary sources include almanacs, abstracts, dictionaries, encyclopedias and handbooks.

2.4 Evaluation Criteria for Information Sources

There are standard criteria that the users can rely on to appraise the relevance and usefulness of particular information source. The criteria for evaluation of information sources include the following: authority, scope and depth of coverage, accuracy/objectivity, currency, citations, quality of writing, audience.

1. Authority

It is pertinent to know the scholarly pedigree of the author or publisher of the information resource. The author or publisher refers to the person or organisation that is chiefly responsible for the creation of the intellectual or artistic content. There could be co-authorship or joint authorship and the publisher could be a university, academic or professional association, government, international organisation, trade or commercial publisher. The reputation of the publisher is critical considering that we are in the era of the existence of predatory publishers. Authors with outstanding previous publications and profound expertise in their area of specialisation are preferred to relatively unknown other. However, unknown authors whose works are published by reputable publishers enjoy good visibility and reliability. Beyond the reliance on renowned authors and publishers, there is need to verify the academic and professional qualifications of authors, as well as expertise in the subject area under consideration. Library users will want to retrieve at all cost, materials associated with authors of outstanding reputation or profound expertise, even if it involves inter-library lending.

2. Scope and Depth of coverage

The user should ascertain whether the coverage of the material is in-depth or superficial. The basic questions in this regard are: does the publication extensively cover the topic? Is it a general treatment or detailed in terms

of specialisation? Is it an update of other sources, is there a definite contribution to knowledge. It is important to be assured that the scope and coverage of the material will meet the information need of the user. Relevant reviews about the publication appearing in journals, magazines or newspapers or credible online reviews can be used to determine the coverage as such reviews provide critical perspectives about the information source.

3. Accuracy and Objectivity

The user of information is interested in information that is credible and reliable. The publication must be based on verifiable and demonstrable evidence capable of replication. It excludes subjective assertions, biases, hearsays, stereotypes, opinions and propaganda. It must be rooted in facts, authority and flawless logic. It is imperative that eclectic perspective and balance are maintained in the interpretation of facts especially when controversial topics like religion, gender, politics, race, ethnicity and ideologies are concerned. Information users will be inclined to evaluate the authenticity and integrity of information retrieved from the collection for utilisation. Availability of credible links and references validating the information is central to defining its accuracy.

4. Currency

The currency or recency or up-to-date-ness of contents of any publication is crucial to determining appropriateness and relevance of same to users. Knowledge is not static but dynamic hence the need to ensure timeliness to accommodate new inventions, innovations, latest findings and current trends. Information sources are usually revised to reflect the current realities and new discoveries. It is important that users retrieve the latest edition of this information to avoid perpetuation of stable or obsolete knowledge. The authenticity and integrity of information must be ascertained to ensure reliability and effectiveness. However, disciplines like history, anthropology and archaeology find old information useful and useable.

5. Citations

Since knowledge is cumulative any reputable source must show proper attribution to other sources. This helps to authenticate the information contained in the information sources. It is done through provision of footnotes, bibliographies, hyperlinks to electronic resources. Citation indexes could be used to determine the extent of citedness or impact of the information source to scholarship, which is a measure of the quality of the source. Some useful citation indexes include Social Sciences Citation Index, Science Citation Index, Web of Science, Arts &

Humanities Citation Index, Scopus Citation Index, Conference Proceedings Citation Index, Citeseerx, Crossref, etc. Other criteria to be considered in the evaluation of information sources include quality and style of writing as well as target audience. Adherence to basic grammatical competence and use of language in clear communication of ideas are indicators of quality of writing. The sophistication of language and deployment of appropriate terminologies will be determined by the target audience. The audience could be undergraduates, postgraduate students, professionals, scholars and the general public.

2.5 Summary

The Unit explored the various criteria for evaluating information sources. This stems from the need to imbue in the students and library users the capability to determine the relevance of retrieved information with a view to meeting or exceeding their information needs. The standard criteria discussed in this unit for evaluation of information sources are: authority or pedigree of the author or publisher, scope and depth of coverage of the information, citations, quality and style of writing as well as target audience.

Self-Assessment Exercise(s)

1. What are the major criteria for evaluating information sources?
2. Differentiate between primary, secondary and tertiary sources of information with appropriate examples.

2.6 References/Further Reading/Web Resources

Ojedokun, A.A. (2007). *Information Literacy for Tertiary Education Students in Africa*. Third World Information Services Limited, Ibadan, Nigeria. Pp 129-143

Ifidon, S. E. (2006). *Modern Theory and Practice of Library Collection Development*.Department of Library and Information Science, Ambrose Ali University, Ekpoma, Nigeria. Pp 46-55

Adeyoyin, S. O. (2011). *Library and Information Resources Management: The Beginners' Text*. Eagle Publishers, Abeokuta, Nigeria.

Unit 3: Problems and Challenges of Information Representation and Retrieval in Libraries and Information Centers in Nigeria

Unit Structure

3.1 Introduction

Information representation and retrieval system are designed to facilitate easy and speedy access to information thereby satisfying the information need of users. It has been observed generally that there are problems and challenges associated with such systems globally as well as peculiar issues affecting developing nations in this regard. This unit will discuss such problems and challenges affecting information representation and retrieval in libraries and information centres in Nigeria.

3.2 Objectives

By the end of this unit, you will be able to:

- discuss challenges of information retrieval in Nigerian context
- proffer solution to improving the information retrieval system.

3.3 Problems and Challenges of Information Representation and Retrieval in Libraries and Information Centres in Nigeria

There are plethora of problems and challenges militating against information retrieval in Nigerian libraries and information centres. The users are discussed under the following sub-headings:

3.3.1 Dynamic Information Environment

Information representation and retrieval operates in a heightened dynamic digital environment. The advent of emerging technologies has necessitated new formats of information, new standards of representation and retrieval. International standards in library information description is now based on the principle of Resource Description and Access (RDA) which is designed for digital environment and aimed at helping users to find, identify, select, obtain and use information. Unfortunately, most libraries and information centers in Nigeria have not metamorphosed into the new era and are still using the metadata provided under Anglo-American Cataloguing Rules 2 (AACR2).

3.3.2 Poor Infrastructure and High Cost of Equipment

Modern information retrieval systems are based on emerging technologies. The low level of development in Nigeria characterised by weak telecommunication network and poor Internet broad band resulting in frequent internet downtime and outright disruption in some cases pose a major challenge to information retrieval in Nigerian libraries. Corollary to this is the fact that most technological equipment is imported and the astronomical exchange rate disparity makes affordability difficult. This also affects the maintenance of existing equipment or procurement of spare-parts.

3.3.3 Absence of Indigenous Content in most Databases

It has been observed that vast amount of indigenous contents are not represented in most databases as a result of either lack of documentation or lack of digital versions. Users searching for relevant information in their quest to understand and address Nigerian issues are usually disappointed for inability to retrieve information that matches their needs. Most indigenous contents are within the purview of tacit knowledge or embedded in the cultural ethos and oral tradition. The problem of recording, packaging and disseminating such information is herculean. It requires skills, expertise and financial resources. Scientific validation to provide documented evidence of the efficacy of oral assertions and the inability to capture all aspects of indigenous knowledge as artifacts using technology constitute challenges to libraries and information centers.

3.3.4 Inadequate Information Retrieval Skills

Most users do not possess requisite retrieval skills to navigate successfully the retrieval systems. They engage on trial and error approach leading to frustration. While the digital natives are receptive to modern information retrieval models, the preponderance of users (the

adult population) shows aversion and preference for the traditional system. This is predicated on the obvious lack of appropriate skills by this category of users. The current librarians' skills in the country have been adjudged to be below the global average in providing digital services to meet the demands of twenty-first century information retrieval imperatives.

The librarians should embrace new competencies that help to provide exceptional experiences for users, such as social media skills for instant communication to various users about retrieval of accurate information as well as deployment of Web 2.0 tools for effective virtual operations and interactions with their clientele in the libraries and information centers.

3.3.5 Overdependence on One Information Retrieval System

The structure, interface and several techniques vary from one retrieval system to the other. There is need to study each system for its peculiarity to enhance higher retrieval rate. This lack of familiarity with the basic features of the available retrieval systems leads to difficulty in locating relevant information. It is also gratifying to note that despite some of the drawbacks, there is growing incidence of automation of Nigerian libraries and information centres. It is becoming fashionable to have Online Public Access Catalogues (OPACs) as well as subscription to online databases. There is good prospect for information representation and retrieval in Nigerian libraries.

There is need to mount regular trainings to improve the retrieval skills of the users. Virtually all Nigerian libraries and users now have internet connectivity which enhance searching and retrieval of information from the global information network. However, there should be concerted effort at improving network infrastructure and guaranteeing regular electricity supply for effective retrieval systems.

3.3.6 Competition with other Information Retrieval Systems

Libraries and Information Centres face stiff competition with other information retrieval systems. The first point of call for users in the current dispensation for information retrieval is Google and other search engines that provide fast and credible alternatives to the hitherto celebrated traditional tools. By mere click of the mouse, millions of information could be sourced from disparate global information network. The libraries are under heightened pressure to provide services comparable to these other retrieval systems that characterise the digital era.

3.3.7 Power Supply

One of the critical challenges of digital information retrieval in Nigerian university libraries is incessant power outages. Information and communication technology infrastructures are hundred percent dependent of power supply for functioning. In Nigeria, there is serious poor electricity generation, distribution and management. Consequently, it results in intermittent power supply in most universities which seriously hampers digital information retrieval in libraries. This discourages a researcher who intends to visit the library for the retrieval of digital information.

3.3.8 Funding

Adequate funding is a critical precondition for effective provision of services in the library and information centres. The libraries are grossly underfunded and consequently incapacitated in approximating to international standards. Capacity building initiatives are hampered as many librarians do not attend international conferences and lack requisite exposure as well as benchmarking experiences that ought to engender astuteness, excellence, global competitiveness and best practices.

There is absolute and dire need to provide sufficient funding to ensure continuous relevance of the institutions.

3.3.9 Leadership Challenges

Leadership is critical to the growth and development of any institution or organisation. The extent to which librarians will adapt to technologies and mainstream to global standards is a reflection of their leadership. The famous maxim by John Maxwell that everything rises and falls on leadership is inconvertible. Many libraries and information centres in Nigeria appear very traditional in their outlook devoid of the effect of global transformation characterising the librarianship landscape. The LIS institutions in Nigeria do not have strategic plans that guide future pathways but engage in routine practices that are no longer consistent with best practices. Many of the office holders in the libraries are not necessarily leaders but managers who feel contented maintaining the status quo rather than inspiring changes and catalysing creativity and innovations as well as actualising institutional shared vision.

Appointment of heads of libraries must be based strictly on merit devoid of unnecessary clamouring for quota and lobbying which tend to foist mediocrity on the system. It takes a vibrant leadership to maintain relevance and grapple with the issues of change management in a globalised world.

3.4 Summary

This unit reviewed some of the problems and challenges associated with information representation and retrieval in Nigerian libraries and information centres. The challenges are not exhaustive but include; dynamic information environment, absence of indigenous content in most databases, inadequate information retrieval skills, overdependence on one information retrieval system, poor infrastructure and high cost of equipment

Self-Assessment Exercise(s)

1. Discuss the problems and challenges of information representation and retrieval in Nigerian libraries and information centres in Nigeria.
2. Provide solutions to improving information retrieval in Nigeria libraries and information centres.

3.5 References/Further Reading/Web Resources

Aina, L. O. Mutula, S.M. Tiamiyu, M. A. (2008). Information and knowledge management in the digital age: concepts, technologies, and African perspectives. Third world.

Chimah, J. N., Unagha, A. O. and Nwokocha U. (2010). Information retrieval in libraries and information centres: concepts, challenges and search strategies. Journal of Applied information science and technology, 4 (1) pp.20-24

Mabawonku, I.M. (2002). The Systematic Management of IndigenousKnowledge: A Review of OralInformation Projects in a Library School. *Proceedings of SCECSALConference. Pretoria. From Africa To The world- the Globalization of Indigenous Knowledge Systems*, South Africa, April15- 19. Pp. 6-7

Msuya, J. (2007). Challenges and Opportunities in the Protection and Preservation of Indigenous Knowledge in Africa. *International Review of Information Ethics*, 7. Retrieved from http://www.ir-i e.net/inhalt/007/38-msuya.pdf

Okore, A.M., Ekere, J.N and Eke, H.N. (2009). Promoting Access to Indigenous Knowledge in Digital Age: Libraries as Facilitators. Libraries Create Future: A Paper presented at the Nigerian Libraries Association, 47th Annual General Conference 2009, Ibadan, Oyo State, July 26-31, Pp.55-56

Zaid, Y. and Abioye, A. (2009). Museums, Libraries and Archives:Collaborating for the Preservation of Heritage Materials in Nigeria.Paper presented at the World Library and Information Congress: 75th IFLA General Conference and Council, August 23-27, Milan, Italy.Retrievedfromhttp://www.ifla.org/annual-conference/ifla75/index.htm