INFORMATION PROCESSING RETRIEVAL

M.L.I.Sc., Semester – I, Paper-III

Complied by

Dr. Md. Gouse Riyazuddin
M.A., M.L.I.Sc., PGDLAN., M.Phil., Ph.D.,
Library
Government Women's College
Guntur

Director
Dr. NAGARAJU BATTU
MBA., MHRM., LLM., M.Sc. (Psy).,MA (Soc)., M.Ed., M.Phil., Ph.D
CENTRE FOR DISTANCE EDUCATION
ACHARAYA NAGARJUNA UNIVERSITY
NAGARJUNA NAGAR – 522 510

Ph: 0863-2293299, 2293214, ,Cell:9848477441
0863-2346259 (Study Material)
Website: www.anucde.info
e-mail:anucdedirector@gmail.com

M.L.I.Sc.,

First Edition : 2021

No. of Copies :

Published by :
Dr. NAGARAJU BATTU,
Director
Centre for Distance Education,
Acharya Nagarjuna University

Printed at :

FOREWORD

Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A' grade from the NAAC in the year 2016, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over 443 affiliated colleges spread over the two districts of Guntur and Prakasam.

The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the door step of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.A., and B.Com courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.

To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.

It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson- writers of the Centre who have helped in these endeavors.

Prof. P. Raja Sekhar
Vice-Chancellor (FAC)
Acharya Nagarjuna University

INFORMATION PROCESSING RETRIEVAL

M.L.I.Sc., Semester – I, Paper-III

Syllabus

Objectives :

1. To make the students aware with the latest developments and trends in the field of advanced library classification.
2. To train the students in the practical application of Universal Decimal Classification.
3. To acquaint the students with recent developments in computerized bibliographic records and communication formats
4. To train the students in the cataloguing of non-book materials and complex serial publications according to AACR-2, 1988 revised edition

UNIT I
Information Processing and Retrieval - Concept – An overview of UDC & BSO

UNIT II
Indexing Languages: Characteristics, Types – Vocabulary Control: Thesaurus Construction

UNIT III
Bibliographic Standards and formats - ISBD, AACR, CCF, MARC21, ISO 2709 Metadata-Dublin Core.

UNIT IV
Indexing Systems- Types – Search Statement and Search Process - Search Strategies – Tools, Techniques and Methods

UNIT V
Evaluation of Information Retrieval System– Parameters for Evaluation Information Retrieval System – Information Retrieval Models.

Books for study and reference:

1. Fosket, A.C. Subject approach to Information. 5th Rev. Ed. London, Bingley, 1996
2. Lancaster, F.W. Indexing and Abstracting in Theory and Practice. 2nd Ed. London, Lib. Assoc.,1998
3. Satyanarayana, V.V.V. Universal Decimal Classification: A Practical Primer. New Delhi, Ess Pub, 1998
4. UDC Consortium. Universal Decimal Classification, International MEDIUM Edition, 1993.
5. Raju, A.A.N. Universal Decimal Classification IME 1993: Theory and practice (A self instructional manual). Delhi, Ess Ess Publications, 2007
6. Soma Raju, P. Universal Decimal Classification IME 1993. Visakhapatnam, Author,

19971.

7. A course in Information consolidation: a handbook for education and training in analysis, synthesis and repackaging of Information. General Information Programme and NISIST, UNESCO, PGI, Paris. 1986.
8. Alberico, R. and Micco M.(1990). Expert systems for reference and Information retrieval. West Port : Meckler.
9. Atchison, J. & Alan G. A.(1072). Thesaurus construction: a practical manual. London: Aslib.
10. Atchison, J. & Gilchrist, A.(1972). Thesaurus construction: a practical manual. London: Aslib.
11. Austin, D.(1984). PRECIS: A manual of concept analysis and subject Indexing. 2nd ed.
1. Chowdhruy, G.G.(2003). Introduction to modern Information retrieval. 2nd Ed. London, Facet Publishing.
12. Ghosh, S.B. and Biswas, S.C. (1998). Subject Indexing systems: Concepts, methods and techniques. Rev. ed. Calcutta: IASLIC.
13. Lancaster, F. W. (1968). Information retrieval systems, characteristics, testing and evaluation. London: Facet Publishing.
14. Pandey, S.K. Ed.(2000).Library Information retrieval. New Delhi: Anmol.
15. Vickery, B.C.(1970). Techniques of Information retrieval. London: Butterworths
16. Kumar, P.S.G.(2004). Information Analysis, repackaging, consolidation and information retrieval.Delhi: B.R.Publishing.
17. Kumar, P.S.G. (2002) : A student manual of library and infoemation science. Delhi: B.R. publishing.
18. Narayana,G.J.(1991).Library and Information Management.New Delhi:Prientice Hall.
19. Information today and tomorrow (Hyderabad) (1999). Working papers. Delhi: NISSAT.

INFORMATION PROCESSING RETRIEVAL

CONTENTS

LESSON - 1

INFORMATION PROCESSING AND RETRIEVAL

AIMS AND OBJECTIVES

The objective of this lesson is to explain Information processing and retrieval concept, definition and its development. This lesson also explains design of Information Storage and Retrieval systems. Different types of Information Storage and Retrieval Systems are also discussed.

After studying this lesson you can understand

- What is Information Processing and Retrieval Systems?
- Objectives of Information Storage and Retrieval systems.
- Factors influencing the design of ISARS.
- Types of Information Storage and Retrieval systems and
- Components of ISARS

Structure

1.1 INTRODUCTION

Information is an essential ingredient in decision making. The need for Information systems in recent years has been made critical by the steady growth in size and complexity of organizations and data. As a matter of fact it has been said that information has a synergising effect in several areas of human activities.

The information is recorded and stored only when it is expected to have potential importance. But, its importance may decline with time necessitating a weeding process. Information may be available but not accessible which is more frustrating than having no information. Information professionals have resorted to organisation of information into some meaningful groupings of topics for easy retrieval as and when necessary**.**

1.2 DEFINITION

The "Information processing and retrieval" can be easily understood when we define and delimit the use of the terms in the concept and consider some of the resulting implications. This is especially because the meaning of terms depends upon the context in which they are used.

Information: Information is some meaningful message recorded in conventional or non-conventional media and stored and processed by systems and services with a view to providing a more or less permanent memory of the messages and their dissemination to users.

Processing: The term processing involves the important input – storage – output of information. It also involves necessary computer programmed operations. Processing also involves numerous other non-computer technologies such as reprography, communication etc.

According to Doyel (1975) processing can be thought of as a family operations, each of which acts uniformly on every iten of information presented and includes the idea of having the items in a place available and in a form available for ultimate use,

According to Ranganathan, processing includes both technical processing and book preparation for display on the shelves. So, Technical processing involves classifying and cataloguing the documents as well as preparation of documents which involves stamping, labelling, numbering etc.

Retrieval: According to "Harrod's Librarian's Glossary". Retrieval is defined as

- The act of finding again, recovery, retrospective search and securing of documents. The act of going to a specific location or area where the document is available and retrieving the same.
- The act and means of obtaining
- Facts and other information which is recorded and indexed in some way by subject.
- The documents contain the required facts

Information Retrieval: According to Salton and Magil (1983) Information Retrieval is concerned with the representation, storage, organization and accessing of Information items.

According to Lancaster (1979) Information Retrieval is really synonymous with literature searching. It is the process of searching some collection of documents using the terms dictionary in its widest sense, in order to identify those documents which deal with a particular subject.

1.3 OBJECTIVES OF INFORMATION STORAGE AND RETRIEVAL SYSTEMS

The main objective of any Information storage and retrieval system is to provide right information to the right user at the right time.

According to M.L. Pao (1989) an Information storage and retrieval system shoud consider the following objectives:

a) **Information content:** Information Storage and Retrieval system exist to provide information to its users. It may be subject oriented or mission oriented. For example the AGRIS Information storage and retrieval system contain all information relating to agriculture. Similarly INIS contain information pertaining to nuclear science.

b) **Utility:** Utility of the ISARS depends up on the continuous acquisition of up to date information resources and sophisticated indexing and abstracting services offered by the system. The system must be capable of providing CAS and SDI services.

c) **Users:** The ISARS should identify actual users and its intended users while designing the system.

d) **Documentary Resources:** The ISARS should identify the subject coverage, types of documents, period and languages of documents etc. Need to be identified and decided by the designers of ISARS.

e) **Performance criteria:** In order to make the ISARS cost effective different performance criteria such as subject coverage, types of documents included and indexing techniques used etc. are the key elements to design ISARS. The system must provide pinpointed, exhaustive and expeditious service which is the demand of the user.

f) **Economics:** Cost is the deciding factor in designing any ISARS. However most of the services provided by the system are qualitative and not quantifiable to assess the cost-effectiveness. The ISARS should be designed and operated with societal value rather than profit motive.

1.4 FACTORS INFLUENCING THE DESIGN OF ISARS

The factors that influence the design of the ISARS are as follows:

- Users requirement
- Types of documents which meets the requirement of users
- Source of data – internal or external or both
- Compatibility of the system if the system needs to exchange data with others
- Kinds of documents to be included in the system. For example monograph, theses, patents, reports, journal articles etc.
- level of bibliographic description.
- search techniques adopted by the system.

1.5 TYPES OF INFORMATION STORAGE AND RETRIEVAL SYSTEMS

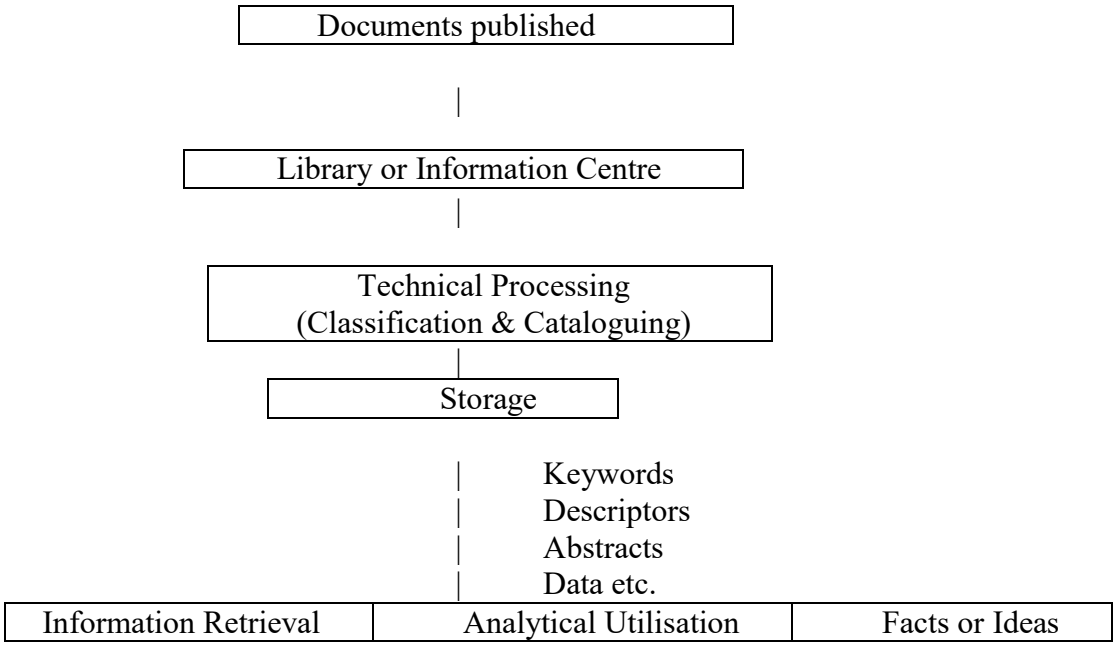
Information storage and retrieval systems may be grouped into two types based on the content of the system

1.5.1 Reference Retrieval System: This type of Information Storage and Retrieval System (ISRS) contains the database of records of documents instead of documents itself. Each record is made of with brief description of the document and its location in the library or Information centre. For example if the record is of the book, it includes the details such as its title, author/s, publication details, collation, subject entries, call number etc. If the database is of journal articles, it include the title, author and abstract details along with journal name, volume details, page numbers and year of publication. So reference retrieval systems are those systems which help to search, retrieve and locate the documents which contain the information required.

1.5.2 Data Retrieval Systems: This type of Information Storage and Retrieval System (ISRS) contains the database of records of actual information such as physical or chemical or other types of data. These systems directly provide information required by the users, while reference retrieval systems provide details about containers of information. This type of systems directly answers the user's query, as these stores data itself rather than data surrogates. The stored data may be referred to as a "data bank".

1.6 INFORMATION STORAGE AND RETRIEVAL PROCESS

In any Library or Information centre the Information storage and retrieval system (ISRS) involves the processes such as the creation, searching, and modification and retrieval of stored data. The following figure explains the process of an ISRS in a schematic and simplified way:



Lancaster identifies two types of Information retrieval activities:

1) Question-answering service i.e. services that attempt to produce the answer to a particular factual question. The result delivered to the user is the answer to the question posed to the system together with an indication of the source in which the answer was found.

2) Literature search services, i.e. services that attempts to identify documents that deal with a particular subject area of interest to the person requesting the search. The result delivered to the user will sometimes be a document or group of documents dealing with the subject matter sought.

The process in information storage and retrieval system starts from the moment the document is accessioned by the library or information centre. The documents are then classified, catalogued, indexed and database of bibliographic records or actual information itself is stored to facilitate immediate retrieval as and when required.

The retrieval and dissemination of information originates by a query put to the system by user. Sometimes the query is already built into the system through Selective Dissemination of Information (SDI), wherein the subscribers of SDI Service with their pertinent interest profiles receive bibliography or extracts or the full information.

Besides providing bibliographic and/or information to the user depending upon the query, the Information Storage and Retrieval System also provides state-of art-reports, trend reports etc. by evaluating, analysing and synthesising the information received by the system. The final step of any ISRS is feed back to assess the satisfaction of user and to assess the usage and performance of the system.

1.7 COMPONENTS OF INFORMATION STORAGE AND RETRIEVAL SYSTEM

Following are the various components of ISARS

Document selection
Indexing
Vocabulary control
Data input and validation
Searching
Output and
Usage

1.7.1 Document selection

Documents to be selected for inclusion in the system depend up on the scope of the system. Subject coverage and kinds of documents (books, journals, theses etc.) to be selected must meet the objective of the system.

1.7.2 Indexing

This sub-system deal with indexing techniques used natural language or assigned indexing. The designer of the system must decide different options such as: whole field, sub field, any word or phrase etc.

1.7.3 Vocabulary control

The system must provide online thesaurus to be used while indexing as well as searching the system for documents. This thesaurus helps the user in broadening or narrowing the query depending up on the documents retrieved.

1.7.4 Data input and validation

Data can be entered offline or online. In online interactive mode the data is usually entered through work sheets. A set of rules are embedded in the work sheet to validate the data entered. A set of authority files may also be used for consistency in indexing and to validate the data entered.

1.7.5 Searching

The ISARS must support various search techniques such as:

- single key search
- multiple key search with Boolean logic
- free text search
- wild card search
- truncated search
- proximity search
- browsing
- search refining by combining search sets
- searching within the search results
- sorting the records of search results
- profile matching in batch mode for providing SDI

1.7.6 Output

The system should allow the user to get the results of the search in the desired form. One or more types of output formats designed and built into the system for the end user. The system can also provide facility to design the output format as required by the user.

1.7.7 Usage

The system must be so designed to monitor the users, their logs, timings, search terms and strategies used, the number of records retrieved and so on. Statistical analysis of these details will help to assess the usage of the system and improve the performance of the system.

1.8 SUMMARY

Information Retrieval Systems are concerned with selection, processing, storage and retrieving of information itself or information containers to meet the requirements of the users. The Information Retrieval systems may be manual or computer-based systems. For example catalogue is a manual Information Retrieval System. As the libraries have started using the information and communication technologies, computer-based Information Storage and Retrieval Systems came into existence. For example Online Public Access Catalogue (OPAC) is one such computer-based Information Storage and Retrieval System. The Information Storage and Retrieval Systems may be Reference Retrieval Systems, Document Retrieval systems and Fact retrieval systems depending upon the content it serves. The Information Storage and Retrieval System consist of sub-systems such as selection subsystem, input and data validation subsystem, processing subsystem, search subsystem, output subsystem, and a system usage monitoring subsystem. The performance of the Information Storage and Retrieval

System depends up on the indexing techniques used, vocabulary control adopted, knowledge of the subject covered by the system and needs of users.

1.9 TECHNICAL TERMS

ISARS : Information storage and Retrieval
SDI : Selective Dissemination of Information
OPAC : Online Public Access Catalogue

1.10 SUGGESTED READINGS

1. Lancaster, F.W. Vocabulary control for Information Retrieval. Washington: Information Resources Press, 1972.
2. Pao, M.L. Concepts of Information Retrieval Systems. Englewood, Colombus: Libraries Unlimited, 1989.
3. Satyanarayana, B. et al.Information Technology : issues and trends. Delhi : Vedam books, 1998
4. Vickery, Brian C. And Alan Vickery. Information science in theory and practice. London: Butterworths, 1987

LESSON 2

UNIVERSAL DECIMAL CLASSIFICATION

AIMS AND OBJECTIVES

The objective of this lesson is to explain Universal Decimal Classification. The outline of UDC also given for use in classifying the library documents

After studying this lesson you can understand

- What is UDC
- Basic features and structure of UDC
- Application of UDC to classify library documents

Structure

2.1 INTRODUCTION

The **Universal Decimal Classification** (UDC) is a bibliographic and library classification developed by the Belgian bibliographers Paul Otlet and Henri La Fontaine at the end of the 19th century. UDC provides a systematic arrangement of all branches of human knowledge organized as a coherent system in which knowledge fields are related and inter-linked.

Originally based on the Dewey Decimal Classification, the UDC was developed as a new analytico-synthetic classification system with a significantly larger vocabulary and syntax that enables very detailed content indexing and information retrieval in large

collections. In its first edition in 1905, the UDC already included many features that were revolutionary in the context of knowledge classifications: tables of generally applicable (aspect-free) concepts - called common auxiliary tables; a series of special auxiliary tables with specific but re-usable attributes in a particular field of knowledge; an expressive notational system with connecting symbols and syntax rules to enable coordination of subjects and the creation of a documentation language proper. Although originally designed as an indexing and retrieval system, due to its logical structure and scalability, UDC has become one of the most widely used knowledge organization systems in libraries, where it is used for shelf arrangement, content indexing or both. UDC codes can describe any type of document or object to any desired level of detail. These can include textual documents and other media such as films, video and sound recordings, illustrations, maps as well as realia such as museum objects.

The first edition of UDC in French "Manuel du Répertoire bibliographique universel" was published in the year 1905. Since then, UDC has been translated and published in various editions in 40 languages. UDC Summary, an abridged Web version of the scheme is available in over 50 languages.[10] The classification has been modified and extended over the years to cope with increasing output in all areas of human knowledge, and is still under continuous review to take account of new developments.

2.2 APPLICATION OF UDC

UDC is used in around 150,000 libraries in 130 countries and in many bibliographical services which require detailed content indexing. In a number of countries it is the main classification system for information exchange and is used in all type of libraries: public, school, academic and special libraries. UDC is also used in national bibliographies of around 30 countries. Examples of large databases indexed by UDC include:

NEBIS (The Network of Libraries and Information Centers in Switzerland) - 2.6 million records

COBIB.SI (Slovenian National Union Catalogue) - 3.5 million records

Hungarian National Union Catalogue (MOKKA) - 2.9 million records

VINITI RAS database (All-Russian Scientific and Technical Information Institute of Russian Academy of Science) with 28 million records

Meteorological & Geoastrophysical Abstracts (MGA) with 600 journal titles

PORBASE (Portuguese National Bibliography) with 1.5 million records

UDC has traditionally been used for the indexing of scientific articles which was an important source of information of scientific output in the period predating electronic publishing. Collections of research articles in many countries covering decades of scientific output contain UDC codes. Examples of journal articles indexed by UDC

2.3 UDC STRUCTURE

2.3.1 Notation

A notation is a code commonly used in classification schemes to represent a class, i.e. a subject and its position in the hierarchy, to enable mechanical sorting and filing of subjects. UDC uses Arabic numerals arranged decimally. Every number is thought of as a decimal fraction with the initial decimal point omitted, which determines the filing order. An advantage of decimal notational systems is that they are infinitely extensible, and when new subdivisions are introduced, they need not disturb the existing allocation of numbers. For ease of reading, a UDC notation is usually punctuated after every third digit:

Notation	Caption (Class description)
539.120	Theoretical problems of elementary particles physics. Theories and models of fundamental interactions
539.120.2	Symmetries of quantum physics
539.120.22	Conservation laws
539.120.222	Translations. Rotations
539.120.224	Reflection in time and space
539.120.226	Space-time symmetries
539.120.23	Internal symmetries
539.120.3	Currents
539.120.4	Unified field theories
539.120.5	Strings

In UDC the notation has two features that make the scheme easier to browse and work with:

- **Hierarchically expressive** - the longer the notation, the more specific the class: removing the final digit automatically produces a broader class code.
- **Syntactically expressive** - when UDC codes are combined, the sequence of digits is interrupted by a precise type of punctuation sign which indicates that the expression is a combination of classes rather than a simple class

  e.g. the colon in 34:32 indicates that there are two distinct notational elements: 34 Law. Jurisprudence and 32 Politics;

  the closing and opening parentheses and double quotes in the following code 913(574.22)"19"(084.3) indicate four separate notational elements: 913 Regional geography, (574.22) North Kazakhstan (Soltüstik Qazaqstan); "19" 20th century and (084.3) Maps (document form)

2.3.2 Basic features and syntax

UDC is an analytico-synthetic and faceted classification. It allows an unlimited combination of attributes of a subject and relationships between subjects to be expressed. UDC codes from different tables can be combined to present various aspects of document content and form, e.g. 94(410)"19"(075) History *(main subject)* of United Kingdom *(place)* in 20th century *(time)*, a textbook *(document form)*. Or: 37:2 Relationship between Education and Religion. Complex UDC expressions can be accurately parsed into constituent elements.

UDC is also a disciplinary classification covering the entire universe of knowledge. This type of classification can also be described as *aspect* or *perspective*, which means that concepts are subsumed and placed under the field in which they are studied. Thus, the same concept can appear in different fields of knowledge. This particular feature is usually implemented in UDC by re-using the same concept in various combinations with the main subject, e.g. a code for language in common auxiliaries of language is used to derive numbers for ethnic grouping, individual languages in linguistics and individual literatures. Or, a code from the auxiliaries of place, e.g.*(410) United Kingdom*, uniquely representing the concept of United Kingdom can be used to

express *911(410) Regional geography of United Kingdom* and *94(410) History of United Kingdom*.

2.3.3 Organization of classes

Concepts are organized in two kinds of tables in UDC viz. Common auxiliary tables and Main tables

2.3.3.1 Common auxiliary tables:

These tables contain facets of concepts representing, general recurrent characteristics, applicable over a range of subjects throughout the main tables, including notions such as place, language of the text and physical form of the document, which may occur in almost any subject. UDC numbers from these tables, called common auxiliaries are simply added at the end of the number for the subject taken from the main tables. There are over 15,000 of common auxiliaries in UDC.

2.3.3.2 The main tables or main schedules:

These tables are meant for accommodating the various disciplines and branches of knowledge, arranged in 9 main classes, numbered from 0 to 9 (with class 4 being vacant). At the beginning of each class there are also series of special auxiliaries, which express aspects that are recurrent within this specific class. Main tables in UDC contain more than 60,000 subdivisions.

2.4. MAIN CLASSES

- Science and Knowledge. Organization. Computer-Science. Information Science. Documentation. Librarianship. Institutions. Publications
- Philosophy. Psychology
- Religion. Theology
- Social Sciences
- *vacant*
- Mathematics. Natural Sciences
- Applied Sciences. Medicine, Technology
- The Arts. Entertainment. Sport
- Linguistics. Literature
- Geography. History

The vacant class 4 is the result of a planned schedule expansion. This class was freed by moving linguistics into class 8 in the 1960s to make space for future developments in the rapidly expanding fields of knowledge; primarily natural sciences and technology.

2.5. COMMON AUXILIARY TABLES

Common auxiliaries are aspect-free concepts that can be used in combination with any other UDC code from the main classes or with other common auxiliaries. They have unique notational representations that make them stand out in complex expressions. Common auxiliary numbers always begin with a certain symbol known as a facet indicator, e.g. = (equal sign) always introduces concepts representing the language of a

document; (0...) numbers enclosed in parentheses starting with zero always represent a concept designating document form. Thus (075) Textbook and =111 English can be combined to express, e.g.(075)=111 Textbooks in English, and when combined with numbers from the main UDC tables they can be used as follows: 2(075)=111 Religion textbooks in English, 51(075)=111 Mathematics textbooks in English etc.

- =... Common auxiliaries of language. Table 1c
- (0...) Common auxiliaries of form. Table 1d
- (1/9) Common auxiliaries of place. Table 1e
- (=...) Common auxiliaries of human ancestry, ethnic grouping and nationality. Table 1f
- "..." Common auxiliaries of time. Table 1g helps to make minute division of time e.g.: "1993-1996
- -0... Common auxiliaries of general characteristics: Properties, Materials, Relations/Processes and Persons. Table 1k.
- -02 Common auxiliaries of properties. Table 1k
- -03 Common auxiliaries of materials. Table 1k
- -04 Common auxiliaries of relations, processes and operations. Table 1k
- -05 Common auxiliaries of persons and personal characteristics. Table 1k this table is repeated

2.6. CONNECTING SIGNS

In order to preserve the precise meaning and enable accurate parsing of complex UDC expressions a number of connecting symbols are made available to relate and extend UDC numbers. These are:

Symbol	Symbol name	Meaning	Example
+	plus	coordination, addition	e.g. 59+636 zoology and animal breeding
/	stroke	consecutive extension	e.g. 592/599 Systematic zoology (everything from 592 to 599 inclusive)
:	colon	relation	e.g. 17:7 Relation of ethics to art
[ ]	Square brackets	subgrouping	e.g. 311:[622+669](485) statistics of mining and metallurgy in Sweden (the auxiliary qualifiers 622+669 considered as a unit)
*	asterisk	Introduces non-UDC notation	e.g. 523.4*433 Planetology, minor planet Eros (IAU authorized number after the asterisk)
A/Z	alphabetical extension	Direct alphabetical specification	e.g. 821.133.1MOL French literature, works of Molière

2.7 UDC OUTLINE

UDC classes in this outline are taken from the Multilingual Universal Decimal Classification Summary released on the by the UDC Consortium under the Creative Commons Attribution Share Alike 3.0 license (first release 2009, subsequent update 2012).

Main Tables

Science and knowledge. Organization. Computer science. Information. Documentation. Librarianship. Institution. Publication

00 Prolegomena. Fundamentals of knowledge and culture. Propaedeutics
001 Science and knowledge in general. Organization of intellectual work
002 Documentation. Books. Writings. Authorship
003 Writing systems and scripts
004 Computer science and technology. Computing
004.2 Computer architecture
004.3 Computer hardware
004.4 Software
004.5 Human-computer interaction
004.6 Data
004.7 Computer communication
004.8 Artificial intelligence
004.9 Application-oriented computer-based techniques
005 Management
005.1 Management Theory
005.2 Management agents. Mechanisms. Measures
005.3 Management activities
005.5 Management operations. Direction
005.6 Quality management. Total quality management (TQM)
005.7 Organizational management (OM)
005.9 Fields of management
005.92 Records management
005.93 Plant management. Physical resources management
005.94 Knowledge management
005.95/.96 Personnel management. Human Resources management
006 Standardization of products, operations, weights, measures and time
007 Activity and organizing. Information. Communication and control theory Genallay (Cybernetics)
008 Civilization. Culture. Progress
01 Bibliography and bibliographies. Catalogues
02 Librarianship
030 General reference works (as subject)
050 Serial publications, periodicals (as subject)
06 Organizations of a general nature
069 Museums
070 Newspapers (as subject). The Press. Outline of journalism
08 Polygraphies. Collective works (as subject)
09 Manuscripts. Rare and remarkable works (as subject)

1 Philosophy. Psychology

101 Nature and role of philosophy
11 Metaphysics
111 General metaphysics. Ontology
122/129 Special Metaphysics
13 Philosophy of mind and spirit. Metaphysics of spiritual life
14 Philosophical systems and points of view
159.9 Psychology
159.91 Psychophysiology (physiological psychology). Mental physiology
159.92 Mental development and capacity. Comparative psychology
159.93 Sensation. Sensory perception
159.94 Executive functions
159.95 Higher mental processes
159.96 Special mental states and processes
159.97 Abnormal psychology
159.98 Applied psychology (psychotechnology) in general
16 Logic. Epistemology. Theory of knowledge. Methodology of logic
17 Moral philosophy. Ethics. Practical philosophy

2 Religion. Theology

The UDC tables for religion are fully faceted. Indicated in italics below, are special auxiliary numbers that can be used to express attributes (facets) of any specific faith. Any special number can be combined with any religion e.g. *-5 Worship* can be used to express e.g. *26-5 Worship in Judaism*, *27-5 Worship in Christianity*,*24-5 Worship in Buddhism*. The complete special auxiliary tables contain around 2000 subdivisions of various attributes that can be attached to express various aspects of individual faiths to a great level of specificity allowing equal level of detail for every religion.

2-1/-9 Special auxiliary subdivision for religion
2-1 Theory and philosophy of religion. Nature of religion. Phenomenon of religion
2-2 Evidences of religion
2-3 Persons in religion
2-4 Religious activities. Religious practice
2-5 Worship broadly. Cult. Rites and ceremonies
2-6 Processes in religion
2-7 Religious organization and administration
2-8 Religions characterised by various properties
2-9 History of the faith, religion, denomination or church
21/29 Religious systems. Religions and faiths
21 Prehistoric and primitive religions
22 Religions originating in the Far East
23 Religions originating in Indian sub-continent. Hindu religion in the broad sense
24 Buddhism
25 Religions of antiquity. Minor cults and religions
26 Judaism
27 Christianity
28 Islam

29 Modern spiritual movements

3 Social Sciences

303 Methods of the social sciences
304 Social questions. Social practice. Cultural practice. Way of life (Lebensweise)
305 Gender studies
308 Sociography. Descriptive studies of society (both qualitative and quantitative)
311 Statistics as a science. Statistical theory
314/316 Society
314 Demography. Population studies
316 Sociology
32 Politics
33 Economics. Economic science
34 Law. Jurisprudence
35 Public administration. Government. Military affairs
36 Safeguarding the mental and material necessities of life
37 Education
39 Cultural anthropology. Ethnography. Customs. Manners. Traditions. Way of life

4 Vacant

This section is currently vacant.

5 Mathematics. Natural sciences

502/504 Environmental science. Conservation of natural resources. Threats to the environment and protection against them
502 The environment and its protection
504 Threats to the environment
51 Mathematics
510 Fundamental and general considerations of mathematics
511 Number theory
512 Algebra
514 Geometry
517 Analysis
519.1 Combinatorial analysis. Graph theory
519.2 Probability. Mathematical statistics
519.6 Computational mathematics. Numerical analysis
519.7 Mathematical cybernetics
519.8 Operational research (OR): mathematical theories and methods
52 Astronomy. Astrophysics. Space research. Geodesy
53 Physics

531/534 Mechanics
535 Optics
536 Heat. Thermodynamics. Statistical physics
537 Electricity. Magnetism. Electromagnetism
538.9 Condensed matter physics. Solid state physics
539 Physical nature of matter
54 Chemistry. Crystallography. Mineralogy
542 Practical laboratory chemistry. Preparative and experimental chemistry
543 Analytical chemistry
544 Physical chemistry
546 Inorganic chemistry
547 Organic chemistry
548/549 Mineralogical sciences. Crystallography. Mineralogy
55 Earth Sciences. Geological sciences
56 Palaeontology
57 Biological sciences in general
58 Botany
59 Zoology

6 Applied sciences. Medicine. Technology

Class 6 occupies the largest proportion of UDC schedules. It contains over 44,000 subdivisions. Each specific field of technology or industry usually contains more than one special auxiliary table with concepts needed to express operations, processes, materials and products. As a result, UDC codes are often created through the combination of various attributes. Equally, some parts of this class enumerate concepts to a great level of detail e.g. *621.882.212 Hexagon screws with additional shapes. Including: Flank screws. Collar screws. Cap screws*

60 Biotechnology
61 Medical sciences
611/612 Human biology
613 Hygiene generally. Personal health and hygiene
614 Public health and hygiene. Accident prevention
615 Pharmacology. Therapeutics. Toxicology
616 Pathology. Clinical medicine
617 Surgery. Orthopaedics. Ophthalmology
618 Gynaecology. Obstetrics
62 Engineering. Technology in general
620 Materials testing. Commercial materials. Power stations. Economics of energy
621 Mechanical engineering in general. Nuclear technology. Electrical engineering. Machinery
622 Mining
623 Military engineering
624 Civil and structural engineering in general
625 Civil engineering of land transport. Railway engineering. Highway engineering

626/627 Hydraulic engineering and construction. Water (aquatic) structures
629 Transport vehicle engineering
63 Agriculture and related sciences and techniques. Forestry. Farming. Wildlife exploitation
630 Forestry
631/635 Farm management. Agronomy. Horticulture
633/635 Horticulture in general. Specific crops
636 Animal husbandry and breeding in general. Livestock rearing. Breeding of domestic animals
64 Home economics. Domestic science. Housekeeping
65 Communication and transport industries. Accountancy. Business management. Public relations
654 Telecommunication and telecontrol (organization, services)
655 Graphic industries. Printing. Publishing. Book trade
656 Transport and postal services. Traffic organization and control
657 Accountancy
658 Business management, administration. Commercial organization
659 Publicity. Information work. Public relations
66 Chemical technology. Chemical and related industries
67 Various industries, trades and crafts
68 Industries, crafts and trades for finished or assembled articles
69 Building (construction) trade. Building materials. Building practice and procedure

7 The arts. Recreation. Entertainment. Sport

7.01/.09 Special auxiliary subdivision for the arts
7.01 Theory and philosophy of art. Principles of design, proportion, optical effect
7.02 Art technique. Craftsmanship
7.03 Artistic periods and phases. Schools, styles, influences
7.04 Subjects for artistic representation. Iconography. Iconology
7.05 Applications of art (in industry, trade, the home, everyday life)
7.06 Various questions concerning art
7.07 Occupations and activities associated with the arts and entertainment
7.08 Characteristic features, forms, combinations etc. (in art, entertainment and sport)
7.091 Performance, presentation (in original medium)
71 Physical planning. Regional, town and country planning. Landscapes, parks, gardens
72 Architecture
73 Plastic arts
74 Drawing. Design. Applied arts and crafts
745/749 Industrial and domestic arts and crafts. Applied arts
75 Painting
76 Graphic art, printmaking. Graphics

77 Photography and similar processes
78 Music
79 Recreation. Entertainment. Games. Sport
791 Cinema. Films (motion pictures)
792 Theatre. Stagecraft. Dramatic performances
793 Social entertainments and recreations. Art of movement. Dance
794 Board and table games (of thought, skill and chance)
796 Sport. Games. Physical exercises
797 Water sports. Aerial sports
798 Riding and driving. Horse and other animal sports
799 Sport fishing. Sport hunting. Shooting and target sports

8 Language. Linguistics. Literature

Tables for class 8 are fully faceted and details are expressed through combination with common auxiliaries of language (Table 1c) and a series of special auxiliary tables to indicate other facets or attributes in Linguistics or Literature. As a result, this class allows for great specificity in indexing although the schedules themselves occupy very little space in UDC. The subdivisions of e.g. *811 Languages* or *821 Literature* are derived from common auxiliaries of language =1/=9 (Table 1c) by substituting a point for the equals sign, e.g. 811.111 English language (as a subject of a linguistic study) and *821.111 English literature* derives from=*111 English language*. Common auxiliaries of place and time are also frequently used in this class to express place and time facets of Linguistics or Literature, e.g. *821.111(71)"18" English literature of Canada in 19th century*

80 General questions relating to both linguistics and literature. Philology
801 Prosody. Auxiliary sciences and sources of philology
808 Rhetoric. The effective use of language
81 Linguistics and languages
81`1/`4 Special auxiliary subdivision for subject fields and facets of linguistics and languages
81`1 General linguistics
81`2 Theory of signs. Theory of translation. Standardization. Usage. Geographical linguistics
81`3 Mathematical and applied linguistics. Phonetics. Graphemics. Grammar. Semantics. Stylistics
81`4 Text linguistics, Discourse analysis. Typological linguistics
81`42 Text linguistics. Discourse analysis
81`44 Typological linguistics
811 Languages

Derived from the common auxiliaries of language =1/=9 (Table 1c) by replacing the equal sign = with prefix *811.* e.g. *=111 English* becomes *811.111* Linguistics of English language

811.1/.9 All languages natural or artificial
811.1/.8 Individual natural languages
811.1/.2 Indo-European languages
811.21/.22 Indo-Iranian languages
811.3 Dead languages of unknown affiliation. Caucasian languages

811.4 Afro-Asiatic, Nilo-Saharan, Congo-Kordofanian, Khoisan languages
811.5 Ural-Altaic, Palaeo-Siberian, Eskimo-Aleut, Dravidian and Sino-Tibetan languages. Japanese. Korean. Ainu
811.6 Austro-Asiatic languages. Austronesian languages
811.7 Indo-Pacific (non-Austronesian) languages. Australian languages
811.8 American indigenous languages
811.9 Artificial languages

82 Literature

82-1/-9 Special auxiliary subdivision for literary forms, genres
82-1 Poetry. Poems. Verse
82-2 Drama. Plays
82-3 Fiction. Prose narrative
82-31 Novels. Full-length stories
82-32 Short stories. Novellas
82-4 Essays
82-5 Oratory. Speeches
82-6 Letters. Art of letter-writing. Correspondence. Genuine letters
82-7 Prose satire. Humour, epigram, parody
82-8 Miscellanea. Polygraphies. Selections
82-9 Various other literary forms
82-92 Periodical literature. Writings in serials, journals, reviews
82-94 History as literary genre. Historical writing. Historiography. Chronicles. Annals. Memoirs
82.02/.09 Special auxiliary subdivision for theory, study and technique literature
82.09 Literary criticism. Literary studies
82.091 Comparative literary studies. Comparative literature
821 Literatures of individual languages and language families

Derived from the common auxiliaries of language =1/=9 (Table 1c) by replacing the equal sign = with prefix *821.* e.g. *=111* English becomes 821.111 English literature

9 Geography. Biography. History

Tables for Geography and History in UDC are fully faceted and place, time and ethnic grouping facets are expressed through combination with common auxiliaries of place (Table 1d), ethnic grouping (Table 1f) and time (Table 1g)

902/908 Archaeology. Prehistory. Cultural remains. Area studies
902 Archaeology
903 Prehistory. Prehistoric remains, artefacts, antiquities
904 Cultural remains of historical times
908 Area studies. Study of a locality
91 Geography. Exploration of the Earth and of individual countries. Travel. Regional geography
910 General questions. Geography as a science. Exploration. Travel
911 General geography. Science of geographical factors (systematic geography). Theoretical geography
911.2 Physical geography
911.3 Human geography (cultural geography). Geography of cultural factors

911.5/.9 Theoretical geography
912 Nonliterary, nontextual representations of a region
913 Regional geography
92 Biographical studies. Genealogy. Heraldry. Flags
929 Biographical studies
929.5 Genealogy
929.6 Heraldry
929.7 Nobility. Titles. Peerage
929.9 Flags. Standards. Banners
93/94 History
930 Science of history. Historiography
930.1 History as a science
930.2 Methodology of history. Ancillary historical sciences
930.25 Archivistics. Archives (including public and other records)
930.85 History of civilization. Cultural history
94 General

Common Auxiliary Tables

Common auxiliaries of language. Table 1c

=1/=9 Languages (natural and artificial)
=1/=8 Natural languages
=1/=2 Indo-European languages
=1 Indo-European languages of Europe
=11 Germanic languages
=12 Italic languages
=13 Romance languages
=14 Greek (Hellenic)
=15 Celtic languages
=16 Slavic languages
=17 Baltic languages
=18 Albanian
=19 Armenian
=2 Indo-Iranian, Nuristani (Kafiri) and dead Indo-European languages
=21/=22 Indo-Iranian languages
=21 Indic languages
=22 Iranian languages
=29 Dead Indo-European languages (not listed elsewhere)
=3 Dead languages of unknown affiliation. Caucasian languages
=34 Dead languages of unknown affiliation, spoken in the Mediterranean and Near East (except Semitic)
=35 Caucasian languages
=4 Afro-Asiatic, Nilo-Saharan, Congo-Kordofanian, Khoisan languages
=41 Afro-Asiatic (Hamito-Semitic) languages
=42 Nilo-Saharan languages
=43 Congo-Kordofanian (Niger-Kordofanian) languages

=45 Khoisan languages
=5 Ural-Altaic, Palaeo-Siberian, Eskimo-Aleut, Dravidian and Sino-Tibetan languages. Japanese. Korean. Ainu
=51 Ural-Altaic languages
=521 Japanese
=531 Korean
=541 Ainu
=55 Palaeo-Siberian languages
=56 Eskimo-Aleut languages
=58 Sino-Tibetan languages
=6 Austro-Asiatic languages. Austronesian languages
=61 Austro-Asiatic languages
=62 Austronesian languages
=7 Indo-Pacific (non-Austronesian) languages. Australian languages
=71 Indo-Pacific (non-Austronesian) languages
=72 Australian languages
=8 American indigenous languages
=81 Indigenous languages of Canada, USA and Northern-Central Mexico
=82 Indigenous languages of western North American Coast, Mexico and Yucatán
=84/=88 Central and South American indigenous languages
=84 Ge-Pano-Carib languages. Macro-Chibchan languages
=85 Andean languages. Equatorial languages
=86 Chaco languages. Patagonian and Fuegian languages
=88 Isolated, unclassified Central and South American indigenous languages
=9 Artificial languages
=92 Artificial languages for use among human beings. International auxiliary languages (interlanguages)
=93 Artificial languages used to instruct machines. Programming languages. Computer languages

(0...) Common auxiliaries of form. Table 1d

(0.02/.08) Special auxiliary subdivision for document form
(0.02) Documents according to physical, external form
(0.03) Documents according to method of production
(0.032) Handwritten documents (autograph, holograph copies). Manuscripts. Pictorial documents (drawings, paintings)
(0.034) Machine-readable documents
(0.04) Documents according to stage of production
(0.05) Documents for particular kinds of user
(0.06) Documents according to level of presentation and availability
(0.07) Supplementary matter issued with a document
(0.08) Separately issued supplements or parts of documents
(01) Bibliographies
(02) Books in general
(03) Reference works

(04) Non-serial separates. Separata
(041) Pamphlets. Brochures
(042) Addresses. Lectures. Speeches
(043) Theses. Dissertations
(044) Personal documents. Correspondence. Letters. Circulars
(045) Articles in serials, collections etc. Contributions
(046) Newspaper articles
(047) Reports. Notices. Bulletins
(048) Bibliographic descriptions. Abstracts. Summaries. Surveys
(049) Other non-serial separates
(05) Serial publications. Periodicals
(06) Documents relating to societies, associations, organizations
(07) Documents for instruction, teaching, study, training
(08) Collected and polygraphic works. Forms. Lists. Illustrations. Business publications
(09) Presentation in historical form. Legal and historical sources
(091) Presentation in chronological, historical form. Historical presentation in the strict sense
(092) Biographical presentation
(093) Historical sources
(094) Legal sources. Legal documents

(1/9) Common auxiliaries of place. Table 1e

(1) Place and space in general. Localization. Orientation
(1-0/-9) Special auxiliary subdivision for boundaries and spatial forms of various kinds
(1-0) Zones
(1-1) Orientation. Points of the compass. Relative position
(1-11) East. Eastern
(1-13) South. Southern
(1-14) South-west. South-western
(1-15) West. Western
(1-17) North. Northern
(1-19) Relative location, direction and orientation
(1-2) Lowest administrative units. Localities
(1-5) Dependent or semi-dependent territories
(1-6) States or groupings of states from various points of view
(1-7) Places and areas according to privacy, publicness and other special features
(1-8) Location. Source. Transit. Destination
(1-9) Regionalization according to specialized points of view
(100) Universal as to place. International. All countries in general
(2) Physiographic designation
(20) Ecosphere
(21) Surface of the Earth in general. Land areas in particular. Natural zones and regions

(23) Above sea level. Surface relief. Above ground generally. Mountains
(24) Below sea level. Underground. Subterranean
(25) Natural flat ground (at, above or below sea level). The ground in its natural condition, cultivated or inhabited
(26) Oceans, seas and interconnections
(28) Inland waters
(29) The world according to physiographic features
(3) Places of the ancient and mediaeval world
(31) Ancient China and Japan
(32) Ancient Egypt
(33) Ancient Roman Province of Judaea. The Holy Land. Region of the Israelites
(34) Ancient India
(35) Medo-Persia
(36) Regions of the so-called barbarians
(37) Italia. Ancient Rome and Italy
(38) Ancient Greece
(399) Other regions. Ancient geographical divisions other than those of classical antiquity
(4/9) Countries and places of the modern world
(4) Europe
(5) Asia
(6) Africa
(7) North and Central America
(8) South America
(9) States and regions of the South Pacific and Australia. Arctic. Antarctic

(=...) Common auxiliaries of human ancestry, ethnic grouping and nationality. Table 1f

They are derived mainly from the common auxiliaries of language =... (Table 1c) and so may also usefully distinguish linguistic-cultural groups, e.g. =111 English is used to represent (=111) English speaking peoples

(=01) Human ancestry groups
(=011) European Continental Ancestry Group
(=012) Asian Continental Ancestry Group
(=013) African Continental Ancestry Group
(=014) Oceanic Ancestry Group
(=017) American Native Continental Ancestry Group
(=1/=8) Linguistic-cultural groups, ethnic groups, peoples [derived from Table 1c]
(=1:1/9) Peoples associated with particular places
e.g. (=111:71) Anglophone population of Canada

"..." Common auxiliaries of time. Table 1g

"0/2" Dates and ranges of time (CE or AD) in conventional Christian (Gregorian) reckoning
"0" First millennium CE
"1" Second millennium CE
"2" Third millennium CE
"3/7" Time divisions other than dates in Christian (Gregorian) reckoning
"3" Conventional time divisions and subdivisions: numbered, named, etc.
"4" Duration. Time-span. Period. Term. Ages and age-groups
"5" Periodicity. Frequency. Recurrence at specified intervals.
"6" Geological, archaeological and cultural time divisions
"61/62" Geological time division
"63" Archaeological, prehistoric, protohistoric periods and ages
"67/69" Time reckonings: universal, secular, non-Christian religious
"67" Universal time reckoning. Before Present
"68" Secular time reckonings other than universal and the Christian (Gregorian) calendar
"69" Dates and time units in non-Christian (non-Gregorian) religious time reckonings
"7" Phenomena in time. Phenomenology of time

-0 Common auxiliaries of general characteristics. Table 1k

-02 Common auxiliaries of properties
-021 Properties of existence
-022 Properties of magnitude, degree, quantity, number, temporal values, dimension, size
-023 Properties of shape
-024 Properties of structure. Properties of position
-025 Properties of arrangement
-026 Properties of action and movement
-027 Operational properties
-028 Properties of style and presentation
-029 Properties derived from other main classes
-03 Common auxiliaries of materials
-032 Naturally occurring mineral materials
-033 Manufactured mineral-based materials
-034 Metals
-035 Materials of mainly organic origin
-036 Macromolecular materials. Rubbers and plastics
-037 Textiles. Fibres. Yarns. Fabrics. Cloth
-039 Other materials
-04 Common auxiliaries of relations, processes and operations
-042 Phase relations
-043 General processes
-043.8/.9 Processes of existence
-045 Processes related to position, arrangement, movement, physical properties, states of matter

-047/-049 General operations and activities
-05 Common auxiliaries of persons and personal characteristics
-051 Persons as agents, doers, practitioners (studying, making, serving etc.)
-052 Persons as targets, clients, users (studied, served etc.)
-053 Persons according to age or age-groups
-054 Persons according to ethnic characteristics, nationality, citizenship etc.
-055 Persons according to gender and kinship
-056 Persons according to constitution, health, disposition, hereditary or other traits
-057 Persons according to occupation, work, livelihood, education
-058 Persons according to social class, civil status

2.8 SUMMARY

The UDC is one of the important developments in the history of classification schemes. It is based on DDC. However, it has included several notational techniques and became fist analytic-synthetic classification. This scheme is popularised as international classification by its features such as use of common axillaries, and several signs and symbols to indicate different aspects of subjects. The scheme is also being published in abridged editions, besides full editions in various languages of the world.

2.9 TECHNICAL TERMS

MGA: Meteorological & Geo Astrophysical Abstracts

2.10 SUGGESTED READINGS

1. Dhyani, P. Universal Decimal Classification International Medium Edition. “Library Review”Vol21, 1989, P.165-172
2. Universal Decimal Classification. London,BSI, 1985
3. Kumar, Krishan. Theiry of Library Classification. 4th ed. New Delhi : Vikas , 1988

LESSON - 3

BROAD SYSTEM OF ORDERING (BSO)

AIMS AND OBJECTIVES

The objective of this lesson is to explain machine oriented classification system called Broad System of Ordering (BSO).

After studying this lesson you can know

- the origin of BSO
- the usage of BSO
- the structure of BSO and
- the applications of BSO

Structure

3.1 INTRODUCTION

BSO (Broad System of Ordering), also termed SRC (Subject-field Reference Code), is a classification system developed within the UNISIST-program for the purpose of interconnection of information systems. It is a disciplinary organized system founded in 1972 as a UNESCO project in the UNISIST-program "World Science Information System" in cooperation with FID. The idea behind BSO is related to the idea of networks and probably represents the latest attempt to create a new universal classification. It is developed by the Englishman Eric Coates in cooperation with others, including the Bliss Classification Association. BSO is a modern machine-held classification system embracing all fields of knowledge.

BSO was meant to be an international switching language, an overall information retrieval language to transfer blocks of information in coarse subject groups between information systems applying different indexing languages.

Coates (1979) states: "theoretically the switching operation requires nothing more than neutral code system in which concepts are represented". Dahlberg (1978) regards it as a positive step towards standardization: "A standard classification assists library rationalization and national and international cooperation on statistics, research and cataloguing".

'Subject indication' is the phrase used to refer to those facilities of an information system which enable it to be interrogated by queries which have a subject as their point of departure. The user supplies to the system the name of a subject with the aim of extracting information on that subject from the system's store. Frequently the system contains in its store, not the information ultimately required, but records of the names and addresses for documents in which the information is likely to be found. In such a case subject indication has as its aim the identification of documents carrying the required information. The tools or languages of subject indication include indexing languages, classification systems, controlled term or keyword lists and thesauri.

The Broad System of Ordering is such a subject indication language, or, more specifically, a classification system, developed for a proposed world-wide information network covering the whole field of knowledge. At first sight there appears to be a little reason for supposing that a subject indication language for a network should be fundamentally different from a subject information language for information system generally, and it is arguable that the schedules of *BSO* bear this out. However, there are areas of such uncertainty surrounding subject indication languages, that it would have been rash indeed not to have put the matter to the test by information requirements in view. It will be seen that the result of the exercise bears a family resemblance to some of the document classifications which have preceded it, despite the fact that during the exercise no reference or recourse was made to literary or documentary warrant in the direct sense. Whether the differences between *BSO* and the document classifications are considered significant or trivial in themselves, they could possibly prove to be essential to the network application.

3.2 ORIGIN OF BSO

BSO originated in the context of the idea which emerged in the 1960s that consideration should be given to the possibility of a global network of scientific information centres, taking into account particularly the needs of developing countries. The

network idea was itself triggered by a technological development, not at that time generally available, but certainly upon the near horizon. This was the possibility of cheap and fast data transmission links. It is notable that thinking about the information network, involving the first steps towards system definition began about a decade before the hardware became generally available. This was an honourable exception to the more usual situation in the mechanisation of information services in which the computer hardware was available well advance of system planning.

The subject indication sub-system of the network was seen as an important part of the whole system. It was vital in such a network that information on the subjects of documentary resources held by any one participating centre should be accessible to all the centres in the network. There were two closely interlocked, but still separable, problems here. The first was that, despite - and perhaps because of - the growth of mechanisation of information services, which in the late 1960s was just getting under way on a substantial scale, a greater amount of subject indication activity depended upon human intuitive skill and know-how than in the pre-mechanisation period ending about 1955. There was, for instance, an unprecedented proliferation of controlled keyword lists and thesauri, for use with mechanised systems, but little sign of common logical rationale in their construction which might otherwise itself be amenable to mechanisation. These human skills produced indexing tools of great diversity for particular subject areas. Did these tools, the detailed construction principles of which were usually not fully communicable, offer a suitable model for the subject indication language of the proposed network?

The second problem arose for the realisation that for often good and sufficient reasons, centres representing the various subject fields would continue to use a variety of subject indication languages, corresponding to a variety of needs. Accordingly, communication through the use of a possible standard indexing language, which all participating centres would use for subject description of documents, was ruled out. On the other hand a solution seemed to lie in a procedure whereby subject information coded in one local indexing language could be converted by clerical means into the codes of another language conveying the same subject information.

3.2.1 SWITCHING INDEXING LANGUAGES

It so happened that this solution involving interconnection of individual local indexing languages, by a mediating or switching language had been under study by the Groupe d'Etude sur l'Information scientifique based at Marseilles, since 1963. Unlike the proposed global scientific information network, the system envisaged by GEIS was composed of centres dealing with the same subject discipline, and the particular discipline used as a study sample was the Science of Scientific Information itself. This difference is of some importance when considering the transfer *en bloc*, of the conclusions of GEIS in their 'Intermediate Lexicon' project to the context of the global scientific information network. A key feature in the GEIS scheme was the 'equivalence' or 'conversion' table in which the code for a given concept as rendered in one indexing language was coupled with the code of the same concept in another indexing language. It was assumed that such coupling was practicable to the extent that both indexing languages were, in fact as well as in pretension, lists of terms each of which corresponded with a definite and unambiguous concept. In fact the term 'indexation' was reserved for the process of concept analysing a document and assigning a code accordingly. (The code could be a notation symbol of a classification, or a descriptor or authorised term drawn from a thesaurus or subject heading list).

For the simplest case of a network in which the participating centres taken in aggregate used only two local indexing languages, all that was required was a pair of 'equivalence tables', one leading from language A to B, and the other from language B to A. If there were more than three indexing languages represented in the network, then it became more economical, in terms of the number of pairs of 'equivalence tables' required, to employ what has been variously termed as switching language, a mediating language, or a communication indexing language. A message (i.e. a subject request, or the answer to a subject request) would thus proceed from centre A (using local indexing language A) via an 'equivalence table' to the switching language, and then outward to a further 'equivalence table' reaching its destination, centre B, coded in the form in which the same subject is rendered in local indexing language B. This system, which is exactly analogous to a telephone network, would require one pair of 'equivalence tables' (one subscriber's line in the telephone analogy) between each centre's local indexing language and the switching language, ; whereas, if there were no switching language employed, each centre would need to construct, and of course maintain, as many pairs of 'equivalence tables' as there were languages in use in the network, minus one. It has been mentioned that such a system was expected to work, subject to the condition that the various local languages were in fact concept controlled. By the same token, the switching language itself would need to be one in which each representation (notation symbol, or term) corresponded to one, and one only, concept; and in which for every concept there was one , and one only preferred representation. The form and arrangement of the switching language could be considered a function of the kind of use for which it was intended. If it were required only for simple matching, its arrangement would be immaterial to those engaged in day-to-day operations of sending and receiving messages. If additionally it were intended to employ such a switching language for hierarchical search, it would be necessary to incorporate the necessary hierarchical linkages into the language. However from the point of view of constructing and updating a switching language under controlled vocabulary conditions, some form of schematic arrangement, on the lines of a classification, is probably mandatory. This schematic arrangement might not, however, be the form in which the language was most conveniently held in a computer store.

It is to be supposed that the idea emanating from GEIS (which was later given quantitative elaboration in a research study carried out at the Polytechnic of North London School of Librarianship) because it was the only one of its kind available, must have affected the thinking of members of the first bodies charged with the task of considering the global science information programme, when they turned to the question of subject indication.

3.3 GUIDING PRINCIPLE OF BSO

3.3.1 The limits of broadness

The first task of the FID/SRC Working Group was that of sharpening the somewhat indefinite terms of the remit entrusted to it. The central question here was to try to decide in the most concrete possible manner what was to be understood by broadness as a feature of the proposed Broad System of Ordering. What should the determining principle be, which would cause some terms to be included in the scheme as sufficiently broad, and others to be rejected on the ground that they were too specialised?

3.3.1.1 Arithmetical Approach

Several possible approaches to an answer to this question could be foreseen. The answer could be purely arithmetical - a stated total number of terms in the system could be settled in advance. The answer could be based on a particular property of terms in relation

to the classification structure yet to be devised - namely the hierarchical level of the term within the structure. Or, it could be based upon some inherent semantic property which a term might or might not possess. Or, yet again, it could be based upon some formal linguistic property of terms. Finally, it might be possibly based upon some sufficiently objective social property or phenomenon associated with the term or with the concept denoted by it. An obvious thought here was that a social property useful as marking cut off of detail might well be one closely related with the purpose which it was hoped BSO would serve.

The first approach to the problem of defining broadness, or cut-off point, for BSO - the laying down in advance of the total number of terms to be used - had some special attractions in relation to cost predictability, especially in the context of mechanised exchange of information within a network. Clearly the cost of the computer processing of such exchange would fairly closely depend upon the size of the interconnection language to be traversed in the passing of each message. This dependence is probably less significant now (1978) than at the outset of BSO development in 1973. With the expected future use of microprocessor elements as customary computing hardware, it is likely to be even less significant in future. At the beginning of the development of BSO, it was provisionally assumed that the full scheme might contain 2000 terms, and in the first draft submitted for comment in 1975 there were in fact 2100 terms. This draft was faulted both on account of its omissions and on account of alleged over-development of detail. Such criticism on mutually opposed grounds might have been crudely interpreted as a justification for the middle position taken by the BSO draft. However, the tenor of the comments themselves pointed to a great weakness of any solution to the cutoff problem based upon a prescribed maximum number of terms. As the approach to the maximum is reached, the question of what is, or is not, to be included in the system, comes to depend upon refined judgments of the relative importance of candidate subjects. Reliability in such judgments or judgments reflecting a real consensus are hardly to be expected from practitioners in the borderline specialties themselves - these specialties are, after all, often in competition among themselves for social recognition and funding. For this reason any purely arithmetical characterisation of cutoff in terms of the total number of subject-terms which the system is to contain is likely to be unsatisfactory and tendentious.

A brief side-glance at the arithmetical size of the scheme as a result of a cutoff criterion to be described later may be in order here. The 3rd revised draft of BSO (1978) contains about 4000 terms. The 18th edition of DDC, an established general classification for books has about 80,000 terms, so in approximate terms an average 'broad block' of information which can be designated by BSO is 20 times 'broader' than typical information at book level, and 3 times 'broader' than the information units which can be designated by the abridged UDC.

3.3.1.2 Hierarchical Approach

Many comments received on the earlier drafts of the scheme assumed without question that cutoff could appropriately be defined by reference to some-hierarchical level in the scheme. The arguments against such a basis for setting the limits of detail of the scheme are formidable. It can be contended that the policy on limit of detail, far from being derivative from the exigencies of the structure of the ordering system itself, should be independent of that structure. The structure is for the purpose of ordering, not for delimitation of acceptable detail. Furthermore hierarchical level of a given term is one of the most unstable features of all classifications in face of necessary changes required by the

arrival of new knowledge. Much new knowledge arises by the fusion, following the discovery of common properties, of two or more hitherto separate subjects on the same hierarchical level. Whenever this occurs the separate subjects and all other subjects subsumed by them change their hierarchical level. Another consideration is that a statement of a hierarchical level is often made for explanatory or presentational purposes (for example in BSO 212 ENERGY INTERACTIONS & FORMS (ANY STATE OF MATTER). Obviously alternative presentational strategies are possible, and they will to some extent depend on available type variations for display. Also a chosen strategy may at some time have to be modified because of the appearance of a new subject remote from the hierarchical statement in question. Hierarchical levels are thus determined both by logical imperatives and presentational nuances. Both factors are subject to necessary change, and their states at a particular moment should not be the determinants of system detail cutoff Finally the practical importance of a subject is by no means necessarily correlated with its hierarchical level. For instance 923,70 BASQUE LANGUAGE, being a unique member of a set is on the same hierarchical level as 921 INDO-EUROPEAN LANGUAGES.

The lack of agreement between natural languages as to the incidence of 'logies' and 'graphies' probably reflects the fact that mental organisation - the central characteristic of the kind of knowledge which constitutes a discipline - is not an all-or-nothing property. While one can perceive intuitively that, for instance, Chemistry is a more highly organised system of thought than Reprography; this is not to say that Reprography is not a discipline. Indeed, it would be quite hard to identify any subject matter which has generated literature, which can confidently be said to possess zero mental organization. On the practical plane of handling subjects found in documents, no hard and fast line can be drawn between disciplines and non-disciplines from the standpoint of mental organisation of the material.

3.4 APPLICATION OF BSO

The primary purpose for which BSO has been compiled is to serve as an exchange or switching language for use in an information network covering all subjects and in principle extending to users anywhere in the world.

3.4.1 Concept representation as the basis of switching

Behind the surface idea of subject indication switching between different indexing languages lies the assumption that despite the fact that individual centres participating in a network may differ from one another in the formalisms of their local indexing languages, there is between them an underlying agreement as to the nature and relations of the concepts represented in the local indexing languages. In other words, diversity belongs to the plane of language and terminology, but agreement to the plane of thought and idea. Switching is accordingly feasible on the plane of thought and idea on which agreements exists.

Different sets of indexing terms, descriptors, or notation symbols, used in different indexing languages to represent the same idea can be made to switch their idea-content between centres, provided that

a) each local indexing language consists of terms and symbols, each of which is the sole representation, in the language, of a particular idea, and also represents that idea alone

b) some neutral representation of the idea, agreed by all concerned, becomes the medium for clerical linkage for switching purposes. The neutral or switching language of

concept representation must, like the local languages involved in the switching process be a controlled language.

Does this mean that a centre using free-text indexing cannot participate in switching? The answer is that in formal terms such a centre could participate, but practically it is unlikely to do so, because, in preparing the necessary concordance tables between its own input and the switching language, it would need to embark upon a vocabulary control exercise no less onerous than the control of the local indexing language itself: this is, however, the burden from which free-text indexing seeks to escape.

3.4.2 Form of switching language

The next question which arises is: what form of controlled indexing language is appropriate for the switching duty? Should it be arbitrary identifying code, a thesaurus, or a classification? An arbitrary code which carries no implicit or explicit information upon relations between vocabulary control itself - namely, the selection of codes to represent concepts uniquely - depends upon prior process of clustering concepts in order to establish near relationships and actual identity. An arbitrary code is no aid to such clustering. ON the other hand thesauri and classifications do display semantic relations - relations between ideas on the plane of meaning.

The choice between universal classification and universal thesaurus for the switching language role follows from the manner in which each displays relationships. A classification attempts to display relationships as a totality by means of tabulation. A thesaurus depicts relationships in a fragmentary manner, in the form of binary linkages, each of which is probably separated from semantically 'next neighbour' binary linkages by the accident of the alphabet. There is very little question as to which manner of relational display is the more useful for the purpose of controlling the vocabulary in face of an incoming flow of candidate new terms. Indeed, it is becoming increasingly common for thesauri themselves to supplement the fragmentary manner of showing semantic relationships, by adding to the alphabetical sequence of keywords, ancillary sections of grouped, categorised, or fully classified terms. Indeed it is becoming increasingly common for thesauri themselves to supplement the fragmentary manner of showing semantic relationships, by adding to the alphabetical sequence of keywords, ancillary section of grouped, categorised, of fully classified terms. From a small thesaurus the clustering process essential to vocabulary control in admitting new terms may be undertaken informally as a purely mental activity. If the thesaurus is large and of wide subject scope, then reliable and economic control of the vocabulary requires that the clustering should be externally formalised as a classification structure. It has been argued earlier that the practicability of a universal switching language depends critically upon its ability to be controlled, revised, and updated with minimum effort. A classification, more than any other form of indexing language, is amenable to easy, predictable, yet at the same time fully controlled updating. This is the essential ground upon which it is the preferred form of indexing language for the universal switching application. That existing universal classifications have failed, or are visibly failing, precisely in this respect does not vitiate the argument. The theoretical developments in classification of the last half-century have been preferentially applied to special subject classifications. BSO is in one sense an attempt to bring many of these developments into the sphere of general classification. It seems likely that these developments, all in the direction of bringing pervasive structural patterns into general classification, may hold the key to resolving the updating/keeping-up-

with-knowledge problem which besets the established systems of universal classification and their users.

3.4.3 New knowledge, new technology and universal classification

On the broadest perspective, the UNISIST requirement of a classification, covering all fields, for exchange or switching purposes, may be seen as a particular concrete manifestation of a more general new need for a universal classification which has emerged only in the present decade. This need has arisen from the conjunction of three separate factors. The first of these concerns the process by which growing points of new knowledge often appear astride of discipline boundaries, and in aggregate have the effect of diminishing the practical significance of these boundaries. This process has been well recognised for many years but its impact has only recently been fully felt. The fringe or marginal subjects of specialised information services are spreading ever more widely over the total field of knowledge. It is not only that some of the socially more significant of the new technologies are of mixed scientific parentage. There is at present a considerable emphasis on what may be termed holistic approaches to all departments of human affairs. The ground 'between' technology, economics and apparently more distantly related social sciences is at present receiving unprecedented attention, as may be seen from the appearance of such interdisciplinary information services as SPLINES. Equally the boundaries between technology and social sciences have become blurred by the integrated concept 'Environment' which ultimately stems from the realm of biology and psychology. The rise of this holistic standpoint has on the one side strained the capacity of the established general classifications for accommodation to near breaking point, and on the other stimulated a new need for a universal classification.

The second factor contributing to a new need for universal classifications is directly technical in character. The limitations of clerical manual methods of manipulation and transfer of information records tended to confine such activities to single disciplines, within which quantities of material to be processed were sometimes manageable. Electronic data processing has vastly relaxed these limitations, and accordingly the significance of the discipline boundaries themselves has relaxed.

The third factor contributing to a new need for universal classifications is directly technical in character. The limitations of clerical manual methods of manipulation and transfer of information records tended to confine such activities to single disciplines, within which quantities of material to be processed were sometimes manageable. Electronic data processing has vastly relaxed these, limitations, and accordingly the significance of the discipline boundaries themselves has relaxed.

The fourth factor leading to a renewed need for a universal classification has been the internationalisation of information processing activities. Access to information is far less than hitherto the prerogative of advanced countries alone. In the developing countries there is at present great activity in the setting up of information centres covering all fields of knowledge, and collecting or arranging access to information from all sources. This flow of information on a global scale has re-animated the whole issue of a universal classification, particularly in its role for indicating the nature of the subject-content of information requests and documents.

3.5 BSO FOR SWITCHING AND MEDIATING

All of the three above factors are clearly related to the universal switching language application of BSO. An information network should be capable of connecting centres individually oriented to different focal disciplines. Its practicability on a large scale depends substantially upon exploiting data processing and transmission technology, and the associated switching language has to be capable of surmounting linguistic and cultural barriers.

There are other possible applications, essentially of the same operational type as the switching language, which may be envisaged for BSO. In all cases they are products of the first and third of the three general factors mentioned above which seem to require a new universal classification, but the second factor concerning the liberation of earlier restraints owing to data processing technology is generally less significant than in the switching language application and may be absent altogether. In most cases, though not exclusively, these additional applications involve users who potentially may be in any part of the world.

Networking is not the only context within which neutral mediating languages come into play. Considerable financial resources are at the moment being applied to the translation, harmonisation, and inter conversion of thesauri. For the minority of thesauri which themselves are no more detailed than BSO it is possible to conceive of BSO as a clerical switching language. For larger thesauri involved in inter conversion projects designed to give users of one indexing language access to documents indexed in a different language, the use of BSO is a mediating, or common reference, language would achieve the necessary preliminary clustering of related terms from both thesauri, and would provide a framework for the higher organisation of the formed clusters, which might not be carried through to the finished product, but in any case would be useful as a provisional concept-holding device while the conversion work was in progress. The advantage of this approach would be both to eliminate decision process in the preliminary clustering, and to enable the broadest view of the overall subject-structure of the thesauri to be available from a very early stage in the project. Thus costly looping back whereby an early decision has to be modified to conform with the implications of a decision taken later - very characteristic of piecemeal operations on a structure of which the integrity is for the time being invisible - would be eliminated. The preliminary clusters thus formed would of course require to be broken down further by human intelligence - this being the inevitable limitation of a 'coarse' ordering system.

Other applications

Another example of a possible application of BSO is as an aid in the routing operations of referral centres and clearinghouses in dealing with inquiries. Compared with the switching application, this use of BSO would exploit communications technology equally, but its involvement with data processing would be less sophisticated.

Ultimately serving the same purposes as the referral centre, but serving individual demand by the mass medium of an older form of communication - the printed world - is the comprehensive directory of specialist organisations and specialist information sources. From the point of view of subject indication, present standards in publications of this kind could be improved to the substantial benefit of users. Such improvement could be realised either by arranging the material by BSO codes or by providing and index from BSO codes to page or item serial numbers.

It is also possible to envisage the use of BSO in purely disseminative modes of communication. As a subject tag supplied on copies of distributed reports and separates of all kinds, BSO codes would serve recipients of this material both as a 'coarse' interest filter, and secondly as a temporary filing system both for purposes of retrieval and subsequent control of disposal of little used material.

In these latter applications and in some others such as the possible use of BSO codes as subject indicators in machine readable records. BSO would to some extent be competitive with existing established general classifications. The seriousness of this competition would perhaps depend upon

a) The inherent advantages and disadvantages, input cost-wise and user-wise or relatively 'coarse' subject specification versus the more detailed specification aiming at book level or documentation level in the established schemes

b) The relative merits of BSO and the established schemes in providing unequivocal placing for subjects, and thus in ease of decision effort in indexing

c) Achievement by BSO of a new style of updating arrangement which would permit prompt assimilation of new knowledge into the scheme at a cost to the user which would be found acceptable

One final question which arises here is whether BSO might conceivably in future infiltrate or invade the territory proper of the established document classifications. In other words, will it ever be used for shelving books or filing documents in libraries? The answer to this question depends upon established systems rather than upon BSO. All that can be said is that if the established systems are found wanting on either of the two issues labelled b) and c) in the foregoing paragraph, then this same question will doubtless be raised repeatedly. It is not entirely unusual for tools of this kind to be used for purposes other than those for which they were originally intended. Furthermore there is nothing in the design of BSO which would inhibit elaboration to a greater depth of detail.

BSO - DESCRIPTION OF THE SCHEME

3.5.1 One System design and user effort

It has been suggested in the preceding chapter that a switching indexing language needs to be economical in usage. The benefits of networking are not obtainable entirely without cost. The indexing of material by a switching language at a centre would, after all, be an addition to indexing effort normally put forth for local purposes. It is therefore essential that the additional cost of communication with other centres in the network should not contain any unnecessary element. It is against this background that the question of the cost of BSO to the user, both in day-to-day operation and in making changes consequent upon changes in the content and structure of knowledge, has been a matter of primary concern at every step in designing the scheme.

A classification user's unnecessary costs arise mainly in two ways. First, day-to-day application of the scheme may demand more decision effort than is necessary. Second, the local implementation of update amendments to the scheme may involve unnecessary effort.

Unnecessary decision effort is the result either of gross mismatch between the subjects found in the material to which the classification is to be applied and the concepts represented in the classification itself, or to lack of structural homogeneity in the scheme

itself. It should be noted that mismatch is the result not only of initial shortcomings of the scheme but also of delays in updating. Lack of structural homogeneity may be paraphrased as unnecessary complexity in the scheme due to absence of overall pattern. An example of an inhomogenous general classification would be one which was prepared simply by bringing together the special classifications corresponding to each included subject area, and listing them sequentially (possibly in some logical or otherwise helpful order). Any discipline, almost by definition, represents a particular viewpoint. A series of classifications, each optimal for the needs of a particular viewpoint, form, when added together, a general classification of great complexity, and consequently demand excessive decision effort in being applied.

Unnecessary effort in implementing, updating, both on the part of the updater and of the user, is demanded when the insertion of a new subject requires not only an addition to the schedule but also a re-notation of adjacent terms representing old knowledge. This may arise either because the area involved was in the first place inadequately structured or because of a constraint offered by the notation.

These considerations are reflected in the general features of BS0, which include a marked incidence of structured pattern, both within and transcending discipline boundaries. The system is also highly prescriptive. There are no alternative placings offered. Completely definitive and embracing procedures are laid down by which indexers deal with the necessary factor of cross-classification in the schedule, which is therefore expected to be non-ambiguous in use and predictable in updating.

3.5.2 Neutrality and value judgments

After the question of the economics of the system comes the matter of its neutrality. All special classifications reflect the special viewpoint partly inherent in the discipline concerned and partly conventional among specialists within the discipline. Likewise, all general classifications are vulnerable to the charge that they reflect some particular world outlook or philosophy. This is obviously a question with potentially serious implications for a scheme intended for global use. Like the material which will be subject to switching in the foreseeable future, BSO reflects in many ways the standpoint of European tradition and culture. Within this limitation, the compilers have tried to stand outside sectional philosophies and to avoid decisions which have sectional philosophical implications. It is perhaps necessary to insist that neither the hierarchical nor ordinal position of any term carries any implication as to the importance of the associated concept.

3.6 OUTLINE OF BSO

The outline of the system is as follows:

FIRST OUTLINE OF BSO

088	Phenomena & entities from a multi or non-disciplinary point of view	460	EDUCATION
		470	HUMAN NEEDS
	SUBJECT FIELDS	475	Household science
		477	Work & leisure
		480	Sports & games

100	KNOWLEDGE GENERALLY		
112	Philosophy	500	HUMANITIES, CULTURAL & SOCIAL SCIENCES
116	Science of science		
118	Logic	510	History
120	Mathematics	526	Area studies
128	Computer science	530	Social sciences
140	Information sciences	533	Cultural anthropology
150	Communication sciences	535	Sociology
160	Systemology	537	Demography
165	Management	540	Political science & politics
182	Research	550	Public administration
188	Metrology	560	Law
200	SCIENCE AND TECHNOLOGY	570	Social welfare
203	Natural sciences	580	Economics
205	Physical sciences	588	Management of enterprises
210	Physics		
230	Chemistry	600	TECHNOLOGY
250	Space& earth sciences	910	LANGUAGE, LINGUISTICS & LITERATURE
300	Life sciences		
300/439	Application of life science		
360	Agriculture	940	ARTS
368	Veterinary science	943	Plastic arts
368	Forestry	945	Graphic fine arts
380	Wild life exploitation	949	Decorative arts & handicrafts
390	Environment & natural resources	950	Music & performance arts
410	Biomedical sciences	970	RELIGION & ATHEISM
445	Behavioural sciences		
450	Psychology		

3.7 SYNTACTIC ASPECTS AND COMBINATORY FACILITIES

In the BSO schedules some subjects may be used as 'tools' in other subjects, and that some contribute 'aspects' to others. 'Tools' and 'aspects' represent certain kinds of syntactic relations. These are relations, at the concept level, between terms which stand together to denote compound or composite subjects. 'Cross-classification' is frequently used to refer to the dilemmas experienced by classifiers attempting to assign places for composite subjects in classification schemes which are inadequately prescriptive on the handling of syntactic relations.

Among organised information sources there are some which are devoted to subjects which are composite in nature. Accordingly BSO has comprehensive facilities for combining notational elements to represent composite subjects. It is, in fact, a fully

synthesising or faceted system, though it has not been thought necessary or even desirable to label facets as such.

Combinatory facilities in classification systems inevitably raise the issue of order in which the elements are combined, also called citation order or facet order. In some working situations this issue may be bypassed or left to intuitive judgment, but for a neutral mediating indexing language covering all subject fields a completely fixed or prescriptive citation order appears to be necessary to ensure reasonably' noise-free transmission of information.

In BSO the order in which notational elements are combined to form codes for composite subjects is in the majority of cases the reverse of the order in which the elements are set down in the classification schedule. Without the qualification in the majority of cases citation order problems would be reduced to purely clerical procedures, and if we can specify those situations to which the reverse-schedule-sequence rule applies without exception, we still have a highly time and effort saving feature of BSO.

It is first of all useful to categorise combinations into internal combinations which comprise notational elements drawn from the same subject field (e.g. 575,32,0,73,50 Child welfare in disaster relief, constructed from the elements 575,32 Child welfare and 573,50 Disaster relief and aid) on the one hand, and external combinations constructed from notational elements taken from different subject fields (e.g. 550-163 Operations research in public administration, using elements 550 Public administration and 163 Operations research) on the other.

In order to make this categorisation completely explicit it is necessary to state unambiguously what is meant in this context by a subject field. Subject fields for defining internal vs. external combinations are enumerated as 'Combination areas' on page xi of the published BSO. A combination with both elements drawn from one of these 'Combination areas' is an internal combination.

Internal combinations without exception obey the reverse-schedule sequence rule for combination order. (In the above example the leading element in the combination, 575,32 is later in the schedule than the second element 573,50).

The structural background to this combination rule is that each subject field is elaborated according to a facet pattern, which, with very slight variations, is repeated over many fields. The following is the commonest facet pattern, given in schedule sequence which would be reversed for combination order:

1) Tools or equipment for carrying out operations
2) Operations (i.e. purposive activities by people)
3) Processes, interactions
4) Parts, subsystems of objects of action or study, or of products
5) Objects of action or study, or products, or total systems

(In the example above the first element in the combination order, namely the concept Child belongs to facet (5), the second element, the process which requires a welfare operation to be undertaken, namely the concept Disaster, belongs to facet (3). Facet (4) is inapplicable to this subject field. Facet (2) is applicable but has no role in this

combination because the operation, Welfare already defines the whole 'combination area'. Facet (1) would be applicable if a particular kind of welfare agency were to be specified). Such regularity of underlying pattern covering the whole scheme is conducive to economy both in the day-to-day use of the scheme by indexers or searchers and in predictable updating.

The reverse-schedule-sequence rule cannot be used in the same clerical or mechanical manner in deciding combination order for external combinations, though more often than otherwise it would give correct and consistent results. The reason why it cannot be employed reliably for external combinations can be shown from a single example. Let us assume that reverse-schedule-sequence is being used as the basis for combination order in the case of Educational psychology. The rule will then give 460-450 (460 is Education, the hyphen or dash is the connecting symbol for external combinations, 450 is Psychology). Educational psychology, may be approximately factored as the Psychological aspects of the Education process. How then do we code Psychological education, the teaching and training in the subject Psychology? If the reverse-schedule-sequence rule were used we should arrive again at 460-450 as for Educational psychology. For any indexing system covering the whole of knowledge this would produce unacceptable noise at output. The example leads to two further considerations. The first is that external combinations should not (in the manner of the UDC colon connecting symbol) be used to indicate any relationship. This would lead not only to output noise, but also to anomalies in file sequence of classified material. The second consideration is that both in Educational psychology and Psychological education, one of the subjects (or rather the phenomena of one of the subjects) is the 'recipient' or 'target' to which the other subject contributes a set of aspects or properties. Thus psychological viewpoints are contributed to the education process in Educational Psychology, and an education process is contributed or applied to the realm of psychology in Psychological Education. An interesting difference to be noted in passing is that the 'recipient' in Educational psychology is the primary phenomena of education, i.e. the education process, while the 'recipient' in the case Psychological education is not the primary phenomena of psychology - there is no reference here to the educating of psychological processes - but the second-order phenomena of people involved in psychology as a field of interest or profession.

The upshot of these considerations is that combination order for external combinations in BSO needs to be determined by reference to the relation between the elements which require connection. The following rule which emphasises the directionality of the 'recipient' element in the relation and the 'aspect contributing' element is believed to be unambiguously applicable to situations which can be represented by external combinations, and is recommended:

Cite first	the notation for the element denoting application area, mission, purpose, end-product or whole system: more generally the subject which 'receives' an action or effect, or is seen according to a particular viewpoint, or has a property attributed to it
Cite second	the notation for the element denoting aspect, approach, action applied, agent, or part of a stated whole: more generally the subject which 'contributes' an aspect, approach or action.

Use of the above relational formula where the 'aspect contribution' element belongs to the area 210 to 450 will normally produce combination orders which reverse the schedule order, as in the case of internal combinations throughout the schedule. This is because in this area the entities and phenomena studied by a particular science include aspects and properties which essentially belong to other sciences located earlier in the schedule sequence. For instance biological entities may have physical or chemical properties: medical, psychological and social phenomena may have biological aspects. In these cases the roles of 'aspect contributor' and 'recipient' elements cannot be reversed, as long as the 'recipient' element is the primary phenomena of the subject field concerned. It has been shown in the case of Psychological Education the 'recipient' element is the second order phenomena associated with the subject field and an apparent reversal of roles results in a combination order which is identical with the schedule order of the elements.

Cases in which the 'aspect contributed' element belongs to the area 460 to 992, and to which the relational formula is applied, more frequently produce combination orders which break the reverse-schedule sequence rule. However, a glance at the outline suggests that the main exceptions to the reverse-schedule-sequence rule fall into a few categories. These are

1. Any social or historical aspect of any subject field 112 to 480 ('social' here is to be understood as including any aspect corresponding to subject fields 530 to 588)
2. Terminological aspect of any subject field 112 to 890.

Finally it should be noted that some departures from the reverse-schedule-sequence rule occur when the relational formula is applied to composite subjects of whom both elements are drawn from the area 112 to 188.

3.8 DISCIPLINE AND PHENOMENA CLASSES

One further feature of the BSO outline is worthy of mention. General classifications are based primarily upon subject disciplines which are methodologies and special points of view usually, but not necessarily, focussing upon a definite set of entities or phenomena. A consequence is that in conventional general classifications there is no way of classing entities or phenomena as such, merely described, or treated from many points of view. An institution dealing with, for example, Fish in all their aspects, zoological, economic, aquacultural, technical, mythological, and as quarry in a pastime, is not appropriately placed in Zoology (345,62). Furthermore there are organised information sources dealing in a multi-disciplinary manner with such topics as Food and Housing. In virtually every general classification before BSO such vital topics of everyday life have been misleadingly assigned to the discipline Sociology. Sociology is also an invariable dumping ground for multi-aspect studies, also reflected in institutional warrant, of social groups, such as Women, Racial Minorities, the Aged and the Disabled. These studies are by no means primarily sociological in viewpoint or treatment. BSO has attempted to deal with this problem by including a few phenomenon- or entity-based classes in its main outline, all containing a marked human reference, and by supplying a special location 088 at the beginning of the classification, for other phenomena or entities not included in the above-mentioned phenomena/entity-based classes. The enumerated phenomena/entity-based classes are 470 Human Needs, covering Food, Clothing and Shelter in their most extended aspects, together with Leisure, 520 Area Studies which are multidisciplinary in character, and 528 Social Groups. In connection with the residual phenomenon/entity class at 088, the problem arises as to how the multifarious phenomena and entities which might need to be assigned here should be individualised and ordered. The solution to this

problem, as used in BSO, may be illustrated by the example of Fish already given. It was pointed out that Zoology offers one point of view upon Fish and that it would therefore be wrong to assign multidisciplinary (or non-disciplinary) material on Fish within the discipline of Zoology. Yet despite this mismatch in aspect or point of view Fish have a relation with the discipline Zoology which is not of the same Kind as their relation to Economics, Aquaculture, Food technology, Sport, or Mythology. The special relation with Zoology consists of the fact that the concept Fish is uniquely defined by the zoological characteristics of fish, namely their anatomical and physiological features. The concept Fish is not similarly defined - though it may well be described - in terms of the characteristics peculiar to Economics, Aquaculture, Food technology, Sport or Mythology. When treated in multidisciplinary manner any entity such as Fish may be linked - though not subordinated - to the discipline within which it is uniquely defined, and this circumstance make available a mechanism whereby an entity may be individualised and ordered at 088. The notation for the entity within the discipline which defines it is simply added to 088,. Thus the notion of Fish is uniquely defined in zoological terms. The notation for zoological aspects of Fish is 345,62. The same notation added to 088 as 088,345,62 then signifies Fish in all their aspects, zoological and other.

A somewhat different basic view, but a similar mechanism, is applied to the problem of individualising technicall products. There is no question here of multiplicity of points of view. The point of view is assumed to be technical, embracing manufacture and the technique for using and maintaining the product. The problem is simply one of individualising the great number of kinds of products which emerge from technical processes. In BSO products defined by purpose or designed for a particular purpose are classed at the end of the Technology schedule at 890, and individualised by reference to the BSO code for the particular purpose, elsewhere in the scheme. It is necessary to emphasise 'elsewhere in the scheme' as the purpose of some products is simply to contribute to more complex technology. Such products (e.g. Switchgear) with a role internal to technology are normally enumerated in the BSO Technology schedules. The scheduled heading covers both their manufacture and use (Manufacture can be distinguished from use by employment of the suffix ,06,20 taken from 620 Production technology). One consequence of the policy for individualising by purpose those products with purposes external to technology is that 877,60 Cloth and fabric technology does not schedule manufacture of clothing as a product. The technology of the purpose-defined product Clothing is classed at 890,472. The 472 is taken from the root Human Needs code for Clothing.

3.9 COMMON FACETS

BSO has Time and Place facets, introduced by notation -01 and -02 respectively, which are functionally similar to Time and Place divisions provided in other general classification schemes. They are applicable to every subject field except those such as 510 History, 520 Area studies, 544 to 546 Political history and Politics of individual states and groupings of states, where Place and Time are specially scheduled facets. The Place facet makes use of ISO two-character alphabetical codes, and can also specify transnational political areas (e.g. EEC countries) areas defined by language, race or religion, and areas defined by the usual physical geographical factors (e.g. Tropical areas).

An Optional facet enabling the type of information source to be specified has also been included as a result of the field test. It was found that data as to type of information source is

often given prominence in descriptive material upon which indexers rely in order to establish the subject field for classifying purposes. However this data does not form part of the subject description and failure to realise this results in codes being applied which give misleading information. For example, lack of attention to this factor could cause such an information source title as British Technology Index to be wrongly coded as 600-26,GB An information source on the technology of Great Britain whereas the correct coding is 600 33-026,GB An index, originating in Great Britain, on technology were it decided not to use the Optional facet, the correct coding in this case would be 600. Though the above example is taken from the area of technology, ambiguity in the use of place designation in the titles of services and institutions is even more commonly encountered in the social sciences. The use of the Optional facet compels the classifier to penetrate such ambiguity in searching for the correct subject description of an item.

3.10 NOTATION

Notation is the last feature of the BSO to be dealt with in this descriptive account of the scheme, and this perhaps reflects the view of the compilers that notation is at all times to be regarded as an ancillary to the structure of the classification. The scheme was in the first place constructed independently of any notation. The present notation could be uncoupled and another used in its place without changing the character of the system, always assuming that any new notation would be no less able than the present one to handle combinations and produce the required order.

The notation given in the published BSO is intended to be read and carried in mind by human users. There could be good reasons why a notation intended to be read and stored by a machine might be rather different. The human user reacts negatively to over length and over-complexity arising from the appearance of symbols from different sets (e.g. alphabetical and numerical). Within the necessarily prescribed size limits of even variable length records, the computer is not seriously troubled by length of notation, and as all species of digits are in any case converted to numerical values for processing, a superficially mixed notation has no terrors for it. The BSO notation is believed to be tolerably brief: over 90% of the un-compounded terms cited in the schedules have codes of 5 numerical symbols length. By relying on the use of numbers as the main symbol set, and using other symbols only sparingly it manages not to he over-obtrusive. Also, by eschewing the secondary functions often accorded to notations it is capable of admitting new subjects, without limit, at their logically correct positions. It fulfils its primary function of mechanising the sequence of subjects in the schedule or in a user's file but it gives no structural information apart from that necessarily implied by the order of subjects alone. Notations of this kind are often termed 'non-expressive' notations, though it should not be overlooked that such notations do express syntactic relations. A 'non-expressive' notation was devised for 850, because all experience of the earlier established general classifications goes to show that notations which express structure, particularly hierarchy, create very difficult and sometimes insoluble problems in the insertion of new subjects in their correct places. Too often new subjects are inserted in the wrong place because of the presence of a notational gap, or the process of inserting it in the correct place involves the re-notating of the neighbouring part of the schedule. It is the dilemma embodied in these two alternatives which causes most of the decision effort and cost entailed in revising the established general classifications. This effort, and the associated delay and cost to users, should not, it was felt, be accepted as a necessity in connection with an ongoing universal switching indexing language. It was here, more perhaps than anywhere else, that the

requirements of the UNISIST switching language demanded a complete break with tradition.

It was stated above that the published notation was intended for the human user. This is not quite the same as saying that it is intended only for manual switching systems. Computer processed switching systems also have human users of switching languages at both input and output ends of the switching system. Also, it is not to say that the notation could not be fed to a computer for switching between symbols having the same meaning in different local indexing languages. However, if facilities for combining switching with interactive computer-aided search were required, it would be preferable to employ for this purpose another notation containing built-in cues enabling the machine to traverse requested search paths. For the human user, the schedule of terms itself, the conceptual pattern implicit in the manner of their ordering, their hierarchical status, and the cross-references directing to related locations in the schedule, together constitute the search aid. For computer search all these matters must be explicit in the notation. Such a fully-expressive computer-oriented notation would be far too long and complex for direct use by human beings. However, given removal of the constraints upon length and complexity necessary for the human user, such a computer-oriented notation could be as hospitable to new knowledge as is the present BSO human-oriented notation.

Arabic numerals were chosen as the base symbol-set of the BSO notation because they are the best known set with elements carrying well-understood sequence values, and because they are invariable throughout the world. The numerical characters are supplemented by two punctuation signs, the hyphen and comma, and by the Roman alphabet A to Z for occasional situations where individualisation rather than grouping is required, as for instance in specifying the names of individual artists. Some notation elements are drawn from outside coding systems, such as the ISO code for names of countries, and the Groups of the Periodic table also employ Roman numerals. The use of characters supplementary to numerals demands a fixed system of ordinal values as between the supplementary characters and numerals. The following sequence gives the recommended ordinal value system for files organised by BSO:

Spaces after last symbol of notation	
Two spaces, followed by further numerical characters	This occurs when the Optional facet for type of source is used.
- followed by further numerical characters	This is the connecting symbol for external combinations of notation.
, followed by further characters	This is a semantically empty character which introduces intercalated numbers filing between consecutive members of a notational array.
00 to 99 or 000 to 999	These two sets never occur

	together in a file in such a way as to require ordinal preference between them.
A to Z	

Turning now from the ordinal value of individual symbols to the make-up of a notation or code for a given subject, all codes begin with a member of the millesimal array 000 to 999. Between any two subjects represented by consecutive members of this millesimal array, further subjects may be interposed by adding to the first of the two consecutive numbers concerned a comma followed by a member of the two-digit centesimal array 00 to 99. In similar fashion further subjects may be interposed between consecutive numbers of the 00 to 99 array, by adding a comma followed by members of a further 00 to 99 array. Accordingly a typical code structure comprises a single group of 3 numbers followed by an indefinite number of groups of 2 numbers, all groups being separated by commas(e.g. 915,15,50.....)

In a few well-defined situations this typical 3,2,2.... pattern may be varied. In notations which contain the hyphen (external combinations, and Time and Place Facets) the 3,2,2.... pattern may appear on both sides of the hyphen. However, in many cases the hyphen links two groups of 3 numbers (e.g. 642-580 Nuclear reactor economics). Internal notation combinations contain the single separated number a as a connecting symbol (e.g. 978,0,72,37 The Koran). The numbers 088 and 890 are the leading number groups of untypical 3,3,2,2,... patterns.

3.10.1 Notational combination

A tabulation of the procedures for combining notation is given on page xiii of BSO, and is reproduced here on page 51. It should be added that in some fields where composite subjects are expected to arise frequently by comparison with unitary enumerated subjects, the schedules themselves provide for intersecting concepts by Expand like instructions. Notations produced by this mechanism are always shorter than the combined notation in which elements are linked by the connecting symbols mentioned in the tabulation. In deciding when to use Expand like instructions, the BSO Panel were obliged to balance the advantage of brevity against the two disadvantages that the Expand like mechanism is more likely to lead to indexing errors than the connection of two notational elements by a set of connecting symbols, and will also use up more of the available brief notation, thus in the long run causing an increase in the length of notation of future enumerated subjects. At the level of notation manipulation the difference between an internal combination (outside the area 600 to 890) and an Expand like instruction is that while the internal combination adds a connecting symbol and deletes the first numeral of the second notational element, an Expand like instruction omits any connecting symbol and at the same time deletes two or more of the first numerals of the second notational element. At the concept level Expand like instructions can be more versatile than internal combinations. This versatility is manifested in the BSO schedules where an Expand like instruction adds the legend 'with incorporated additions marked +'. These 'incorporated additions' are concepts which arise only in subordination to a facet combination. For instance the BSO Physics schedule consists essentially of a list of energy interactions and forms (from which the notion of a specific medium is absent), followed by a list of media or forms of matter. The Expand like instructions cause notions of forms of matter to be combined with notions of energy. Thus for instance we have 224,25 Plasmas and fluids, Mechanics. An important branch of

Plasma and Fluid Mechanics is Magnetohydrodynamics. This branch of mechanics is logically dependent upon the combination of Plasmas and fluids with Mechanics. At the level of generalised energy interactions there is probably no term to comprehend the abstract idea of mechanical motion, magnetic fields and electric fields in triangular interaction. Accordingly 224,34 Magneto-hydrodynamics is entered as an 'incorporated addition' subordinate to 224,25 Mechanics of plasmas and fluids.

3.11 SUMMARY

The Broad System of Ordering is a classification system. It is also known as a subject indication language. It is developed for a proposed world-wide information network covering the whole field of knowledge. BSO can be used

- as an aid to subject searching on the Net or any miscellaneous compilation or collection covering many subject fields
- as a subject tagging code applied to individual items or records in wide angle collections or compilations. In this application the result is an orderly and easily grasped subject arrangement of the items
- as a mediating tool in changing over from one subject indication system to another

3.12 TECHNICAL TERMS

UNISIST : United Nations Information System In Science & Technology

3.13 SUGGESTED READINGS

1. Bliss, H.E. Organization of knowledge in libraries and the subject approach to books. New York, H.W. Wilson, 1933.
2. Broughton, V. and J. Mills.(ed). Bliss Bibliographic Classification: 2nd ed. London, Butterworth, 1977
3. Brown, J.D. Subject classification. 3rd ed. revised and enlarged by J.D. Stewart. London, Grafton, 1939.
4. Coates, E., Lloyd, G. & Simandl, D. (1978). BSO: Broad System of Ordering: schedule and index. Prepared by the FID/BSO Panel. 3rd revision. The Hague : Federation Internationale de Documentation (FID) : United Nations Educational, Scientific and Cultural Organization (UNESCO). (FID Publication 564).
5. Dahlberg, I. (1977). Major developments in classification. IN: Advances in librarianship, volume 7. Edited by M. J. Voigt & M. H. Harris. New York: Academic Press, 41-103.
6. Foskett, D. J. (1975). Classification. IN: Handbook of special librarianship and information work, 4th ed. Edited by W. E. Batten. London, Aslib, 153-197.

LESSON -4

INDEXING LANGUAGES

AIMS AND OBJECTIVES

The objective of this lesson is to explain Indexing language and its definition, characteristics, advantages and disadvantages. This lesson also explains types of indexing languages with examples.

After studying this lesson you can understand

- What is Indexing language?
- Natural language vs indexing language.
- Types of indexing languages and
- Characteristics of indexing language

Structure

4.1 INTRODUCTION

One of the important functions of an information retrieval system is to match the content of documents with the user's requirements or queries. The content of each document is to be analysed and represented by terms in such a way that it becomes convenient for matching. The process of building document surrogates by assigning suitable terms to the subject content of the documents is known as "subject indexing". When the librarians or documentalists make subject approach to information, they are confronted with the difficult task of subject indexing. Subject indexing is a method of information retrieval. The subject index helps the searcher from an unclear or a rough statement to an extensive standard one. The basic objective of subject indexing is to match the content of documents with the user's queries. According to Lancaster subject indexing involves two important steps. They are Conceptual analysis and Representation. These two steps require indexers intellect. The basic objective of subjective classification is to arrange the documents on the shelves basing on the subject content. The results of the conceptual analysis are represented by artificial language or notational symbols. A number of systems by name DDC, UDC, LC, CC etc., are in use. In subject indexing the results of conceptual analysis are represented by natural language terms that represent the thought content of the document. A number of systems namely chain indexing, POPSI, PRECIS etc have been developed for preparing subject index entries of documents.

In order to solve various complex indexing problems many forms of controlled vocabulary have been developed such as thesaurus, classaurus, thesaurofacet etc. These tools provide help to both indexer while indexing and the user while formulating the query to search for documents.

4.2 INDEXING LANGUAGE

One of the important functions of information processing is to specify the subject of a document. This information is available in different parts of the document. Title usually carries this information. Sometimes titles are illusive. In such cases preface, introduction and content pages are helpful in determining the subject of the document. But whatever may be the source of information, once it is ascertained; this is to be recorded in some languages. These languages are called indexing languages. These may be natural or artificial.

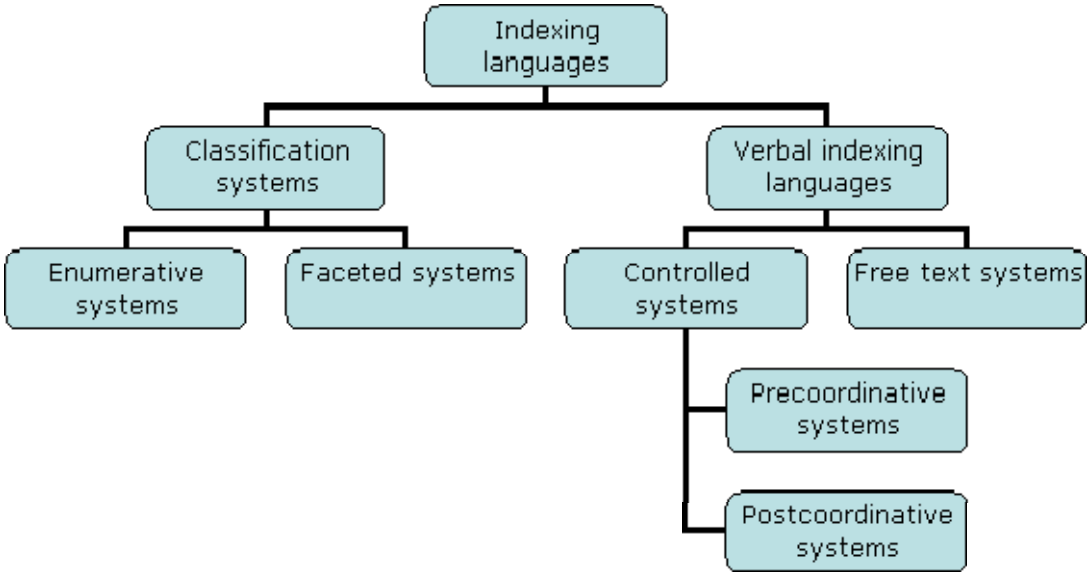
The phrase "Indexing Language" is generally defined as all the words permitted either to describe a specific document or to construct a query to search a document file, along with the rules describing how the terms are to be used and in what relation to each other in that indexing language. It is the complete list of terms in the natural language that are employed in the collection of documents. The list includes all required synonyms that are used in the process of indexing a set of documents. But it does not mean that all the terms in the list can be used to actually index the documents.

4.3 TYPES OF INDEXING LANGUAGES

An indexing language is a "language[1]" used for subject classification or indexing of documents. Indexing languages may be divided into "classification systems" and "verbal indexing languages". Lancaster argues that one should not speak of assigning classification codes as "classification" as opposed to the assignment of indexing terms as "indexing". These terminological distinctions are quite meaningless and only serve to cause confusion.

That this distinction is superficial is also evident from the fact that a classification system may be transformed to a thesaurus and vice versa.

Classification systems may be divided into enumerative systems and faceted systems. Verbal indexing systems may be divided into "controlled vocabularies" and "free text systems". Controlled vocabularies may be divided into "pre-coordinate indexing systems" and "post-coordinate indexing systems".



Traditional View of the Types of Indexing Languages

Indexing languages have been categorized into a number of fundamental types. Broadly they are designated as Assigned and Derived indexing systems.

4.3.1 Assigned Indexing:

In the Assigned Indexing, an indexer assigns terms or descriptors on the basis of subjective interpretation of the concepts implied in the documents. Assigned indexing is an intellectual method involving the finding out the specific subject of the document and assigning an appropriate subject heading. It requires more time and money at input stage, Indexer's lack of knowledge and sense in indexing will results in irrelevant output while searching. All indexing languages with vocabulary control devices such as Subject Heading lists, Thesauri and classification schemes are assigned term systems.

4.3.2 Derived Indexing:

In Derived indexing system all terms or descriptors are taken from the document itself. In other words derived term systems are almost clerical and can be easily mechanized. Author indexes, title indexes, citation indexes and natural language indexes are derived Indexing systems. For example indexes constructed by KWICK, KWOCK are derived indexes, eg. Chemical Abstracts. Science Citation index and Social Sciences Citation indexes developed by Institute of Scientific Information, Philadelphia are good examples of derived indexing.

4.3.3. Natural Language Indexing:

Any information retrieval system without vocabulary control is referred to as a "Natural Language" or sometimes as a "free–text" system because the system allows the indexer to select the terms to be used directly from the text being indexed. In automatic systems, the terms are selected by the computer. The terms are chosen from the text itself. This approach may also be called indexing by extraction. The 'Uniterm' indexing systems in the early days are the examples of natural language system.

Natural language has some advantages. Its vocabulary is up to date and it keeps on growing assimilating new concepts as soon as they come into being. The natural language has its own syntax to convey the exact meaning. Importance of natural language syntax is evident from the following example. Teachers, Students and Assessment are three significant terms in the subject of the document dealing with "Assessment of students by teachers". But these three significant terms freed from prepositions used, projects a altogether different view of the subject and not the exact one conceived by the author of the document. Correct view of the subject is obtained by the use of prepositions 'of' and 'by' in proper places. Any wrong association of the terms and prepositions may convey a different meaning. If we want to conduct a highly specific search, the natural language system is more useful. However it has some disadvantages in information retrieval. Homonyms and synonyms are the major problem which results in representing the same subject with different names. Specifying a subject is not enough for organising an index file. This is to be recorded in a system for information retrieval by users. A concept may be identified by different terms by different authors in natural language for example Mansion may be specified by the terms, 'Sky-scrapper' or 'high-rise building'. Sometimes same concept may be referred by experts and others differently. For example Philately and stamp collection are the two terms that denote same concept, while the former term is used by experts. This is the problem of synonyms in natural language reflects in scattering the same subject in different places. Besides this another major problem of natural language is Homonyms. The same spelling is sometimes used in natural language to represent different concepts, e.g.

Order (command),
Order (sequence)
Order (indent)

These terms are context dependant. In natural language the term is understood by context, but in index it is to be solved by providing context in parenthesis. Another problem is word-forms. e.g. Sail-ing. Sai-lor, Heat-ing, Heat-ed etc. These words are constructed by free morphemes with different suffixes. The user may be interested in broader, related or more specific topic for comprehensive study. This problem is not taken care by natural language while indexing language will solve by incorporating super ordinate, subordinate relations. The purpose of indexing is not only to represent the subject but also retrieval. This feature demands pre-coordination or post-coordination of terms. The different procedures adopted in libraries to perform these functions are now called indexing languages.

4.3.4. Controlled vocabulary or Artificial Language.

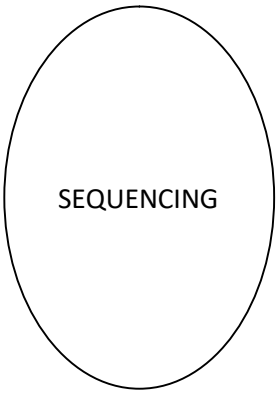
The Subject Heading Lists, Classification schemes and Thesauri are representatives of controlled indexing languages. The vocabulary of an Indexing Language in an information retrieval system exerts considerable influence on its performance. It greatly influences the construction of search strategy. The controlled vocabulary of indexing language makes the search strategy exact and precise. The problem of synonyms and

homonyms will be taken care in controlled vocabulary systems by ‘See”, ‘See also’ and ‘Use’ references.

The controlled vocabulary has indeed a number of obvious advantages. It controls synonyms and near synonyms and binges semantically related terms together. If properly constructed control vocabulary avoids many problems of false coordination or in correct term relationship. The sequencing of the terms in constructing indexes is governed by strict rules of syntax of a particular indexing technique. For example Chain procedure, PRECIS, POPSI have their own syntax to formulate subject representation of the document.

4.4 CHARACTERISTICS OF INDEXING LANGUAGE

Indexing language is designed for a special purpose. Other than subject description indexing language is also used in organising a searchable file to be used by the users. A match between the description of the document done by the indexer and description of the request made by the user only will yield a positive result. This match is possible if the file is organised in a predetermined order and the users are aware of it. Semantic and syntax of the indexing language plays an important role in successful organisation of information storage and retrieval system.



4.4.1. Vocabulary control:

Semantic aspect of indexing refers to the size of vocabulary and devices used for vocabulary control. Vocabulary of indexing language must be precise and exact. A complete one to one relationship between the concepts and terms should be established. Synonyms, homonyms, homographs etc. are controlled in indexing language. Out of the synonyms only one term is accepted in indexing language and this is used in subject description and query formulation. References are made from rejected terms to selected ones facilitating search by the users. In subject indexing this mechanism is provided with the help of ‘see’ references. In a Thesaurus this facility is provided with the help of ‘Use’ reference. In a classification scheme synonyms are guided to the notation, allotted in the scheme for the concept.

4.4.2. Coordination of concepts:

Most of the documents of present day cover compound subject comprising number of concepts. In these cases the complete subject of the document can be represented not by a single term but by a coordination or combination of number of terms. A subject class generated by coordination of two or more terms, representing different concepts, will differ from the class represented by the individual terms or by the terms in some other

combination. For example the terms Library, School and Building may be coordinated in different ways like, Library school building and School library building. The meaning of each coordinated form differs from the other. In indexing language this coordination is done at input stage in pre coordinate indexing and at searching stage in post coordinate indexing.

4.4.3 Sequencing of terms:

The coordination or combination of terms raises the problem of sequencing of terms. This sequencing or ordering of terms is very important in indexing. The four terms, a, b, c, d may be arranged in 24 possible ways. Out of these possibilities we are to choose the appropriate one, which should not only convey the correct relationship of the terms but also helpful in search. The syntax of the indexing language provides rules for sequencing of terms. However, it becomes difficult sometimes to show the correct relationship of the terms. To solve this problem Indexing language has made the provision of relational symbols known as role operators, role indicators, categories etc. in different systems. Dr. S.R. Ranganathan introduced different punctuation marks among Fundamental categories (PMEST). Similarly different role operators and role indicators are used in PRECIS and POPSI indexing systems.

4.4.4 Syntax of Indexing Language:

Importance of sequencing of terms is in all indexing languages. But the rules prescribed by them for ordering of terms differ. J. Kaiser in his Systematic indexing pointed out that compound subjects may be analysed into a combination of 'concrete', 'process' and 'time'. Thus subject formulation for 'Extraction of iron ore from Andhra Pradesh' will be rendered as:

Iron ore-India, Andhra Pradesh-Extraction

Or

India, Andhra Pradesh-Iron ore-Extraction

As these three categories suggested by Kaiser are not enough to represent modern compound subjects, S.R. Ranganathan suggested five Fundamental catgories, viz. Personality, Matter, Energy, Space and Time (PMEST). The citation order of PMEST represent decreasing concreteness of the concepts in a document. E.J. Coates has prescribed another citation order in his British Technology Index. The full citation order according to this device is Thing-Part-Material-Property-Action. E.g. 'Roofs of wood' will be represented as ROOFS : wood. J.E.L. Farradane has created a new type of syntax in indexing language. Farradane highlighted nine types of relationships that exist between each pair of terms. They are: association, comparison, concurrence, dimensional, distinctness, equivalence etc. The relationships are indicated by different symbols or operators by placing them in between the pair of terms. E.g.: 'Treatment of tuberculosis of lungs' will be represented as lungs/; tuberculosis/-treatment. The syntax of PRECIS initiated by Derek Austin is governed by the Role operators. The subject formulation representing a document is constructed with the help of the Main line operators, interposed operators, differencing operators etc. e.g. Assessment of students of medical colleges of India the concepts are Assessment (action) (operator 2) Medical colleges (key system) (operator 1)Students (part) (operator p) India(environment) (operator 0)Hence the string is : India Medical college Students Assessment

4.4.5 Rotation of Component terms

Rules of syntax of indexing language help us to formulate subject representation of a document. But in linear representation index statement can provide only a single access in

the searchable index file. To overcome this limitation, indexing languages have introduced a mechanism of rotation of component terms. The rotation is carried in such a way that each of the component terms is placed at the privileged position on the search line as lead term. Lead terms are not used in isolation but in association with other component terms. Thus context is maintained and the correct meaning of index statement is available in all such cases. This rotation of component terms is a special feature of indexing language.

In chain indexing of Ranganathan, this rotation appears in rudimentary form. It completely lacked the facility of projecting context as the terms were deleted in successive stages. Application of computer in indexing has accelerated this rotation of terms. KWICK, PRECIS and POPSI are good index systems that followed the basic principle of providing access from component terms and projecting the context, even though the format, presentation, designing of search line etc. differ from one another.

e.g. India Textile industries Skilled personnel Training

According to PRECIS the following entries will be generated from the above string of above index statement:

INDIA
Textile industries. Skilled personnel. Training
TEXTILE INDUSTRIES. India
Skilled personnel. Training
SKILLED PERSONNEL. Textile industries. India
Training
TRAINING. Skilled personnel. Textile industries. India

According to KWIC the following entries will be generated out of the above statement:

India. Textile industries. Skilled Personnel. Training
Textile industries. Skilled Personnel. Training/India
Skilled Personnel. Training/India. Textile industries
Training/ India. Textile industries. Skilled Personnel.

4.4.6 Syndatic devices:

Indexing language is an artificial one. Because of this artificiality the user needs guidance how to use it. This guidance to users is provided in indexing language with the help of various types of syntactic devices. These include guides, cross references, inversion of headings, glossaries, introduction etc. Guides and introduction to indexes provide information regarding scope and structure of the index, terminology and symbols used, rendering of headings, depth of indexing, format and typography, exceptions etc. The cross references of different types is the most important of all the syntactic devices. They correlate similar concepts scattered throughout the index due to its alphabetical overtone. The references are of two types – "See" and "See also". These references guide the users from synonym to the preferred term of the index. In thesaurus 'See' reference is provided by 'Use' reference.

e.g. Anonymous classics *See* Classics, Anonymous
Brinjal *See* Eggplant
Index terms *See* Headings

Grain crops
UF Barley

Mineralogy *See also* Petrology
Communication *See also* Telephones

4.4.7 Relational Symbols:

In a natural language correct relationship between two or more terms can easily be established with the help of prepositions, conjunctions etc. Thus photographs of albums and albums of photographs convey different meanings. This facility is absent in indexing language, which is an artificial one. To overcome this difficulty some indexing languages introduce relational symbols or indicator digits to bring out the correct relation between the words. For example Colon Classification used ", ; : . ," as role operators. British Technology Index used "= ., , " .. " as role operators.

4.4.8 Paradigmatic and Syntagmatic relations:

Paradigmatic and Syntagmatic relationships between different concepts in indexing language were introduced by J.C. Gardin. Paradigmatic relations are those which are known in advance before scanning any particular document, while syntagmatic relations are understood only after scanning a particular document. Paradigm means a set of words having the same stem. In the context of indexing language the genus-species and broader-narrower and associative relations are called paradigmatic relations. A syntagmatic relation is restricted to a particular document. On analysing the subject of a document we may establish relation of concepts covered in the document. Indexing language should be able to denote the subjects and show their relations and directions. While paradigmatic relations take care of showing the general relations, specific relations and direction of subjects are shown by syntagmatic relations.

4.4.9 Structuring of Indexing language:

Indexing languages are designed with the ultimate objective of meeting the subject approach of users. To meet this objective indexing languages consider different approaches of users. The user may need to broaden or narrow his search depending upon the availability of documents. At times users may also be interested in collateral subjects during their search for information. Indexing language has to take care of all these characteristics by showing all these relationships of components in different fashions. Thus indexing languages are structured. In an enumerative classification this relationship is focussed with the help of notations and their display. Subject headings list show 'see also' references. Thesaurus uses BT, NT, RT to show this relationship. Thus all indexing languages like thesaurus, subject headings lists and classification schemes possess the mechanism of showing relationships of subjects in one way or other and thus all of them are structured. This structuring helps in describing subject contents of documents at input stage and also during search stage by broadening or narrowing range of search.

4.5 SUMMARY

The main objective of Information storage and retrieval system is to retrieve the positive results when it searched for information. This involves representing the subject content of the documents precisely and exactly and recorded in the file to be searched as and when the user requires information. Indexing language discussed in this lesson helps in designing and establishing efficient and effective Information and storage retrieval system. Best indexing language just like natural language should have its own semantics and syntax.

4.6 TECHNICAL TERMS

PRECIS : Preserved contest Indexing System
POPSI : Postulate Based Permited Subject Indexing
PMEST: Personality, Mater, Energy, Spaces and Time

4.8 SUGGESTED READINGS

1. BR Ambedkar Open University, Information Process and Retrival (MLIS Block-II) BRA, Hyderabad 1998
2. Fosketac The subject approach to information, 4 th ed. London: clive Bingley.1982
3. Guha,B . Documentation and information; services, techniques and systems, 2nd ed. Calcutta; the work press, 1983.
4. Indiragandhi National open University, Information procession and retrieval (MLIS-3 Block -1; Unit 4; thesaurus) New Delhi IGNOU, 1985.
5. Lancaster, F.W. Vocabulary control for information Retrieval. Washing for D.C; Information resource press, 1972.
6. Riaz, Muhammad, Advanced Indexing and abstracting practices New Delhi: Atlantic publisher 1989.
7. Vickery, B-C Technique of information retrieval, London: Butterworths, 1970.

LESSON- 5
VOCABULARY CONTROL

AIMS AND OBJECTIVES

This lesson seeks to examine retrieval within two different contexts:

- a monolingual context where the language of the query is the same as the indexing language.
- Subject heading lists typically include separate listings of standardized
- A controlled list is a simple list of terms used to control terminology

Structure

5.1 INTRODUCTION

The efficiency of information retrieval system depends on the indexing language adopted. The efficiency of indexing language depends on its capability to handle two fundamentally different but interdependent types of relationship between the terms used for representing the subject matter of the documents, viz. syntactic and semantic. Syntactical relationships are built through a set of rules of the system and using a variety of relational

operators. Semantic relationships are controlled by vocabulary control device such as thesaurus.

5.2. VOCABULARY CONTROL

A controlled vocabulary is an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms and has a defined scope or describes a specific domain. The semantic aspect of indexing language is concerned with vocabulary and devices for vocabulary control. Vocabulary of indexing language is precise and exact. Synonyms, homonyms, homographs are controlled in indexing language.

5.3. PURPOSE OF CONTROLLED VOCABULARIES

The purpose of controlled vocabularies is to organize information and to provide terminology to catalog and retrieve information. While capturing the richness of variant terms, controlled vocabularies also promote consistency in preferred terms and the assignment of the same terms to similar content. Controlled vocabularies are essential and necessary at the indexing phase because without them catalogers will not consistently use the same term to refer to the same person, place, or thing. In the retrieval process, various end users may use different synonyms or more generic terms to refer to a given concept. End users are often not specialists and thus need to be guided because they may not know the correct term.

The most important functions of a controlled vocabulary are to gather together variant terms and synonyms for concepts and to link concepts in a logical order or sort them into categories. Are a *rose window* and a *Catherine wheel* the same thing? How is *pot-metal glass* related to the more general term *stained glass?* The links and relationships in a controlled vocabulary ensure that these connections are defined and maintained, for both cataloging and retrieval.

5.4 TYPES OF CONTROLLED VOCABULARIES

Most controlled vocabularies are structured vocabularies. A structured vocabulary emphasizes relationships between and among the concepts represented by the terms or names in a vocabulary.

5.4.1. Relationships in General

The term *relationship* means a state of connectedness or an association between two things in a database—in this case, fields or tables in a database for a controlled vocabulary. One important type of relationship is between equivalents; for example, *Polyglot and Multilingual* refer to the same though they are of different linguistic origin. Other relationships in a structured vocabulary include hierarchical and non-hierarchical.

5.4.2. Subject Heading Lists

Subject headings, or simply *headings,* are uniform words or phrases intended to be assigned to books, articles, or other documents in order to describe the subject or topic of the texts and to group them with texts having similar subjects. The most commonly used Subject Headings are the *Library of Congress Subject Headings* (*LCSH*), which form a

comprehensive list of preferred terms or strings, often with cross-references. Another well-known set of subject headings is the *Medical Subject Headings* (*MeSH*), which is used for indexing journal articles and books on medical science. *MeSH* incorporates a thesaurus structure with subject headings.

Subject heading lists are typically arranged in alphabetical order, with cross-references between the preferred, non-preferred, and other related headings. This emphasis on a preferred entry and links to synonyms may be found in other types of authorities. However, subject headings differ from the other vocabularies in the following fundamental way: pre-coordination of terminology is a characteristic of subject headings in that they combine several unique concepts together in a string. For example, the heading *Medieval bronze vessels* combines a period, a material, and a work type in one heading.

Subject heading lists typically include separate listings of standardized subheadings (e.g., geographic locations) that may be combined with designated headings according to prescribed rules. Various styles of subject heading displays are included in the examples below. *LCSH* displays two dashes and parentheses or periods as required, while other styles may omit punctuation or use colons or dashes for compound phrases. In *LCSH, MeSH,* and other authorities, the parts of a compound heading may be stored in separate MARC format subfields to allow variations in displays as desired.

Bicycle racing--United States
Cat family (Mammals)--Literary collections
South Africa. Arts and Culture Task Group
Architecture — Ancient Egypt
Film history: Movements and styles
Embryonic and Fetal Development
Medieval bronze vessels
Great Britain Description and travel 1801–1900

5.4.2.1. Other Headings

Other types of headings or labels may be used to uniquely identify or disambiguate one vocabulary entry from another. That is, the vocabulary record itself represents a single unique person, place, or thing, but its name is displayed with information in addition to the name. For example, the name of a creator may be listed with a short biographical string (e.g., *Flemish painter, 1423–1549*) to form a heading or label for display in a work record.

5.4.3. Controlled Lists

A controlled list is a simple list of terms used to control terminology. In a well constructed controlled list, the following is true:

- each term is unique;
- terms are not overlapping in meaning;
- terms are all members of the same class (i.e., having the same level of rank in a classification system);
- terms are equal in granularity or specificity; and
- terms are arranged alphabetically or in another logical order.

These lists are also called *flat term lists* or *pick lists,* referring to the typical method of

their implementation in an information system. Controlled lists are usually designed for a very specific database or situation and may not have utility outside that context. They are best employed in certain fields of a database where a short list of values is appropriate and where terms are unlikely to have synonyms or ancillary information. However, as with any vocabulary for cataloging, it is preferable that definitions of the terms be made available to ensure consistency among catalogers. Below is an example of a controlled list for the Classification field in a work record.

architecture	manuscripts
armor	miscellaneous
books	paintings
coins	photographs
decorative arts	sculpture
drawings	site installation
implements	texts
jewelry	vessels

The advantage of such lists is that the cataloger or indexer has only a short list of terms from which to choose, thus ensuring more consistency and reducing the likelihood of error.

5.4.4. Synonym Ring Lists

A synonym ring is a simple set of terms that are considered equivalent for the purpose of retrieval. Equivalence relationships in most controlled vocabularies should be made only between terms and names that have genuine synonymy or identical meanings. However, synonym rings are different. Even though they are classified as controlled vocabularies, they are almost always used in retrieval rather than indexing. They are used specifically to broaden retrieval (this is often referred to as query expansion): thus, synonym rings may in fact contain near-synonyms that have similar or related meanings, rather than restricting themselves to only terms with true synonymy.

Typically, synonym rings occur as sets of flat lists and are used behind the scenes of an electronic information system. They are most useful for providing access to content that is represented in texts and other instances of natural, uncontrolled language. Even though catalogers do not use synonym rings for indexing, subject experts should be involved in the creation of synonym rings for retrieval. The most successful synonym rings are constructed manually by subject matter experts who are also familiar with the specific content of the information system, user expectations, and likely searches.

In the example below, synonym rings (each represented in an individual row) represent true synonyms as well as more generic terms and other terms that are related within the specific context of a given text. The example could represent a partial synonym ring list

for a text about art depicting certain migrating birds. If a user enters *crows,* the search mechanism returns any text containing *birds* or any of the other terms in the same synonym ring as *crows.* Even though these terms are not synonyms, the implementer has judged that these links make sense for broad retrieval in this particular text. Other automated retrieval strategies may be in place as well; for example, the search algorithm may automatically truncate the *s* to allow matches in English on both singular and plural forms.

birds, avian, storks, crows, ravens, herons, Ciconiidae, Corvus, Ardeidae
migration, nonmigratory, migratory, travel, flying, altitude clouds, cumulus, nimbus, storm clouds, cloudy
wind, windy, windstorm, wind damage, air flow, jet stream

5.4.5. Authority Files

An authority file is a set of established names or headings and cross-references to the preferred form from variant or alternate forms. Following illustration is the *LCNAF(*the *Library of Congress/NACO (Name Authority Cooperative Program)) Authority File*—an authority widely used in libraries in North America.

LC Control Number: n 79003969
HEADING: Moses, Grandma, 1860-1961
000 00578cz a2200193n 450
001 1418836
005 19910703055707.6
008 790117n| acannaab |a aaa
010 __ |a n 79003969
035 __ |a (DLC)n 79003969
040 __ |a DLC |c DLC |d DLC-R
100 10 |a Moses, |c Grandma, |d 1860-1961
400 00 |a Grandma Moses, |d 1860-1961
400 10 |w nna |a Moses, Anna Mary Robertson, |d 1860-1961
400 10 |a Mōzesu, |c Guranma, |d 1860-1961
670 __ |a Her Grandma Moses ... 1946.
670 __ |a Her Guranma Mōzesu ten, 1990: |b t.p. (Grandma Moses)
952 __ |a RETRO
953 __ |a xx00 |b zz00

The *LCNAF* record for *Grandma Moses*, illustrating the established heading and cross-references for this artist.

Common types of authority files are name authority files and subject heading authority files. However, any listing of terms, names, or headings that distinguishes between a preferred term, name, or heading and alternate or variant names may be used as an authority. In other words, almost any type of controlled vocabulary—with the exception of a synonym ring list—may be used as an authority.

Authority control refers as much to the methodology as to a particular controlled vocabulary. If a controlled vocabulary is accepted by a given community as authoritative, and if it is used in order to provide consistency in data, it is being used as an authority. A local authority file is often compiled from terminology from one or more published standard controlled vocabularies.

5.4.6. Taxonomies

Taxonomy is an orderly classification for a defined domain. It may also be known as a *faceted vocabulary.* It comprises controlled vocabulary terms (generally only preferred terms) organized into a hierarchical structure. Each term in Taxonomy is in one or more parent/child (broader/ narrower) relationships to other terms in the Taxonomy. There can be different types of parent/child relationships, such as whole/part, genus/ species, or instance relationships. However, in good practice, all children of a given parent share the same type of relationship.

Taxonomy may differ from a thesaurus in that it generally has shallower hierarchies and a less complicated structure. For example, it often has no equivalent (synonyms or variant terms) or related terms (associative relationships). The scientific classifications of animals and plants are well-known examples of taxonomies. A partial display of Flavobacteria in the taxonomy of the U.S. National Center for Biotechnology Information is above.

FIGURE

5.4.7. Alphanumeric Classification Schemes

Alphanumeric classification schemes are controlled codes (letters or numbers, or both letters and numbers) that represent concepts or headings. They generally have an implied taxonomy that can be surmised from the codes. The Dewey Decimal Classification (DDC) system is an example of a numeric classification scheme with which many people are familiar, given that it is one of the two major systems used in libraries in the United States (the other is the Library of Congress Classification [LCC] system). In the Dewey system, the universe of knowledge is divided into sets of three-digit numbers. The arts are represented in the 700-number series; sculpture is represented by numbers between 730 and 739. For example, the number 735 has been established to indicate sculpture after the year 1400 ce. To that number may be added additional decimal indicators to further specify the topic by geographic or other categories. For example, 735.942 refers to sculpture dating after 1400 in England, because the extension 9 indicates geographic area, 4 indicates Europe, and 2 indicates England.

5.4.8. Thesauri

A thesaurus combines the characteristics of synonym ring lists and taxonomies, together with additional features. A thesaurus is a semantic network of unique concepts, including relationships between synonyms, broader and narrower (parent/child) contexts, and other related concepts. Thesauri may be monolingual or multilingual. Thesauri may contain three types of relationships: equivalence (synonym), hierarchical (whole/ part, genus/species, or instance), and associative.

Thesauri may also include additional peripheral or explanatory information about a concept, including a definition (or scope note), bibliographic citations, and so on. A thesaurus is more complex than a simple list, synonym ring list, or simple taxonomy.

Thesauri employ the versatile and powerful vocabulary control generally recommended for use in all databases.

5.4.9. Ontologies

In common usage in computer science, ontology is a formal, machine-readable specification of a conceptual model in which concepts, properties, relationships, functions, constraints, and axioms are all explicitly defined. Such ontology is not a controlled vocabulary, but it uses one or more controlled vocabularies for a defined domain and expresses the vocabulary in a representative language that has a grammar for using vocabulary terms to express something meaningful. Ontology generally divides the realm of knowledge that they represent into the following areas: individuals, classes, attributes, relations, and events. The grammar of the ontology links these areas together by formal constraints that determine how the vocabulary terms or phrases may be used together. There are several grammars or languages for ontology, both proprietary and standards-based. Ontology is used to make queries and assertions.

Ontology has some characteristics in common with faceted taxonomies and thesauri, but ontology use strict semantic relationships among terms and attributes with the goal of knowledge representation in machine-readable form, whereas thesauri provide tools for cataloging and retrieval.

Ontology is used in the Semantic Web, artificial intelligence, software engineering, and information architecture as a form of knowledge representation in electronic form about a particular domain of knowledge.

FIGURE

In the example above, each item in the ontology belongs to the subclass above it. Items can also belong to various other classes, although the relationships may be different. For example, a watercolor is a painting, but it may also be classified as a drawing because it is a work on paper. Van Gogh's *Irises* could be classified with oil paintings (with the relationship type *medium is*) but also with Post-Impressionist art (with relationship type *style/period is*). Relationships in ontology are defined according to strict rules, which are different than the equivalence, hierarchical, and associative relationships used for thesauri and other vocabularies.

5.4.10. Folksonomies

Folksonomy is a neologism referring to an assemblage of concepts represented by terms and names (called *tags*) that are compiled through social tagging. *Social tagging* is the decentralized practice and method by which individuals and groups create, manage, and share tags (terms, names, etc.) to annotate and categorize digital resources in an online social environment. This method is also referred to as *social classification, social indexing, mob indexing,* and *folk categorization.* Social tagging is not necessarily collaborative, because the effort is typically not organized; individuals are not actually working together or in concert, and standardization and common vocabulary are not employed.

Folksonomies do not typically have hierarchical structure or preferred terms for concepts, and they may not even cluster synonyms. They are not considered authoritative because they are typically not compiled by experts. Furthermore, they are by definition not applied to documents by professional indexers. Given that it is impossible for the large and

varied community of creators and users of Web content to independently add metadata in a consistent manner, folksonomies are generally characterized by nonstandard, idiosyncratic terminology. Although they do not support organized searching and other types of browsing as well as tags from controlled vocabularies applied by professionals, folksonomies can be useful in situations where controlled tagging is not possible: they can also provide additional access points not included in more formal vocabularies. There may be great potential for enhanced retrieval by linking terms and names from folksonomies to more rigorously structured controlled vocabularies.

5.5 SUMMARY

Once information is ascertained, it is to be recorded in some indexing language. Indexing are two types. They are natural language and vocabulary controlled. The lists of subject headings, classification schemes, and thesaurus are representative of controlled vocabulary indexing language. Controlled vocabulary makes the indexing language precise and exact. The indexing languages adopt various devices for vocabulary control. Controlled vocabularies are essential for effective Information storage and retrieval systems.

5.6 TECHN ICAL TERMS

LCSH : Library of Congress Subject Headings
MESH : Medical Subject Headings
LCC : Library of Congress Classification
NACO : Name Authority Cooperative Programme

5.7 SUGGESTED READINGS

1. Fosketac The subject approach to information, 4 th ed. London: clive Bingley.1982
2. Guha,B . Documentation and information; services, techniques and systems, 2nd ed. Calcutta; the work press, 1983.
3. Lancaster, F.W. Vocabulary control for information Retrieval. Washing for D.C; Information resource pross, 1972.
4. Vickery, B-C Technique of information retrieval, London: Butterworths, 1970.

LESSON - 6

THESAURUS CONSTRUCTION

AIMS AND OBJECTIVES

Thesaurus, as a vocabulary control devise, has been identifies as an important tool in information processing and retrieval. The present lesson aims to provide an overview of its structure, functions and constriction. After studying this lesson, you will be in a position to explain the need, purpose and functions of a thesaurus.

- Describe the structure of a thesaurus
- Explain the construction of a thesaurus through manual methods as well as computerized systems.
- Discuss the role of thesaurus in information storage and retrieval systems

Structure

6.1. INTRODUCTION

A thesaurus is a tool for vocabulary control. By guiding indexers and searchers about which terms to use, it can help to improve the quality of retrieval. Usually, a thesaurus is designed for indexing and searching in a specific subject area indicating preferred terms, non-preferred terms, and semantic relations between terms; the terms are in ordinary human language.

The word "thesaurus" is derived from Greek and Latin words, which means a "Treasury". According to the Oxford English Dictionary, the earliest usage of word thesaurus was known in 1565 from the title "*Thesaurus Language Romance et Britanie*". Modern usage may be said to date from 1852 when the first edition of the *Thesaurus of English words and phrases* was published by Peter Mark Roget. The addition of "and phrases" in the title had great significance. It is considered as the "Father of all thesauri". Roget himself described his work as "a collection… arranged not in alphabetical order…but according to the ideas which they express…the objective aimed at, the idea being given, to find the word or words, by which the idea may be fitly and aptly expressed. Obviously Roget's thesaurus has nothing to do with information retrieval, but his novel idea was profitably utilized in compilation of modern information retrieval thesauri.

The concept of thesaurus in Library and information Science used by Helen Brownson for the first time in connection with information retrieval at the Dorking conference on classification on May 14th 1957. The first thesaurus used in information retrieval system was developed by Du Pont in USA around 1959 and since then a large number of thesauri have been brought out in different subject fields.

6.2 THESAURUS – DEFINITIONS

The important definitions of thesaurus and meaning of them are as follows:

The 'Oxford English Dictionary' defines the thesaurus as a "Treasury" or "Store house of Knowledge".

'Webster's Third New English Language thesaurus' defines thesaurus as "a book which contains a store of words about a particular field or set of concepts, specifically, a dictionary of synonyms".

The second revised edition of the 'Guidelines for the establishment and development of monolingual thesauri' defines the thesaurus as the vocabulary of a controlled indexing language, formerly organized so that the "priori relationship between the concepts are made explicit."

All the above definitions show that the thesaurus means a treasury or store house in its general usage.

The usage of the term thesaurus, in Library and Information Science reference tools denotes a special function in information retrieval through controlling the terminology."

According to Kent, it is "a compilation of terms of a given information retrieval system's vocabulary, arranged in some meaningful and which provides information relating to each term.

A thesaurus can be defined either in terms of its function or its structure. Functionally a thesaurus is a terminological control device used in representing the subject content of the documents. In terms of its structure, a thesaurus a controlled and dynamic vocabulary of semantically and generally related terms which covers a specific domain of knowledge.

6.3 PURPOSE OF THESAURUS

The major purposes of a thesaurus include the following:

1. To provide a map of a given field of knowledge, indicating how concepts or ideas about concepts are related to one other, which helps an indexer or a searcher to understand the structure of the field.
2. To provide a standard vocabulary for a given subject field.
3. To provide a guide for users of the system so that they choose the correct term for a subject search; this stress the importance of cross references. If an indexer uses more than one synonym in the same index for example; "abroad" "Foreign" and "overseas" – then documents are liable to be scattered under all these; a searcher who chooses one and finds document indexed there, will assume that he has found the correct term and will stop his search without knowing that there are other useful documents indexed under the other synonyms.
4. To provide classified hierarchies so that a search can be broadened or narrowed systematically.

6.4 INTERNAL STRUCTURE

The arrangement of different components of an entry and the arrangement of different entries in relation to one another constitute the structure of a thesaurus. Cross references makes explicit the ways in which entries relate to each other in a network of concepts.
The different terms in an entry are displayed in the following format:

Descriptor
(with scope notes wherever needed)
Synonyms and quasi-synonyms
(displaying equivalence relationships)
Broader terms
(displaying hierarchical – super-ordinate relationships)
Narrower terms
(displaying hierarchical – sub-ordinate relationships)
Related terms
(displaying Associative relationships)
Top term
(displaying the broader class of the descriptor)

The basic elements in a thesaurus are the individual words, terms, or phrases and these are often called "descriptors" or Keywords". The descriptors are arranged in thesaurus in alphabetical order. The indexer assigns the descriptor terms to describe the content of documents. The terms which are not preferred to be used in indexing are "Non-descriptors". They are proper names of corporate bodies, government agencies, institutions and firms, geographical names etc.

6.4.1 Relationships Between Terms

One of the key points of a thesaurus is that it indicates the relationship among terms. They are :

1. Hierarchical or structural Relationship.
2. Equivalent or preferential Relation.
3. Associate or Affinitive Relation.

The symbols to express the relationships in thesauri have become more or less standardized as follows;

SN Scope Note.
USE Equivalent to "see" reference.
UF Use for, the reciprocal of use.
BT Broader Term.
NT Narrow Term.
RT Related Term.

For example:

INTELLIGENCE
BT: Ability.
NT: Comprehension
RT: Aptitude

By having these relationships displayed, both the indexer and the searcher are in a better position to cover the full range of options that may be possible in either indexing or searching.

6.4.1.1 Hierarchical relationships:

The hierarchical relation expresses super/subordinate of concepts. There are two types of relationships in this category; Genus – Species and Whole – Part relationships. This relationship is indicated in order that users may make the transition from a first access point to related terms or access point. It helps the indexer to select the most specific terms to describe a concept that is available in the thesaurus. The relationships are displayed in the thesaurus by using the symbol BT and NT.

e.g. SNAKES – COBRA
TELECOMMUNICATION – TELEGRAPHY

HUMAN BODY – CHEST
TELEVISION – PICTURE TUBE

6.4.1.2 Equivalence or Preferential Relationships:

The equivalence relationship may be described as "the relationship between preferred and non-preferred terms in an indexing language in which each of two or more terms is regarded for indexing purpose, as referring to the same concept. In other words it is the relationship between synonyms and/or quasi-synonyms. These generally indicate preferred terms or descriptors and thus perform the necessary function of controlling index vocabulary, by specifying which terms are to be used in the index and which are not. Preferential relationship in thesaurus is indicated by UF (use for) and USE.

e.g. CYTOLOGY
UF Cell Biology
Cell Biology
Use CYTOLOGY

6.4.1.3 Associative or Affinitive Relationships:

Affinitive relationships exist between terms that are not necessarily connected to one another in any fixed hierarchical manner. This relationship in a thesaurus is displayed by the

symbol RT (Related Term). Associative relationship is a relationship which is neither hierarchical nor equivalent.

For example, instruction is a related concept to education, Teaching, and Courses:

Instruction
RT Education
Teaching
Courses

The UNISIST Guidelines State that "the associative relation is usually employed to cover the other relations between concepts that are related but are neither consistently hierarchical nor equivalent ".
An alternative means of expressing an affinitive relationship is found in MESH, i.e. Medical Subject headings. The term "See" relation is the equivalent of "RT".
Most thesauri make a clear distinction. But when BT and NT are reasonably simple to identify, related terms present a much more complex issue and no system has yet succeeded in defining precisely which terms should be enumerated as RT to any other terms.
Thus a thesaurus provide, the control of terminology by showing a structured display of concepts supplying for each concept all terms that might express that concept and presenting the associative and hierarchical relationships of the vocabulary. A list of terms which do not include structural and relational information is not a thesaurus. It is merely an alphabetical list of descriptors or subject headings.

6.5 PARTS OF THESAURUS

A thesaurus usually has at least two major parts. They are main part and an auxiliary part.

6.5.1 Main part:

The main part in a thesaurus is normal alphabetical list of all descriptors giving complete information of each descriptor, including the concept relationship. This part includes both descriptors and non-descriptors along with scope notes and definitions.

6.5.2 Auxiliary part:

In order to improve the access to the main part a thesaurus may contain several auxiliary parts. i.e., permutation subject index, systematic listings like Hierarchical Index, and subject category Index as in TEST (Thesaurus of Engineering and Scientific Terms). The thesaurus may also contain a faceted classification along with alphabetical terms, which is known as Thesaurofacet.

6.6 CONSTRUCTION OF THESAURUS

Thesauri have been compiled in a variety of different ways.

Aitcheson and Gilchrist have given some practical steps for construction of a thesaurus. They are:

- Identify the needs of the users.
- Define the subject field.
- Decide the type and design of thesaurus layout.
- Collect terms from subject literature, users, and specialist.
- Screen and edit the terms as per the rules of the thesaurus.
- Record the terms in the term cards thesaurus form.

- Sorting and grouping of thesaurus cards.
- Prepare a hierarchical structure and other associated parts.
- Test the thesaurus against a selected collection of document.
- Get the thesaurus evaluated by subject specialists and users.

Let us further discuss these ten steps in detail:

1. Identification of the subject field: The nature of the users, their, type, the needs of a library and information system influence the construction of terminology.
2. Definition of subject field: the boundaries of topic coverage and the depth to which various aspects of the subject are to be treated may be decided. Study of the general or technical thesauri already available may give some idea about the main sections of the subject. Form and geographical divisions could be taken from any major classification system or thesauri.
3. Decide the type and design of the thesaurus layout: what amount of natal is expected from thesaurus? At this stage, the compiler should clarify his ideas on the type of thesaurus to be constructed. Decision should be taken on the characteristics of thesaurus hierarchical and other interrelation ships of the terms.
4. Collect the terms from subject literature, users and subject specialists. Once the design and layout are decided, the collection of terms for these subject areas may be taken up. Sources like descriptor lists or thesauri, subject heading lists, classification schemes, and nomenclatures of indexes of journals and obstructing periodical etc. should be scanned for an initial list of the terms.
5. Screening and editing the terms as per rules of thesaurus construction: Soergel suggests that the indexing would be most effective if done by a number of different subject experts in the same field using terms of their choice. The subject experts may be shown the list of terms in their own subject fields and asked to comment, making amendments and adding term. This will help to screen and edit the terminology. The rules of thesaurus construction (Ex. UNISIST Guidelines for the establishment and Development of monolingual Thesauri) as decided in the beginning should be followed consistently.
6. Recording of terms: In a machine-selected thesaurus, the listing of terms will be printed or displayed by the computer. If the thesaurus includes humanly selected terms, it is convenient to record term on a card. Each term written on card should show RT BT NT VF etc and SN when necessary to determine the status of term for inclusion.
7. Sorting and Grouping of thesaurus cards: All the thesaurus cards are to be sorted out and grouped according to their subject groups and sub-groups. Duplicate entries are to be eliminated in the preliminary scanning.
8. Prepare the Hierarchical structure and other Associated posts: From the group of terms, inter term relationships are to be identified and hierarchical structures are to be developed among the descriptors. Facet analysis helps to identity and display of underlying structures.
9. Test the thesaurus against a selected collection of Documents: any thesaurus, thus compiled should be tested against a selected collection of documents of the concerned subject field to examine its efficiency and use in information retrieval.
10. Evaluation of thesaurus: user satisfaction may be helpful in determining the quality of a thesaurus; based on the feedback of the user the thesaurus could be refined.

6.6.1 Salton's five principles of thesaurus construction:

1. No very rare concepts should be included in the thesaurus since they could be expected to produce many matches between documents and search requests.
2. Very common high frequency terms should also be excluded from the thesaurus.

3. Non-significant words should be studied carefully before they are included in the list of words to be eliminated.
4. Ambiguous terms should be included only for the senses that are likely to be present in the document collection to be treated.
5. Each concept class should include only terms of roughly equivalent frequency so that the matching characteristics are approximately the same for each term within a category.

6.6.2 Relationships among terms:

Thesaurus is a controlled list terminology for a given subject area and shows relationships among the terms.
These may be categorized as
i. hierarchical Relationship; and
ii. Non- hierarchical relationship.
In hierarchical relationship, there are two types, namely, genus species and part- whole. The non- hierarchical relations can be further divided into equivalence and associative relationship.
These relationships can be shown in the following chart;

Concept relationship

↓

↓	↓
Hierarchical Relationship (HR)	Non-Hierarchical relationship (NHR)
Genus Species Relationship	equivalence Relationship
Part-whole Relationship	Associate Relationship

6.6.2.1 Hierarchical Relationships:

The HR for a given concept arises from its super ordinate and subordinate links that is either genus species or whole- part. In the thesaurus the super ordinate link is represented as Broader Term (BT) and the subordinate link as Narrower Term (NT).

Example:
Genus – species hierarchical relationship:
RADIATION
NT Electromagnetic radiation
ELECTRONIC RADIATION
BT Radiation
Whole – part hierarchical relationship:
INDIA
NT Tamilnadu
TAMILNADU
BT India

6.6.2.2 Non – Hierarchical Relationship (NHR)

6.6.2.2.1 Equivalence Relationship:

The equivalence relationship implies the control of the synonymy through preferred and non- preferred terms. The preferred terms are the terms used consistently to represent the concepts when indexing. The non-preferred terms are the synonyms or quasi – synonyms of terms, which are not is used for preferred terms and UF (Used for) for non- preferred terms

Ex: DEVELOPING COUNTRIES
UF Underdeveloped Countries.
Underdeveloped Countries
USE DEVELOPING COUNTRIES.

6.6.2.2.2 Associate Relationship:

The association relationship is the non- hierarchical relationships among the preferred terms.

Two kinds of terms can be linked by the associative relationship:

a) Those that belong to same category, and
b) Those belonging to different categories.
a) Terms belonging to the same category:

Example: SHIPS	BOATS
BT: VEHICLES	BT:VEHICLES
RT: BOATS	RT SHIPS

b) Terms belonging to different categories:

1. Process and resulting product.
i. Ex: Cooking
ii. RT Food.
2. Situation or condition and what may occur in that situation
i. Ex: Inflation
ii. RT Price rise.
3. An action and the product of the action:
i. Ex: WEAVING
ii. RT: Cloth
4. Effect and cause
i. Ex. Poverty
ii. RT Unemployment

5. Process and person usually associated with the process
Ex. Administration of Justice
RT Judge

6.7 ROLE OF THESAURUS IN INFORMATION RETRIEVAL

Thesauri have been used for over decodes to improve precession and recall in information retrieval. The performance of an information retrieval system can be improved by the use of controlled vocabularies from a thesaurus in the concerned field of knowledge. Thesauri will have a keystone role to play in systems of integrated database. Databases cannot be regarded as integrated unless vocabulary can be used to achieve a degree of compatibility.

6.8. FOLLOWING IS A SHORT LIST OF WORLD FAMOUS THESAURUS

1. Thesaurus of engineering and scientific terms (TEST).
2. Thesaurofacet. A thesaurus and façade classification for engineering and related subjected.
3. Information retrieval thesaurus of education terms.
4. Roots thesaurus.

5. OECD macro thesaurus.
6. UNESCO thesaurus.
7. INIS thesaurus.
8. MESH Medical subject heading.
9. SPINES thesaurus published by UNESCO.
10. INSPEC Thesaurus.

6.9 SUMMARY

The efficiency of an information retrieval system largely depends on the indexing language adopted. The efficiency of indexing language depends on its capability to handle two types of relationship between the terms used for representing the subject matter of the documents, viz. syntactic and semantic. Semantic relationships are controlled by a vocabulary control device such as thesaurus. The indexer and the user often do not use the same language. Thesaurus provides for indispensible link between their vocabularies, thereby eliminating both 'redundant indexing' and 'redundant searching.

6.10 TECHNICAL TERMS

HR : Hierarchical Relationship
BT : Broader Term
NR : Narrow Term
RT : Related Term
TEST: Thesaurus of Engineering and Scientific Terms

6.11 SUGGESTED READINGS

1. Aitchison,J and Gilchrist. Thesaurus construction: a practical manual. London: ASLIB 1972.
2. Fosket, A.C. The subject approach to information, 4 th ed. London: clive Bingley.1982
3. Guha, B . Documentation and information; services, techniques and systems, 2nd ed. Calcutta; the work press, 1983.
4. Tonley, Helen M. and Ralph.D Gee. Thasaurus making, London; Andrew dutsch 1980.
5. Indira Gandhi National Open University, Information procession and retrieval (MLIS-3 Block -1; Unit 4; thesaurus) New Delhi IGNOU, 1985.
6. BR Ambedkar Open University, Information Process and Retrival (MLIS Block-II, Unit 8 ; Thesaurus – its structure, functions and construct) BRA, Hyderabad 1998
7. Riaz, Muhammad, Advanced Indexing and abstracting practices New Delhi: Atlantic publisher 1989.
8. Lancaster, F.W. Vocabulary control for information Retrieval. Washing for D.C; Information resource press, 1972.
9. Vickery, B-C Technique of information retrieval, London: Butterworths, 1970.
10. UNESCO: Guidelines for the Establishment and development of monolingual thesauri, 2nd Rev. ed. Paris: UNESCO, 1973.

LESSON - 7
I S B D

AIMS AND OBJECTIVES

The objective of this lesson is to explain the basic features of International Standard Bibliographic Description (ISBD). It explains the evolution of ISBD for different forms of library materials. The basic structure of the ISBD is clearly dealt with in this lesson. The organization of different bibliographic elements and their order with prescribed punctuation is explained in detail.

After studying this lesson you will be able to

- What is ISBD
- History of ISBD
- Outline of ISBD

Structure

7.1 INTRODUCTION

The **International Standard Bibliographic Description** (**ISBD**) is intended to serve as a principal standard to promote universal bibliographic control. ISBD is a set of rules produced by the International Federation of Library Associations and Institutions (IFLA) to create a bibliographic description of all published literature, in a standard, Internationally acceptable human-readable form, especially for use in a bibliography or a library catalog. The ISBD main goal is to offer consistency when sharing bibliographic information. The ISBD is the standard that determines the data elements to be recorded or transcribed in a specific sequence as the basis of the description of the resource being catalogued. In addition, it

employs prescribed punctuation as a means of recognizing and displaying these data elements and making them understandable independently of the language of the description. A consolidated edition of the ISBD was published in 2007 and revised in 2011, superseding earlier separate ISBDs for different types of documents such as monographs, serials, cartographic materials, electronic resources, non-book materials, and printed music. IFLA's ISBD Review Group is responsible for maintaining the ISBD.

.

7.2 HISTORY OF ISBD

The International Standard Bibliographic Description (ISBD) dates back to 1969, when IFLA Committee on Cataloguing sponsored an International Meeting of Cataloguing Experts. The first ISBD(M) for Monographs appeared in 1971. The ISBD(S) for Serials was published in 1974. In the same year 'First standard edition' of ISBD(M) was also published incorporating the suggestions received from experts. The ISBD(G) a general International standard bibliographic description suitable for all types of materials was published in the year 1977. The ISBD(M) was also revised and published as 'First standard edition revised' in the year 1978. The other ISBDs as detailed below are published subsequently:

ISBD(CM) for cartographic materials	1977
ISBD(NBM) for non-book materials	1977
ISBD(S) revised for Serials	1977
ISBD(A) for Antiquarian (older monogrphs)	1980
ISBD(PM) for printed music	1980

7.2.1 Revision

The IFLA committee met in 1977 and decided to review all ISBDs for every five years and an ISBD Review committee was formed. This committee met in 1981 to plan for reviewing and revising the ISBDs. Consequently the ISBDs were republished as follows:

ISBD(M)	1987
ISBD(CM)	1987
ISBD(NBM)	1987
ISBD(S)	1988
ISBD(CF) for computer file	1990
ISBD(A)	1991
ISBD(PM)	1991
ISBD(G)	1992

ISBD(ER) for Electronic resources in 1997 as replacement of ISBD(CF)

Following publications are the outcomes of "Second General Review project"

ISBD(CR) for serials and other continuing resources in 2002 as a replacement of ISBD(S)

ISBD(M)	2002
ISBD(G)	2004

A consolidated edition of the ISBD was published in 2007 and revised in 2011, superseding earlier separate ISBDs for different types of documents such as monographs, serials, cartographic materials, electronic resources, non-book materials, and printed music. IFLA's ISBD Review Group is responsible for maintaining the ISBD.

7.3 SCOPE, PURPOSE AND USE

7.3.1 Scope

Area	Prescribed punctuation	Element	Usage	Repeatability
1. Title and statement of responsibility area		1.1Title proper	M	R
	[ ]	1.2 General material designation. GMDs	O	
	=	1.3 Parallel title	C	R
	:	1.4 Other title	C	R
		1.5 Statements of Responsibility		
	/	First Statement	M	R
	;	Subsequent statements	C	
2: Edition area		2.1 Edition statement	M	R
	=	2.2 Parallel edition	O	
	/ ;	2.3 Statement of responsibility relating to edition First statement Subsequent statement	M O	R
3.Material or type of resource specific area			M	R
4.Publication,production, distribution, etc., area		4.1 Place of publication	M	
	:	4.2 Name of publisher	M	
	,	4.3 Date of publication	M	
5.Physical description area		5.1 Specific material designation and extent	M	
	:	5.2 Other physical details	M	
	;	5.3 Dimensions	M	

	+	5.4Accompanying material	O	R
6: Series area		6.1 Series	M	
	=	6.2 Parallel title of the series	C	R
	:	6.3 Other title of the series	C	R
	/	6.4 Statement of responsibility		
	;	6.5 ISSN	C	R
	,	6.6 Numbering within series	M	
7: Notes area			C	R
8.Resource identifier (e.g. ISBN, ISSN) and terms of availability area		8.1 Resource identifier		
	=	8.2 Key title		
	:	8.3 Terms of availability and price		
	( )	8.4 Qualifications		

ISBD specifies the requirements for the description and identification of the most common types of published resources that are likely to appear in library collections. It also assigns an order to the elements of description and specifies a system of punctuation for description of resources.

7.3.2 Purpose

The primary purpose of the ISBD is to provide stipulations for compatible descriptive cataloguing worldwide in order to facilitate International exchange of bibliographic records.

ISBD aims to:

- make records from different sources interchangeable
- assist in the interpretation of records across language barriers
- assist in the conversion of bibliographic records to electronic form
- enhance interoperability with other content standards

7.3.3 Use

The ISBD provides maximum number of elements to describe the resources for various bibliographic activities. However, ISBD designated three categories of elements as detailed below:

Mandatory elements: These are required in all bibliographic activities and presence of these elements is compulsory.

Optional elements: These elements may be included or omitted depending up on the discretion of the bibliographic agency.

Conditional: These elements are required under certain conditions. If the condition does not warranty the presence of these elements, their use is optional.

7.4 STRUCTURE OF AN ISBD RECORD

The ISBD prescribes eight areas of description. Each area, except area 7, is composed of multiple elements with structured classifications. Elements and areas that do not apply to a particular resource are omitted from the description. Standardized punctuation (colons, semicolons, slashes, dashes, commas, and periods) is used to identify and separate the elements and areas. The order of elements and standardized punctuation make it easier to interpret bibliographic records when one does not understand the language of the description.

7.4.1 Outline of the ISBD

7.4.2 Punctuation

Each element of the description, except the first element of area 1, is either preceded or enclosed by prescribed punctuation. Prescribed punctuation is preceded and followed by space except for coma (,) and point (.) which are only followed by space.

Each area of the ISBD other than area 1 is preceded by a point, space, dash, space (. -). Each area can also be written as separate paragraph without ant preceding punctuation.

When the first element of an area is not present in a description, the prescribed punctuation of the first element that is present is replaced by a point, space, dash, space (. -), preceding the area.

When an element or area ends with a point and the prescribed punctuation for the element or area that follows begins with a point, both points are to be given

Eg. 3rd ed.. –
Not 3rd ed. –

The mark of omission of information is indicated by three points (…). The mark of omission is preceded and followed by a space.

7.4.3 Sources of Information

For all types of material the resource itself constitutes the basis of the description. The ISBDs provided the preferred sources of information for describing all types of resources. For monographs the title page is the preferred source of information.

7.4.4 Example Records of ISBD (M)

A typical ISBD records:

Record 1

A manual for writers of research papers, theses, and dissertations : Chicago style for students and researchers / Kate L. Turabian ; revised by Wayne C. Booth, Gregory G. Colomb, Joseph M. Williams, and University of Chicago Press editorial staff. — 7th ed. — Chicago : University of Chicago Press, 2007. — xviii, 466 p. : ill. ; 23 cm. — (Chicago guides to writing, editing, and publishing). — Includes bibliographical references (p. 409-435) and index. — ISBN 978-0-226-82336-2(cloth : alk. paper) : USD35.00. — ISBN 978-0-226-82337-9 (pbk. : alk. paper) : USD17.00

Record 2

Theory of classification / Krishan Kumar. – Delhi : Vikas, 1979. – xii, 510p ; 22 cm.

ISBN 0 7069 0797 3 Cloth: Rs. 60

Record 3

Education and the social order / Bertrand Russel. – Revised and enlarged ed. – London : George Allen & Unwin, 1975. – viii, 301p ; 22 cm.

Previous edition published as: 'Education and the modern world.'

London : Van Nostrand, 1932

ISBN 0 486 22876 9 Paperback: $5.00

7.5 SUMMARY

The application of computer has brought home the great importance of uniformity; precision; and compatibility of bibliographic tools like library catalogues. This emphasizes the need for standardization. International Standard Bibliographic Descriptions (ISBDs) developed by IFLA serve a good example of an attempt towards uniform cataloguing practices and achieving successful and convenient international exchange of bibliographic information in written as well as machine-readable form. The acceptance of ISBD by all libraries at national and international level is one step towards making Universal Bibliographic Control (UBC) a reality

7.6 TECHNICAL TERMS

ISBD : International Standard Bibliographic Description

IFLA : International Federation of Library Association and Institutions

7.7 SUGGESTED READINGS

1. Chan, Lois Mai. Cataloging and Classification: an introduction. New York: McGraw-Hill Humanities, 1994.
2. International Standard Bibliographic Description (ISBD). Preliminary consolidated ed. München: K.G. Saur, 2007. (IFLA series on bibliographic control, vol. 31)
3. Svenonius, Elaine. The Intellectual Foundation of Information Organization. Boston: The MIT Press, 2000.
4. Willer, Mirna; Dunsire, Gordon; Bosancic, Boris (2010). "ISBD and the Semantic Web". JLIS.it (University of Florence) 1 (2).doi:10.4403/jlis.it-4536. Retrieved 29 June 2013.

LESSON - 8
A A C R 2

AIMS ABD OBJECTIVES

The objective of this lesson is to explain the basic features of Anglo-American Cataloguing Rules 2. It explains the basic structure of the code, organization of description of the library materials and levels of description etc.

After studying this lesson you will be able to

- What is AACR2
- Governance of AACR
- Organization of description of documents
- Levels of description recommended by AACR2
- structure of AACR2

Structure

8.1 INTRODUCTION

The **Anglo-American Cataloguing Rules** (AACR) are a national cataloging code first published in 1967. The rules in the code are based on "Statement of Principles" adopted by the International Conference of Cataloguing Principles in 1961. The rules in AACR1 have been formulated primarily to meet the requirements of general research libraries. Though it is not developed as an International code, it is being widely used in various countries. AACR2 stands for the *Anglo-American Cataloguing Rules, Second Edition*. Despite the claim to be 'Anglo-American', the first edition of AACR was published in 1967 in somewhat distinct North American and British texts. The second edition of 1978 unified the two sets of rules (adopting the British spelling 'catalog**u**ing') and brought them in line with the International Standard Bibliographic Description.

AACR2 exists in several print versions, as well as an online version. Gorman has edited several revisions of AACR2 including a concise edition. Print versions are available from the publishers. The online version is available only via Cataloguer's Desktop from the Library of Congress. Various translations are also available from other sources.

Principles of AACR include cataloguing from the item 'in hand' rather than inferring information from external sources and the concept of the 'chief source of information' which is preferred where conflicts exist.

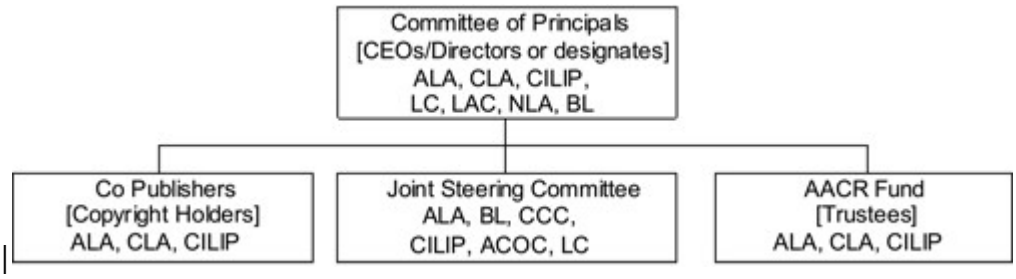
Over the years AACR2 has been updated by occasional amendments, and was significantly revised in 1988 (2nd edition, 1988 revision) and 2002 (2nd edition, 2002 revision). The 2002 revision included substantial changes to sections for non-book materials. A schedule of annual updates began in 2003 and ceased with 2005. AACR is published in English and has been translated into other languages

AACR2 has been succeeded by Resource Description and Access (commonly referred to as RDA), which was released in June 2010. This new code is informed by the Functional Requirements for Bibliographic Records and was conceived to be a framework more flexible and suitable for use in a digital environment. In the fall of 2010, the Library of Congress, National Library of Medicine, National Agricultural Library, and several other institutions and national libraries of other English-speaking countries performed a formal test of RDA, the results of which were released in June 2011.

8.2 AACR GOVERNANCE STRUCTURE

AACR2 is published under the auspices of the AACR Fund**.** The publication of the Anglo-American Cataloguing Rules (AACR) is governed by the Committee of Principals, which coordinates three subordinate groups: The Co-Publishers of AACR, The Joint Steering Committee for Revision of AACR, and The AACR Fund Committee (Trustees). Following is the AACR governance structure:

AACR Governance Structure



The **Anglo-American Cataloguing Rules2** (AACR2) are designed for use in the construction of catalogues and other lists in general libraries of all sizes. The rules cover the description of, and the provision of access points, for all library materials commonly collected at the present time. The second edition of the Anglo-American cataloguing Rules appeared in 1978 is based on a reconciliation of the British and North American texts of the 1967 edition. It is published jointly by the American Library Association, the Canadian Library Association, and the Chartered Institute of Library and Information Professionals in the UK. The editor is Michael Gorman. AACR2 is designed for use in the construction of catalogues and other lists in general libraries of all sizes. The rules cover the description of, and the provision of access points for, all library materials commonly collected at the present time.

8.3 ORGANIZATION OF DESCRIPTION

The description is divided into the following areas:

Title and statement of responsibility

Edition

Material Specific details

Publication, distribution, etc

Physical description

Series

Note

Standard number and terms of availability

Each area is further divided into a number of elements.

8.4 LEVELS OF DETAIL IN DESCRIPTION

The AACR2 code has prescribed three levels of description. First level provides the minimum information which is necessary to identify a given document. Second level can be called standard description. It provides all the data which may be considered necessary for the description of documents. The third level provides information covering every descriptive element described in the code. The choice of level of description would depend uponj the purpose to be satisfied by a given catalogue in the library.

8.5 STRUCTURE OF AACR2

The structure of AACR2 is based on ISBD, which make distinction between bibliographic description and the access points. The AACR2 consists of two parts. Part I deals with the provision of information describing the item being catalogued, and Part II deals with the determination and establishment of headings (access points) under which the descriptive information is to be presented to catalogue users, and with the making of references to those headings. This part contains rules for choice of main and added entries in chapter 21 and form of headings and uniform titles in chapters 22 to 25 and references in chapter 26 of the code. In both parts the rules proceed from the general to the specific.

8.5.1 Part I of AACR2

AACR2 has used general framework for all bibliographic descriptions called ISBD(G). This has enabled it to achieve an effective and consistent approach to bibliographic description. All the rules for description confirm to a single set of punctuation conventions. The same principles have been adopted for description of different kinds of library materials. The rules in part I deal with description of all types of library information materials.

Part I of AACR2 consists of 13 chapters. Chapter 1 provides general rules for description of all materials collected now-a-days by libraries. Chapters 2 to 10 provide rules for describing specific types of materials as given below:

Chapter 2 : Books, pamphlets and printed materials
Chapter 3 : Cartographic materials

Chapter 4 : Manuscripts
Chapter 5 : Music
Chapter 6 : Sound Recordings
Chapter 7 : Motion pictures and video recordings
Chapter 8 : Graphic materials
Chapter 9 : Machine-readable data files
Chapter 10 : Three dimensional artifacts and Realia
Chapter 11 : Microforms
Chapter 12 : Serials
Chapter 13 : Analysis

In chapters 2-10 the areas and elements prescribed by ISBD(G) and their prescribed punctuations have been described in terms of the particular type of library material. Content of an element which is special to a particular type of material has been dealt in detail in the concerned chapter. In case, general rule is applicable reference is made to that of chapter 1. Chapters 11-13 deal with microforms, serial and analysis which are of partial general in nature. In some cases the rules provided in these chapters have modified the provisions given in preceding chapters. In other cases the rules given in these chapters are supposed to be used along with the rules described in preceding chapters. Analysis is conceived as a process by which a part of publication is described is related to the whole document. The rules for preparing analytical entries and multi-level descriptions are provided in chapter 13 which is entitled 'Analysis'.

The rules in each chapter consist of:
0. Preliminary rules
1. Title and statement of responsibility
2. Edition
3. Material Specific details
4. Publication, distribution, etc
5. Physical description
6. Series
7. Note
8. Standard number and terms of availability
9. Supplementary items
10. Items made up of several types of material
11. Facsimiles, photocopies, and other reproductions

8.5.2 Part II of AACR2

Part II deals with Headings, Uniform titles, and References. It consists of 6 chapters numbered 21 to 26. Chapter 21 deals with access points. Chapters 22-24 deal with headings for persons, Geographic names and corporate bodies respectively. Chapter 25 deal with Uniform titles and rules for giving different types of references are dealt in chapter 26. In each chapter general rules precede special rules. In case, no specific rules exist for a given specific problem, then general rules are to be applied.

Rules in chapter 21 provide an integrated approach to all library materials in deciding access points. The concept of mixed responsibility is also dealt in detail in this chapter. Rules for added entries also provided in this chapter.

8.6 APPENDICES

AACR2 code provides four appendices dealing with capitalization, abbreviations, the treatment of numerals and glossary of terms used in cataloguing.

8.7 SUMMARY

Anglo-American Cataloguing Rules2 (AACR2) is widely used catalogue code throughout the world for describing the documents for the purpose of catalogue. It aims to respond to the changes that have been taking place in different aspects of librarianship. The code made an attempt to make the rules as contemporary as could be possible in the circumstances. The entire code is based on ISBD and in accordance with "Paris Principles" on cataloguing rules. An attempt has been made to make the code more accessible to cataloguers and bibliographers in language and articulation. AACR2 has not been able to make all the provisions required for library automation. However, built on foundations established by the Anglo-American Cataloguing Rules, RDA is being developed as a new standard designed for use in a digital environment. RDA will be co-published by the American Library Association, the Canadian Library Association and the Chartered Institute of Library and Information Professionals (CILIP).

8.8 TECHNICAL TERMS

AACR : Anglo American Cataloguing Rules

RDA : Resource Description and Access

CILIP : Charted Institutions of Library and Information Professions

8.9 SUGGESTED READINGS

1. Anglo-American Cataloguing Rules ; 2nd edition. ALA, 1978
2. Kumar, Girija and Krishan Kumar. Theory of Cataloguing. 5th ed. New Delhi: Vikas publishing house, 1988

LESSON - 9

COMMON COMMUNICATION FORMAT (CCF)

AIMS AND OBJECTIVES

The objective of this lesson is to explain the basic features of Common Communication Format (CCF). It explains the basic structure of the format, organization of description of the library materials in the format.

After studying this lesson you will be able to

- What is CCF
- Origin of CCF
- Structure of CCF
- Criticism on CCF

Structure

9.1 INTRODUCTION

CCF (Common Communication Format) is a format for the exchange of bibliographic records. CCF is devised by taking into consideration the major existing International exchange formats. There were different practices in record creation using different formats such as UNIMARC, ISDS, MEKOF-2, ASIDIC, USSR-US CCF etc., resulting in records which, when merged into a database, will show their different origins. To achieve homogeneity and avoid ambiguity a small number of data elements were identified which were used by virtually all information-handling communities, including both libraries and abstracting and indexing organizations. These commonly used data elements formed the core of the CCF. The Common Communication Format (CCF) is to

provide a detailed and structured method for recording a number of mandatory and optional data elements in a computer-readable bibliographic record for exchange purposes between two or more computer-based systems. However, it can also be useful within non-computerized bibliographic systems.

A technique was developed to show relationships between bibliographic records, and between elements within bibliographic records. The concept of the record segment was developed and refined, and a method for designating relationships between records, segments, and fields was accepted. The first edition of CCF: The Common Communication Format was published by UNESCO in 1984.

CCF provides rules to achieve consistency, uniformity and compatibility between more than one computer systems. Within an information system, the record which forms the database will usually exists in a number of separate but highly compatible formats. At the very least there will be:

- a format in which records will be input to the system.
- a format best suited to long term storage.
- a format to facilitate retrieval, and
- a format in which records will be displayed.

In addition, if two or more organizations wish to exchange records with one another, it will be necessary for each of these organizations to agree upon a common standard format for exchange purposes. Each must be able to convert to an exchange format record from an internal-format record, and vice-versa.

9.2 ORIGIN OF CCF

In April 1978 the Unesco/PGI (i.e. Unesco General Information Programme) sponsored an International symposium on Bibliographic exchange formats, which was held in Taormina, Sisily. The symposium was convened to study the desirability and feasibility of establishing the maximum compatibility between existing bibliographic exchange formats.

Immediately after this symposium UNESCO/PGI formed an Adhoc group to establish a Common Communication Format. At the start itself, the Adhoc group has decided certain principles which the CCF still follows like:-

1. The structure of the format to confirm the international standard ISO – 2709.
2. The core record which consists of a small number of mandatory data elements essential for bibliographic description are identified in a standard manner.
3. The mandatory elements are augmented by the well identified optional data elements.

9.3 CCF RECORD STRUCTURE

Each CCF record consists of four major parts, i.e.

1. Record Label (having 24 Characters);
2. Directory (having Variable Length);
3. Data fields (having Variable Length);
4. Record Separator (having 1 Character).

9.3.1 Record Label

Each CCF record begins with a fixed label of 24 characters. The contents of the Label are as follows:-

Character Positions	**Assigned No.of Characters**	**Contents**
0-4th	5 Chrs	Record Length- The length of the record includes the label, directory, data fields and record separator. (Use of 5 characters for the record length permits records as long as 99,999 characters)
5th	1 Chr.	Record Status- Using a code taken from the list of Record Status Codes from CCF manual (Pg. No.143)
6th	1 Chr.	Blank
7th	1 Chr.	Bibliographic Level- Codes are given in CCF manual on Page No. 144.(Character position for bibliographical level is not used in factual record & is filled with space).
8-9th	2 Chrs	Blank
10th	1 Chr.	Indicator Length- Used to fix the length of the indicator.
11th	1 Chr.	Subfield Identifier Length- eg.,^a,^b,^c,^d,…etc.
12-16th	5 Chrs	Base Address of Data- the location within the record at which the first data field begins.
17-19th	3 Chrs	Blank
20th	1 Chr.	Length of the 'Length of data field' in each directory entry- use of 4 characters here permits data field as long as 9,999 characters.
21st	1 Chr.	Length of 'Starting Character position in each directory entry- normally we use 5 because for getting 10,000th (9,999+1th) position we need 5 characters only.

22nd	1 Chr.	Length of 'Implementation defined' section of each directory entry- Normally **unused** (Blank i.e., Zero)
23rd place	1 Chr.	Blank.

9.3.2 Directory

A single Directory Entry Contains:

24-26th	3 Chrs	Tag- By using 3 Characters we can use 999 possible tags.
27-30th	4 Chrs	Length of data field- A four digit number showing how many characters are occupied by the data field, including indicators & data field separator but excluding the record separator code if the data field is the last field in the record.
31-35th	5 Chrs	Starting Character Position- It gives the position of the first character of the next data field relative to the base address of the data.
36th	1 Chr.	Segment Identifier- Choosen from 0-9 &/or A-Z to designate the data field as being a member of particular segment**(Normally not used).**
37th	1 Chr.	Occurrence Identifier- Choosen from 0-9 &/or A-Z, which differentiate multiple occurrences of data fields that carry the same tag within same record **(Normally not used)**.

9.3.3 Data Fields

A single Data Field Contains:

38-39th	2 Chrs	For Indiators
40-41st	2 Chrs	For Subfield Identifiers
Variable	Variable	Subfield
	1 Chr	Field Separator

9.3.4 Record Separator

Each record is separated by one character record separator

At End	1 Chr.	Record Separator.

9.3.5 Sample CCF Record

Sample CCF Record Structure is as follows:

00161000002200072000450 □ Record Label200004000000 □ Tag(200), Length of Data field (0040), Starting Character Position of data(00000).300002000040 Tag(300), Length of Data field (0020), Starting Character Position of Data(00040).400001600060 □ Tag(400), Length of Data field (0016), Starting Character Position of data(00060).440001100076 □ Tag(440), Length of Data field (0011), Starting Character Position of Data(00076).# □ Field Separator
^aProlegomena to Library Classification#^aRanganathan^bS.R.#^aDelhi ^bJaypee#^a19990000# □ Field Separator# □ Record Separator.

9.4 CRITICISM ON CCF

Many of the disadvantages of CCF are based on the disadvantages of the Cataloguing Codes. The CCF is basically a tag code to facilitate data exchange between two or more systems. It should be independent of the cataloguing codes.

The field Physical description (460) which describes the physical attributes of the item cannot be used to produce catalogue cards following different codes. For example: AACR insists that if there are papers numbered in roman numerals they should be taken into consideration. However, CCC does not insist on this. The problem is that the catalogue codes not only prescribe what should be descriptive element, they have formulated rules on how they should be represented.

While exchanging the data the field Place and Name of the Distributor (420) is of no use to the other system, because the distributor can vary from place to place. It can be argued that this field is provided for internal use. But there is no field to provide accession number which is frequently used for internal purposes.

The subfield "B" which describes statement of responsibility in the field Title and Statement of responsibility (200) and the field Name of Person (300) is an overlapping concept to each other.

Field Segment linking fields (080,081,082,083,084,085) does not give any clue about the linking of documents or records. Only Field to Field linking (086) makes some sense.

The Field Source of Record (020) is a non-repeatable one. But if the database is merged with a master database, then it may be repeated disputing the concept of non-repeatability.

The major disadvantage in CCF is the different codes used for data elements. In this case CCF seems to be consistently inconsistent.

For example: for the subfield 'A' which describes language of the record in Field Language and Script of the record (031) they have followed a code list from ISDS manual. The codes are in alphabets. The same problem surfaces in the fields with tag numbers 040, 041, 200,201, 210, 220, 221, etc., where ever the question of language arises. The same type of problem is there in Record Status Codes, bibliographic data level codes, completeness of record codes. But these problems does not appear in character set codes like, physical medium codes, role codes, type of material codes, by using numerals instead of alphabets. Here the consistency fails.

Again no Code is given for subfield 'B' of tag 110 which describes national bibliographic agency code in the field National Bibliography Number (110) and sub field 'B' which describes legal deposit agency code in field Legal Deposit Number (111).

9.5 CCF TAGS

Following are the Common Communication Format (CCF) Tag Numbers

Tag	Name
001	**Record Identifier**
010	Record identifier used in secondary segments
010A	Identifier
011	Alternative record identifier (R)
011A	Alternative identifier
011B	Identification of Agency in coded form
011C	Name of agency
015	Bibliographic level of secondary segment
015A	Bibliographic level
020	**Source of record**
020A	Identification of agency in coded form
020B	Name of agency
020C	Name of code set
020D	Rules for bibliographic description
020L	Language of name of agency
021	Completeness of record
021A	Level of completeness code
022	Date entered on file
022A	Date
023	Date and number of record version
023A	Version date
023B	Version number
030	**Character sets used in record**

030A	Alternative Control set (C1)
030B	Default Graphic set (G0)
030C	Second Graphic Set (G1)
030D	Third Graphic Set (G2)
030E	Fourth Graphic Set (G3)
030F	Additional Control Set (R)
030G	Additional Graphic Set (R)
031	Language and script of record (R)
031A	Language of the record(R)
031B	Script of the record(R)
040	**Language and script of item(R)**
040A	Language of item (R)
040B	Script of item
041	Language and Script of Summary (R)
041A	Language of the summary (R)
041B	Script of the summary
050	**Physical medium**
050A	Physical medium code (R)
060	**Type of material**
060A	Type of material code (R)
080	**Segment linking field: general vertical relationship (R)**
080A	Segment relationship code
080B	Segment indicator code
081	segment linking field: vertical relationship from monograph
081A	Segment relationship code
081B	Segment indicator code
082	Segment linking field: vertical relationship from multi-volume monographic
082A	Segment relationship code
082B	Segment indicator code
083	Segment linking field: vertical relationship from serial
083A	Segment relationship code
083B	Segment indicator code
085	Segment linking field: horizontal(R)
085A	Segment relationship code
085B	Segment indicator code
086	Field to field linking(R)
086A	field linked from
086B	Field relationship code
086C	Field linked to
100	I**nternational standard book number**(R)
100A	ISBN
100B	Invalid ISBN (R)
100C	Qualification (R)
101	International standard serial number (issn)

101A	ISSN
101B	Invalid ISSN
101C	Cancelled ISSN (R)
102	Coden
102A	Coden
110	N**ational bibliographic number (R)**
110A	National Bibliographic Number
110B	National Bibliographic Agency Code
111	Legal deposit number(R)
111A	Legal deposit number
111B	Legal deposit agency
120	**Document identification number(R)**
120A	Document identification number
120B	Type of number
200	**Title and associated statement(s) of responsibility (R)**
200A	Title (R)
200B	Statement of responsibility associated with title (R)
200L	Language of title
200S	Script of title
201	Key title
201A	Key title
201B	Abbreviated key title
201L	Language of key title
201S	Script of key title
210	Parallel title and associated statement(s) of responsibility (r)
210A	Parallel title
210B	Statement of responsibility associated with parallel title (R)
210L	Language of parallel title
210S	Script of parallel title
220	Spine title (R)
220A	Spine title
220L	Language of spine title
221	Cover title (R)
221A	Cover title
221L	Language of cover title
222	Added title page title (R)
222A	Added title page title
222L	Language of added title page title
223	Running title (R)
223A	Running Title
223L	Language of running title
230	Other title (R)
230A	Other variant title
230L	Language of title
240	Uniform title (R)
240A	Uniform title
240B	Number of part(s) (R)

240C Name of part(s) (R)
240D Form subheading (R)
240E Language of item (as part of uniform title) (R)
240F version
240G Date of version
240L Language of uniform title
240Z Authority number
260 Edition statement and associated statement(s) of responsibility (r)
260A Edition Statement
260B Statement of responsibility associated with edition (R)
260L Language of edition statement

300 **Name of person (R)**
300A Entry element
300B Other name elements
300C Additional elements to name
300D Date(s)
300E Role (coded) (R)
300F Role (non-coded) (R)
300Z Authority number

310 **Name of corporate body (R)**
310A Entry element
310B Other part(s) of name (R)
310C Qualifier (R)
310D Address of corporate body
310E Country of corporate body
310F Role (coded) (R)
310G Role (non-coded) (R)
310L Language of entry element
310S Script of entry element
310Z Authority number

320 **Name of meeting (R)**
320A Entry element
320B Other part(s) of name (R)
320C Qualifier (R)
320E Country
320G Location of meeting
320H Date of meeting (in ISO format)
320I Date of meeting (in free format)
320J Number of meeting
320L Language of entry element
320S Script of entry element
320Z Authority number
330 Affiliation (R)
330A Entry element
330B Other parts of the name (R)
330C Qualifier (R)
330D Address (R)

330E	Country of affiliation
330L	Language of entry element
400	**Place of publication and publisher (R)**
400A	Place of publication (R)
400B	Name of publisher
400C	Full address of publisher (R)
400D	Country of publisher (R)
410	Place of manufacture and name of manufacturer (R)
410A	Place of manufacture (R)
410B	Name of manufacturer
410C	Full address of manufacturer (R)
410D	Country of manufacturer (R)
420	Place and name of distributor (R)
420A	Place of distributor (R)
420B	Name of distributor
420C	Full address of distributor (R)
420D	Country of distributor (R)
440	**Date of publication (R)**
440A	date in formalized form
440B	date in non-formalized form
441	Date of legal deposit
441A	Date legal deposit
450	Serial numbering
450A	Serial numbering and date
460	**Physical description**
460A	Number of pieces and designation
460B	Other descriptive details
460C	Dimensions
460D	Accompanying material (R)
480	**Series statement and associated statement(s) of responsibility (R)**
480A	Series Statement
480B	Statement of responsibility associated with series statement
480C	Part statement
480D	ISSN
480L	Language of title
480S	Script of title
490	**Part statement (R)**
490A	Volume / part numeration and designation (R)
490B	Pagination defining a part
490C	Other identifying data defining a part
500	**Note (R)**
500A	Note

510	Note on bibliographical relationship (R)
510A	Note
520	Serial frequency note (R)
520A	Frequency
520B	Dates of frequency
530	Contents note (R)
530A	Note
600	**Abstract (R)**
600A	Abstract
600L	Language of abstract
610	C**lassification scheme notation**
610A	Notation (R)
610B	Identification of classification scheme
620	S**ubject descriptor (R)**
620A	Subject descriptor
620B	Identification of subject system

9.6 SUMMARY

CCF is devised by taking into consideration the major existing International exchange formats and was intended to be used for the transfer of records between systems. This is purely an exchange format. It does not give any information about circulation system of any particular library. In spite of its disadvantages it is one of the most widely used formats especially in developing countries. But still some serious, fruitful steps should be taken to overcome its problems.

9.7 TECHNICAL TERMS

CCF: Common Communication Format
UNIMARC: Universal Machine Readable Catalogue

9.8 SUGGESTED READINGS

1. CCF: the Common Communication Format, 2nd Ed., Paris, UNESCO, 1988(PGI-88/WS/2).
2. Ellen, Gradley and Hopkinson, Alan, Exchanging Biliographic data: MARC and other international format. Library Association Publishing Ltd., London, 1990, pp. 209-222.
3. International Standard ISO 2709(E): Documentation- Format for Bibliographic Information Interchange on Magnetic Tape.(In Handbook on International Standards Governing Information Transfer by International Organization for Standards, 1977. pp. 291-294.

LESSON - 10

MACHINE READABLE CATALOGUE 21 (MARC21)

AIMS AND OBJECTIVES

The objective of this lesson is to explain the basic features of Machine Readable Cataloguing (MARC). It explains the evolution of MARC. The basic structure of the MARC Record is clearly dealt with in this lesson. The Field designators and subfield codes and tags for the variable fields are explained. MARCXML version which is developed for web applications is also explained.

After studying this lesson you will be able to know
- What is MARC
- History of MARC
- Structure of MARC
- Field designators and tags

Structure

10.1 INTRODUCTION

MARC (MAchine-readable Cataloging) standards are a set of digital formats for the description of items, such as books, patents, serials etc. catalogued by libraries. It was developed by Henriette Avram at the US Library of Congress during the 1960s to create records that can be used by computers, and to share those records among libraries. By 1971, MARC formats had become the national standard for dissemination of bibliographic data in the United States. MARC standard became the international standard by 1973. There are several versions of MARC in use around the world, the most predominant being MARC 21, created in 1999. The MARC 21 family of standards now includes formats for authority records, holdings records, classification schedules, and community information, in addition to the format for bibliographic records.

10.2 Record Structure

MARC records are typically stored and transmitted as binary files. MARC uses the ISO 2709 standard to define the structure of each record. This includes a marker to indicate where each record begins and ends, as well as a set of characters at the beginning of each record that provide a directory for locating the fields and subfields within the record.

The basic machine readable catalogue record on a MARC tape consists of the Leader, the Record Directory, the Control Fields and the Variable Fields.

Leader	Record Directory	Control Fields	Variable Fields

The control field consists of both variable control number and Variable Fixed Fields. The Leader is fixed in length for all records contain 24 characters. It is a set of fields describing the general structure of the individual entry. The Record Directory is an index to the location of the Control and Variable Fields in the record. It consists of a series of fixed length entries, one for each variable field in the record.

An entry in the Record Directory contains the identification tag, the length and starting character position of each variable field in the record. The record Directory will end with a field-terminator code. Since the number of variable fields in a record can vary, the total length of the Record Directory is also variable. All fields end with field-terminator code except the last variable field which ends with record-terminator code. All Variable Fields are made up of variable length alphanumeric data. Each variable field is identified by three character numeric tag in the record directory. Tags may be repeated as required in a logical record. However, tags associated with control fields will not be repeated in a logical record.

10.2.1 Outline of Leader

The total number of characters in the Leader is 24, and there are nine data elements in the Leader as described below:

Name of Data Element	No. of Characters
Record Length	5
Record Status	1
Type of Record	1
Bibliographic Level	1
Blanks	2
Indicator Count	1
Subfield Count	1
Base Address of Data	5
Blank Character	7

10.2.2 Outline of the Record Directory

Each record directory consist of three elements viz, Tag, Length of the field and Starting character position of the field in the record. Each entry in the directory is 12 character entries as detailed below. The number of entries in the record directory corresponds to the number of variable fields present in the record. The record directory is terminated by field-terminator code.

The sample record directory is given below:

Entry	Element in the entry	No. Characters
Field 1	Tag of the field	3
	Length of the Field	4
	Starting Character Position of the field	5
Field 2	Tag of the field	3
	Length of the Field	4
	Starting Character Position of the field	5

Field n	Tag of the field	3
	Length of the Field	4
	Starting Character Position of the field	5
	Field Terminator (F/T)	1

10.2.3 Outline of Control Fields

Data Element 1	Data Element 2	Data Element 3	----	Data Element n	F/T

Example:

Tag	Name of the Control Field	Data Elements	No. of Characters	Character position in the field
001	Library of Congress Card Number	Year Number supplement **etc**.	2 6 1	3-4 5-10 11
008	Fixed length Data Elements	1 date entered on file 2 Type of publication 3 Date 1 . 10 Govt. pub. Indicator 16 Biography code 17 Language code	6 1 4 1 1 3	0-5 6 7-10 28 34 35-37

10.2.4 Outline of Variable Fields

Variable field consists of indicators, subfield codes, data elements and the field terminator. Further each variable field is assigned a tag and the tag is stored in the directory. The directory, control fields and variable fields are always terminated by a field terminator. Finally the last character in the record is a Record Terminator.

Indicator
Subfield code
Data Element
Subfield code
Data Element
. . .
Subfield code
Data Element
F/T
Indicator
Subfield code
Data Element
. . .
Subfield code
Data Element
F/T
.

. .
F/T
R/T

10.3 FIELD DESIGNATIONS

Indicators: Each variable field will begin with 2 character code which provides descriptive information about the field. The contents of the indicators are specified for the fields in which they are used, If the indicators are not used with a particular field, they will contain blanks.

Subfield codes: Variable fields are made up of a single data element or a group of data elements. The subfield code precedes each data element in a field and identifies the dta element. The subfield code consists of 2 characters. For the purpose of these specifications, the delimiter will be represented by "$".

Data Elements: All the data elements in the variable fields may have variable lengths. Each variable field or data element has a tag. Some fields are repeatable.

Subfield Codes in Variable Fields: The subfield code identifies the constituent data elements of a variable field. For example the imprint field, tag 260, may have the following 3 data elements in its respective subfield codes:

Place$a
Publisher$b
Date$c

Variable Fields: Each data element in the variable field has tag of three characters. Following are few fields with tags for the purpose of illustration.

010	LC Card number
100	Main entry Personal name
245	Title
300	Collation
650	Topical subject heading
700	Personal name added entry

10.4 CONTENT IN A MARC RECORD

MARC transmits information about a bibliographic item, not the content of that item; this means it is a metadata transmission standard, not a content standard. The actual content a cataloger will place in each MARC field is usually governed and defined by standards outside of MARC, except for a handful of fixed fields defined by the MARC standards themselves. The Anglo-American Cataloguing Rules, for example, define how the physical characteristics of books and other item should be expressed. The Library of Congress Subject Headings (LCSH) provides a list of authorized subject terms to describe the main content of the item. Other cataloging rules, subject thesauri, and classification schedules can also be used.

10.5 MARC FORMATS

MARC formats

Name	Description
Authority records	Provide information about individual names, subjects, and uniform titles. An authority record establishes an authorized form of each heading, with references as appropriate from other forms of the heading.
Bibliographic records	Describe the intellectual and physical characteristics of bibliographic resources (books, sound recordings, video recordings, and so forth).
Classification records	MARC records containing classification data. For example, the Library of Congress Classification has been encoded using the MARC 21 Classification format.
Community Information records	MARC records describing a service providing agency. For example, the local homeless shelter or tax assistance provider.
Holdings records	Provide copy-specific information on a library resource (call number, shelf location, volumes held, and so forth).

10.6 MARC 21

MARC 21 was designed to redefine the original MARC record format for the 21st century and to make it more accessible to the international community. MARC 21 has formats for the following five types of data: Bibliographic Format, Authority Format, Holdings Format, Community Format, and Classification Data Format. Currently MARC 21 has been implemented successfully by The British Library, the European Institutions and the major library institutions in the United States, and Canada.

MARC 21 is a result of the combination of the United States and Canadian MARC formats (USMARC and CAN/MARC). MARC21 is based on the ANSI standard Z39.2, which allows users of different software products to communicate with each other and to exchange data. MARC 21 in UTF-8 format allows all the languages supported by Unicode.

10.7 MARCXML

In 2002, the Library of Congress developed the MARC XML schema as an alternative record structure, allowing MARC records to be represented in XML. Libraries typically expose their records as MARC XML via a web service, often following the SRU or OAI-PMH standards.

MARC XML is an XML schema base on the common MARC 21 standards. MARC XML was developed by the Library of Congress and adopted by it and others as a means of facilitating the sharing of, and networked access to, bibliographic information.

10.7.1 MARCXML primary design goals included:

- Simplicity of the schema
- Flexibility and extensibility
- Lossless and reversible conversion from MARC
- Data presentation through XML style sheets
- MARC records updates and data conversions through XML transformations
- Existence of validation tools

10.8 FUTURE

The future of the MARC formats is a matter of some debate among libraries. On the one hand, the storage formats are quite complex and are based on outdated technology. On the other, there is no alternative bibliographic format with an equivalent degree of granularity. The billions of MARC records in tens of thousands of individual libraries (including over 50,000,000 belonging to the OCLC consortium alone) create inertia. The Library of Congress has launched the Bibliographic Framework Initiative (BIBFRAME), that aims at providing a replacement for MARC that provides greater granularity and easier re-use of the data expressed in multiple catalogs.

10.9 SUMMARY

MAchine Readable Catalogue (MARC) is the product of Library of Congress. It is developed mainly to supply Cataloguing in Publication Data (CIP) in tapes. It is also used as standard format for bibliographic data exchange in computer readable form. But, most of the countries have developed their own MARC Formats. For example UK MARC, Canadian MARC and Indian MARC are in existence for representing bibliographic data in computer readable form. MARC 21 is the latest format in use.

10.10 TECHNICAL TERMS

CIP : Cataloguing in Publication Data

10.11 SUGGESTED READINGS

1. Aswal, Rajendra Singh. MARC 21 Cataloguing Format for 21st Century New Delhi, Ess Ess Publications, 2004
2. Bryne, Deborah J. MARC manual : understanding and using MARC records. Englewood, Colo. : Libraries unlimited, 1991
3. Das, Subarna K.Fundamentals of MARC 21 Bibliographical Format New Delhi : Ess Ess Publications, 2009
4. Fritz, Deborah A. Cataloging with AACR2 and MARC21.New York : ALA, 2007
5. Fritz, Deborah A. and Richard J. Fritz. MARC 21 for Everyone: A Practical Guide . New York : ALA, 2003
6. The Library of Congress US MARC home page at:http://lcweb.loc.gov/marc/
7. Website : http://www.marc21.ca/index-e.html
8. Website : http://en.wikipedia.org/wiki/MARC_standards

LESSON - 11
I S O 2709

AIMS AND OBJECTIVES

The objective of this lesson is to explain computer readable bibliographic data exchange format.

After studying this lesson you will be able to understand

- ISO 2709 - the data exchange format?
- history of the ISO 2709 format
- structure of the ISO 2709

Structure

11.1 INTRODUCTION

ISO 2709 is a standard for Record Structure of machine readable bibliographic record. It is an ISO standard for bibliographic descriptions, titled "Information and documentation—Format for information exchange". It is maintained by the Technical Committee for Information and Documentation

11.2 HISTORY

ISO 2709 is a format for the exchange of bibliographic information. It was developed in the 1960s under the direction of Henriette Avram of the Library of Congress to encode the information printed on library cards. It was first created as ANSI Standard Z39.2, and called Information Interchange Format. The 1981 version of the standard was titled Documentation Format for bibliographic information interchange on magnetic tape. The latest edition of that standard is Z39.2 published in 1994 (ISSN: 1041-5653). The ISO standard supersedes Z39.2. The current standard is ISO 2709 released in December 2008.

11.3 BASIC STRUCTURE

ISO 2709 record has three sections:

11.3.1. Record label: The first 24 characters of the record. This is the only portion of the record that is fixed in length. The record label includes the record length and the base address of the data contained in the record. It also has data elements that indicate how many characters are used for indicators and subfield identifiers.

Name of Data Element	No. of Characters
Record Length	5
Record Status	1
Type of Record	1
Bibliographic Level	1
Blanks	2
Indicator Count	1
Subfield Count	1
Base Address of Data	5
Blank Character	7

11.3.2 Directory: The directory provides the entry positions to the fields in the record, along with the field tags. A directory entry has four parts and cannot exceed nine characters in length:

- Field tag (3 characters)
- Length of the field (4 characters)
- Starting character position of the field (5 characters)
- (Optional) Implementation-defined part

Entry	Element in the entry	No. Characters
Field 1	Tag of the field	3
	Length of the Field	4
	Starting Character Position of the field	5
	Field Terminator (F/T)	1
Field 2	Tag of the field	3
	Length of the Field	4
	Starting Character Position of the field	5

Field n	Tag of the field	3
	Length of the Field	4
	Starting Character Position of the field	5
	Field Terminator (F/T)	1
	Record Terminator (R/T)	1

11.3.3 Data fields (Variable fields): A string containing all field and subfield data in the record

Indicator
Subfield code
Data Element

Subfield code
Data Element
. . .
Subfield code
Data Element
F/T
Indicator
Subfield code
Data Element
. . .
Subfield code
Data Element
F/T
. . .
F/T
R/T

11.3.4 Record separator: A single character.

The tags are often displayed as labels on bibliographic fields. Each bibliographic field has an associated tag. The tags are stored in the directory not in the bibliographic field.

12 Fields

There are three kinds of fields in the ISO 2709 record:

- **Record identifier field**: Identifying the record and assigned by the organization that creates the record. The record identifier field has tag 001.
- **Reserved fields**: Reserved fields supply data which may be required for the processing of the record. Reserved fields always have a tag in the range 002–009 and 00A–ZZZ.
- **Bibliographic Fields:** These are in the range 010–999 and 0AA–ZZZ. The bibliographic fields contain data and a field separator. They can also have these optional sub-parts:
    - **Indicator** (0–9 characters, as coded in the Leader): Indicators generally provide further information about the contents of the field, the relationship between the field and other fields in the record, or about action required in certain data manipulation processes (including display labels).
    - **Identifier** (0–9 characters): This identifies data within the bibliographic field. Where used, identifiers are composed of a delimiter (1 char, IS_1 of ISO 646) and an identifying code (1–9 chars, as defined in the leader), plus a variable length string containing the data.

Example

MARC21 library cataloguing record coded in ISO 2709 format. MARC21 is an instance of ISO 2709 that has the following characteristics:

- tags are in the range of 002–999 only
- there is a two-character indicator on each field, and each character is a separately defined data element
- the identifier within data fields (called "subfield code" in MARC21) is a single ASCII character preceded by IS1 of ISO 646.

11.5 SUMMARY

ISO 2709 specifies the requirements for a generalized exchange format which will hold records describing all forms of material capable of bibliographic description as well as other types of records. It does not define the length or the content of individual records and does not assign any meaning to tags, indicators or identifiers, these specifications being the functions of an implementation format. It describes a generalized structure, a framework designed especially for communications between data processing systems and not for use as a processing format within systems.

11.6 SUGGESTED READINGS

1. Gredley, Ellen and Hopkinson, Alan. Exchanging bibliographic data: MARC and other International formats. London : Library Association, 1990
2. Hopkinson, Alan (ed). Bibliographic format. Munchen : Saur, 2007
3. ISO 2709: Information and documentation -- Format for information exchange. International Organization for Standardization, 1996
4. "ISO 2709:2008 - Information and documentation -- Format for information exchange". Retrieved 21 January 2011.
5. "ISO 2709:1981 - Documentation -- Format for bibliographic information inter change on magnetic tape". Retrieved 21 January 2011.

LESSON 12
METADATA

AIMS AND OBJECTIVES

The objective of this lesson is to explain the basic features of Anglo-American Cataloguing Rules 2. It explains the basic structure of the code, organization of description of the library materials and levels of description etc.

After studying this lesson you will be able to

- What is AACR2
- Governance of AACR
- Organization of description of documents
- Levels of description recommended by AACR2
- Structure of AACR2

Structure

12.1 INTRODUCTION

The term metadata refers to "data about data". The term is ambiguous, as it is used for two fundamentally different concepts (types). Structural metadata is about the design and specification of data structures and is more properly called "data about the containers of data; descriptive metadata, on the other hand, is about individual instances of application data, the data content. Metadata (Meta content) are traditionally found in the card catalogues of libraries. As information has become increasingly digitalized, metadata are also used to describe digital data using metadata standards specific to a particular discipline. By describing the contents and context of data files, the quality of the original data/files is greatly increased. For example, a webpage may include metadata specifying what language it is written in, what tools were used to create it, and where to go for more on the subject, allowing browsers to automatically improve the experience of users.

12.2 DEFINITION OF METADATA

Metadata (Meta content) are defined as the data providing information about one or more aspects of the data, such as:

- Means of creation of the data
- Purpose of the data
- Time and date of creation
- Creator or author of the data
- Location on a computer network where the data were created
- Standards used

For example, a digital image may include metadata that describe how large the picture is, the colour depth, the image resolution, when the image was created, and other data. A text document's metadata may contain information about how long the document is, who the author is, when the document was written, and a short summary of the document.

Metadata are data. As such, metadata can be stored and managed in a database, often called a Metadata registry or Metadata repository. However, without context and a point of reference, it might be impossible to identify metadata just by looking at them. For example: by itself, a database containing several numbers, all 13 digits long could be the results of calculations or a list of numbers to plug into an equation - without any other context, the numbers themselves can be perceived as the data. But if given the context that this database is a log of a book collection, those 13-digit numbers may now be identified as ISBNs - information that refers to the book, but is not itself the information within the book.

The term "metadata" was coined in 1968 by Philip Bagley, in his book "Extension of programming language concepts", where it is clear that he uses the term in the ISO 11179 "traditional" sense, which is "structural metadata" i.e. "data about the containers of data"; rather than the alternate sense "content about individual instances of data content" or meta content, the type of data usually found in library catalogues. Since then the fields of information management, information science, information technology, and librarianship have widely adopted the term. In these fields the word *metadata* is defined as "data about data". While this is the generally accepted definition, various disciplines have adopted their own more specific explanation and uses of the term.

12.3 METADATA APPLICATION

12.3.1. Libraries

Metadata have been used in various forms as a means of cataloguing archived information. The Dewey Decimal System employed by libraries for the classification of library materials is an early example of metadata usage. Library catalogues used 3x5 inch cards to display a book's title, author, subject matter, and a brief plot synopsis along with an abbreviated alpha-numeric identification system which indicated the physical location of the book within the library's shelves. Such data help classify, aggregate, identify, and locate a particular book.

12.3.2. Photographs

Metadata may be written into a digital photo file that will identify who owns it, copyright & contact information, what camera created the file, along with exposure information and descriptive information such as keywords about the photo, making the file searchable on the computer and/or the Internet. Some metadata are written by the camera and some is input by the photographer and/or software after downloading to a computer. However, not all digital cameras enable you to edit metadata. This functionality has been available on most Nikon DSLRs since the Nikon D3 and on most new Canon cameras since the Canon EOS 7D.

Photographic Metadata Standards are governed by organizations that develop the following standards. They include, but are not limited to:

- IPTC Information Interchange Model IIM (International Press Telecommunications Council),
- IPTC Core Schema for XMP
- XMP – Extensible Metadata Platform (an ISO standard)
- Exif – Exchangeable image file format, Maintained by CIPA (Camera & Imaging Products Association) and published by JEITA (Japan Electronics and Information Technology Industries Association)
- Dublin Core (Dublin Core Metadata Initiative – DCMI)
- PLUS (Picture Licensing Universal System).

12.3.3. Video

Metadata are particularly useful in video, where information about its contents (such as transcripts of conversations and text descriptions of its scenes) are not directly understandable by a computer, but where efficient search is desirable.

12.3.4. Web pages

Web pages often include metadata in the form of Meta tags. Description and keywords Meta tags are commonly used to describe the Web page's content. Most search engines use these data when adding pages to their search index.

12.4 CREATION OF METADATA

Metadata can be created either by automated information processing or by manual work. Elementary metadata captured by computers can include information about when an object was created, who created it, when it was last updated, file size and file extension.

An "object" refers to any of the following:

- A physical item such as a book, CD, DVD, map, chair, table, and flower pot, etc.
- An electronic file such as a digital image, digital photo, document, program file, database table, etc.

12.5 METADATA TYPES

The metadata application is many fold covering a large variety of fields of application there are nothing but specialized and well accepted models to specify types of metadata. Bretheron & Singley (1994) distinguish between two distinct classes: structural/control metadata and guide metadata.

Structural metadata are used to describe the structure of computer systems such as tables, columns and indexes.
Guide metadata are used to help humans find specific items and are usually expressed as a set of keywords in a natural language.

According to Ralph Kimball metadata can be divided into 2 similar categories: **Technical metadata** and **Business metadata**. Technical metadata correspond to internal metadata, Business metadata correspond to external metadata. Kimball adds a third category named **Process metadata.**

On the other hand, NISO distinguishes three types of metadata: descriptive, structural and administrative. **Descriptive metadata** are the information used to search and locate an object such as title, author, subjects, keywords, publisher; S**tructural metadata** give a description of how the components of the object are organized; and A**dministrative metadata** refer to the technical information including file type. NISO also identifies two sub-types of administrative metadata. They are rights management metadata and preservation metadata.

12.6 METADATA STRUCTURES

Metadata (Meta content), or more correctly, the vocabularies used to assemble metadata (Meta content) statements, are typically structured according to a standardized concept using a well-defined metadata scheme, including: metadata standards and metadata models. Tools such as controlled vocabularies, taxonomies, thesauri, data dictionaries and metadata registries can be used to apply further standardization to the metadata. Structural metadata commonality is also of paramount importance in data model development and in database design.

12.6.1 Metadata syntax

Metadata (Meta content) syntax refers to the rules created to structure the fields or elements of metadata (Meta content). A single metadata scheme may be expressed in a number of different mark ups or programming languages, each of which requires a different syntax. For example, Dublin Core may be expressed in plain text, HTML, XML and RDF.

A common example of (guide) Meta content is the bibliographic classification, the subject, the Dewey Decimal class number. There is always an implied statement in any "classification" of some object. To classify an object as, for example, Dewey class number 514 (Topology) (i.e. books having the number 514 on their spine) the implied statement is:

"<book><subject heading><514>. This is a subject-predicate-object triple, or more importantly, a class-attribute-value triple. The first two elements of the triple (class, attribute) are pieces of some structural metadata having a defined semantic. The third element is a value, preferably from some controlled vocabulary, some reference (master) data. The combination of the metadata and master data elements results in a statement which is a Meta content statement i.e. "Meta content = metadata + master data". All these elements can be thought of as "vocabulary". Both metadata and master data are vocabularies which can be assembled into Meta content statements. There are many sources of these vocabularies, both Meta and master data: UML, EDIFACT, XSD, Dewey/UDC/LOC, SKOS, ISO-25964, Pantone, Linnaean Binomial Nomenclature etc. Using controlled vocabularies for the components of Meta content statements, whether for indexing or finding, is endorsed by ISO-25964: "If both the indexer and the searcher are guided to choose the same term for the same concept, then relevant documents will be retrieved."

12.6.2 Hierarchical, linear and planar schema

Metadata schema can be hierarchical in nature where relationships exist between metadata elements and elements are nested so that parent-child relationships exist between the elements. An example of a hierarchical metadata schema is the IEEE LOM schema where metadata elements may belong to a parent metadata element. Metadata schema can also be one-dimensional, or linear, where each element is completely discrete from other elements and classified according to one dimension only. An example of a linear metadata schema is Dublin Core schema which is one dimensional. Metadata schema is often two dimensional, or planar, where each element is completely discrete from other elements but classified according to two orthogonal dimensions.

12.6.3 Metadata hyper mapping

In all cases where the metadata schema exceeds the planar depiction, some type of hyper mapping is required to enable display and view of metadata according to chosen aspect and to serve special views. Hyper mapping frequently applies to layering of geographical and geological information overlays.

12.6.4 Granularity

The degree to which the data or metadata are structured is referred to as their granularity. Metadata with a high granularity allow for deeper structured information and enable greater levels of technical manipulation however, a lower level of granularity means that metadata can be created for considerably lower costs but will not provide as detailed information. The major impact of granularity is not only on creation and capture, but moreover on maintenance. As soon as the metadata structures get outdated, the access to the referred data will get outdated. Hence granularity shall take into account the effort to create as well as the effort to maintain.

12.7 METADATA STANDARDS

International standards apply to metadata. Much work is being accomplished in the national and international standards communities, especially ANSI (American National Standards Institute) and ISO (International Organization for Standardization) to reach consensus on standardizing metadata and registries.

The core standard is ISO/IEC 11179-1:2004. All yet published registrations according to this standard cover just the definition of metadata and do not serve the structuring of metadata storage or retrieval neither any administrative standardization. It is important to note that this standard refers to metadata as the data about containers of the data and not to metadata (Meta content) as the data about the data contents. It should also be noted that this standard describes itself originally as a "data element" registry, describing disembodied data elements, and explicitly disavows the capability of containing complex structures. Thus the original term "data element" is more applicable than the later applied buzzword "metadata".

The Dublin Core metadata terms are a set of vocabulary terms which can be used to describe resources for the purposes of discovery. The original set of 15 classic metadata terms, known as the 'Dublin Core Metadata Element Set' endorsed in the following standards documents:

- IETF RFC 5013
- ISO Standard 15836-2009
- NISO Standard Z39.85

12.7.1 Library and information science

Libraries employ metadata in library catalogues, most commonly as part of an Integrated Library Management System. Metadata are obtained by cataloguing resources such as books, periodicals, DVDs, web pages or digital images. These data are stored in the integrated library management system, ILMS, using the MARC metadata standard. The purpose is to direct patrons to the physical or electronic location of items or areas they seek as well as to provide a description of the item/s in question.

More recent and specialized instances of library metadata include the establishment of digital libraries including e-print repositories and digital image libraries. While often based on library principles, the focus on non-librarian use, especially in providing metadata, means they do not follow traditional or common cataloguing approaches. Given the custom nature of included materials, metadata fields are often specially created e.g. taxonomic classification fields, location fields, keywords or copyright statement. Standard file information such as file size and format are usually automatically included.

Standardization for library operation has been a key topic in international standardization (ISO) for decades. Standards for metadata in digital libraries include Dublin Core, METS, MODS, DDI, ISO standard Digital Object Identifier (DOI), ISO standard Uniform Resource Name (URN), PREMIS schema, Ecological Metadata Language, and OAI-PMH. Leading libraries in the world give hints on their metadata standards strategies.

Kimball et al. refers to three main categories of metadata: Technical metadata, business metadata and process metadata. Technical metadata are primarily definitional, while business metadata and process metadata are primarily descriptive. Keep in mind that the categories sometimes overlap.

- **Technical metadata** define the objects and processes in a DW/BI system, as seen from a technical point of view. The technical metadata include the system metadata which define the data structures such as: tables, fields, data types, indexes and partitions in the

relational engine, and databases, dimensions, measures, and data mining models. Technical metadata define the data model and the way it is displayed for the users, with the reports, schedules, distribution lists and user security rights.

- **Business metadata** are content from the data warehouse described in more user-friendly terms. The business metadata tell you what data you have, where they come from, what they mean and what is their relationship is to other data in the data warehouse. Business metadata may also serve as documentation for the DW/BI system. Users who browse the data warehouse are primarily viewing the business metadata.
- **Process metadata** are used to describe the results of various operations in the data warehouse. Within the ETL process, all key data from tasks are logged on execution. This includes start time, end time, CPU seconds used, disk reads, disk writes and rows processed. When troubleshooting the ETL or query process, this sort of data becomes valuable. Process metadata are the fact measurement when building and using a DW/BI system. Some organizations make a living out of collecting and selling this sort of data to companies - in that case the process metadata become the business metadata for the fact and dimension tables. Collecting process metadata is in the interest of business people who can use the data to identify the users of their products, which products they are using and what level of service they are receiving.

12.7.2 Metadata on the Internet

The HTML format used to define web pages allows for the inclusion of a variety of types of metadata, from basic descriptive text, dates and keywords to further advance metadata schemes such as the Dublin Core, e-GMS, and AGLS standards. Pages can also be geo-tagged with coordinates. Metadata may be included in the page's header or in a separate file. Micro formats allow metadata to be added to on-page data in a way that users do not see, but computers can readily access.

12.8 METADATA MANAGEMENT

Metadata management is the end-to-end process and governance framework for creating, controlling, enhancing, attributing, defining and managing a metadata schema, model or other structured aggregation, either independently or within a repository and the associated supporting processes (often to enable the management of content). The world Wide Web Consortium (W3C) has identified Governance as a key challenge in the advancement of third generation Web Technologies (Web 3.0, Semantic Web), and a number of research prototypes, such as S3DB, explore the use of semantic modelling to identify practical solutions.

12.8.1 Database management

Each relational database system has its own mechanisms for storing metadata. Examples of relational-database metadata include:

- Tables of all tables in a database, their names, sizes and number of rows in each table.
- Tables of columns in each database, what tables they are used in, and the type of data stored in each column.

In database terminology, this set of metadata is referred to as the catalogue. The SQL standard specifies a uniform means to access the catalogue, called the information schema, but not all databases implement it, even if they implement other aspects of the SQL

standard. For an example of database-specific metadata access methods, see Oracle metadata. Programmatic access to metadata is possible using APIs such as JDBC, or Schema Crawler.

12.9 SUMMARY

Metadata is "data about data". There are two "metadata types" structural metadata, about the design and specification of data structures or "data about the containers of data" and descriptive metadata about individual instances of application data or the data content. Earlier Metadata was traditionally in the card catalogues of libraries. As information has become increasingly available in digital form, metadata are also used to describe digital data using metadata standards specific to a particular discipline. For example, a webpage may include metadata specifying what language it is written in, what tools were used to create it, and where to go for more on the subject, allowing browsers to automatically improve the experience of users. The main purpose of metadata is to facilitate in the discovery of relevant information, more often classified as resource discovery. Metadata also helps organize electronic resources, provide digital identification, and helps support archiving and preservation of the resource. Metadata assists in resource discovery by allowing resources to be found by relevant criteria, identifying resources, bringing similar resources together, distinguishing dissimilar resources, and giving location information.

12.10 TECHNICAL TERMS

XMP: Extremisable Meta Data Plat Form

IPTC : International Press Telecommunications Council

PLUS: Picture Licensing Universal System

DCMI : Dublin Core Metadata Initiative

CIPA : Camera & Imaging Products Association

12.11 SUGGESTED READINGS

1. Bretherton, F. P. and Singley, P.T. Metadata: A User's View, Proceedings of the International Conference on Very Large Data Bases.1994 pp. 1091–1094.
2. Cathro, Warwick (1997). "Metadata: an overview". Retrieved from http://www.nla.gov.au/.
3. Guenther, Rebecca and Jaqueline Radebaugh. *Understanding Metadata*. Bethesda, MD: NISO Press, 2004
4. Haynes, David. Metadata for Information Management and Retrieval. London : Facet Publishing, 2004.
5. Kimball, Ralph . The Data Warehouse Lifecycle Toolkit. (Second Edition d.). New York: Wiley.2008.
6. Kunze, J. and T. Baker (2007). "The Dublin Core Metadata Element Set". ietf.org. Retrieved 17 August 2013.

LESSON - 13

DUBLIN CORE

AIMS AND OBJECTIVES

The objective of this lesson is to explain the basic features of Dublin core Meta data standards for cataloguing web resources. It explains the basic structure of the Dublin core code, organization of description of the web resources of information and levels of description etc.

After studying this lesson you will be able to

- What is Dublin Core
- HISTORY OF Dublin Core
- Levels of the Dublin Core Standard
- Syntax of Dublin Core

Structure

13.1 INTRODUCTION

The Dublin Core metadata terms are a set of vocabulary terms which can be used to describe resources for the purposes of identification/discovery. The terms can be used to describe a full range of web resources (video, images, web pages, etc.), physical resources such as books and objects like artworks. The original set of 15 classic metadata terms, known as the Dublin Core Metadata Element Set are endorsed in the following standards documents:

- IETF RFC 5013
- ISO Standard 15836-2009
- NISO Standard Z39.85

Dublin Core Metadata can be used for multiple purposes, from simple resource description, to combining metadata vocabularies of different metadata standards, to providing interoperability for metadata vocabularies in the Linked data cloud and Semantic web implementations.

13.2 BACKGROUND

"Dublin" refers to Dublin, Ohio, USA where the work originated during the 1995 invitational OCLC/NCSA Metadata Workshop, hosted in by Online Computer Library Center (OCLC), a library consortium based there, and the National Center for Supercomputing Applications (NCSA). "Core" refers to the metadata terms as "broad and generic being usable for describing a wide range of resources".

The semantics of Dublin Core were established and are maintained by an international, cross-disciplinary group of professionals from librarianship, computer science, text encoding, museums, and other related fields of scholarship and practice.

The Dublin Core Metadata Initiative (DCMI) provides an open forum for the development of interoperable online metadata standards for a broad range of purposes and of business models. DCMI's activities include consensus-driven working groups, global conferences and workshops, standardization, and educational efforts to promote widespread acceptance of metadata standards and practices. In 2008, DCMI separated from OCLC and incorporated as an independent entity.

13.3 LEVELS OF THE STANDARD

The Dublin Core standard includes two levels viz. Simple and Qualified.

13.3.1 Simple Dublin Core

The Simple Dublin Core Metadata Element Set (DCMES) consists of 15 metadata elements:

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

Each Dublin Core element is optional and may be repeated. The DCMI has established standard ways to refine elements and encourage the use of encoding and vocabulary schemes. There is no prescribed order in Dublin Core for presenting or using the elements. The Dublin Core became ISO 15836 standard in 2006 and is used as a base-level data element set for the description of learning resources in the ISO/IEC 19788-2 Metadata for learning resources (MLR) -- Part 2: Dublin Core elements, prepared by the ISO/IEC JTC1 SC36.

Example of code

Eg.1.

```
<meta name=”DC.Format” content=”video/mpeg; 10 minutes>
<meta name=”DC.Language” content=”en”>
<meta name=”DC.Publisher” content=”publisher-name”>
<meta name=”DC.Title” content=”HYP”>
```

Eg.2.

```
<meta name="dc.language"CONTENT="UK">
<meta name="dc.source"CONTENT="http://www.your-domain.com/">
<meta name="dc.relation"CONTENT="http://www.second-domain.com/">
<meta name="dc.title"CONTENT="a title">
<meta name="dc.keywords"CONTENT="more keywords">
<meta name="dc.subject"CONTENT="th esubject">
<meta name="dc.description" CONTENT="A description of the content">
```

13.3.2 Qualified Dublin Core

Qualified Dublin Core includes three additional elements (Audience, Provenance and Rights Holder), as well as a group of element refinements (also called qualifiers) that refine the semantics of the elements in ways that may be useful in resource discovery. Subsequent to the specification of the original 15 elements, an ongoing process to develop exemplary terms extending or refining the Dublin Core Metadata Element Set (DCMES) was begun. The additional terms were identified, generally in working groups of the Dublin Core Metadata Initiative, and judged by the DCMI Usage Board to be in conformance with principles of good practice for the qualification of Dublin Core metadata elements.

Elements refinements make the meaning of an element narrower or more specific. A refined element shares the meaning of the unqualified element, but with a more restricted scope. The guiding principle for the qualification of Dublin Core elements, colloquially known as the *Dumb-Down Principle*, states that an application that does not understand a specific element refinement term should be able to ignore the qualifier and treat the metadata value as if it were an unqualified (broader) element. While this may result in some loss of specificity, the remaining element value (without the qualifier) should continue to be generally correct and useful for discovery.

In addition to element refinements, Qualified Dublin Core includes a set of recommended encoding schemes, designed to aid in the interpretation of an element value. These schemes include controlled vocabularies and formal notations or parsing rules. A value expressed using an encoding scheme may thus be a token selected from a controlled vocabulary (for example, a term from a classification system or set of subject headings) or a string formatted in accordance with a formal notation, for example, "2000-12-31" as the ISO standard expression of a date. If an encoding scheme is not understood by an application, the value may still be useful to human reader.

Audience, Provenance and **Rights Holder** are elements, but not part of the Simple Dublin Core 15 elements. Use Audience, Provenance and Rights Holder only when using Qualified Dublin Core. DCMI also maintains a small, general vocabulary

recommended for use within the element Type. This vocabulary currently consists of twelve terms.

13.4 SYNTAXES OF DUBLIN CORE METADATA

Syntax choices for Dublin Core metadata depend on a number of variables. When considering an appropriate syntax, it is important to note that Dublin Core concepts and semantics are designed to be syntax independent. These are equally applicable in a variety of contexts, as long as the metadata is in a form suitable for interpretation both by machines and by human beings.

The Dublin Core Abstract Model provides a reference model against which particular Dublin Core encoding guidelines can be compared, independent of any particular encoding syntax. Such a reference model allows implementers to gain a better understanding of the kinds of descriptions they are trying to encode and facilitates the development of better mappings and translations between different syntaxes.

13.5 SUMMARY

The Dublin Core metadata terms are a set of vocabulary terms used to describe a full range of web resources (video, images, web pages, etc.), physical resources such as books etc. It is an open source standard which can be improved by any one. The Dublin Core Metadata initiative (DCMI) provides an open forum for the development of interoperable online metadata standards for a broad range of purposes and of business models.

13.6 TECHNICAL TERMS

NCSA: National Centre for Super Competing Applications

DCMI: Dublin Core Metadata Initiative

DOMES : Simple Dublin core Metadata Element set

MLR: Metadata for learning resources

13.7 SUGGESTED READINGS

1. http://dublincore.org/
2. "DCMI Metadata Basics". http://dublincore.org/metadata-basics/
3. http://dublincore.org/documents/dcmi-terms/
4. Caplan, Priscilla. Metadata fundamentals for all librarians. ALA, 2003.

LESSON - 14

INDEXING SYSTEMS

AIMS AND OBJECTIVES

The objective of this lesson is to explain what is subject indexing? Kinds of indexing systems and techniques of indexing are also discussed in this lesson. The Chain indexing technique is dealt in detail.

After studying this lesson you will be able to know

- what is subject indexing
- kinds of indexing
- indexing techniques
- chain indexing procedure

Structure

14.1 INTRODUCTION

Subject indexing is the act of describing or classifying a document by index terms or other symbols in order to indicate the content of the document. In other words, it is about identifying and describing the subject of documents. Subject indexing is used in information retrieval especially to create bibliographic databases to retrieve documents on a particular subject. The process of indexing begins with analysis of the subject of the document. The indexer must then identify terms which appropriately identify the subject either by extracting words directly from the document in the case of derived indexing or assigning words from a controlled vocabulary in the case of assigned indexing. The terms in the index are then presented in a systematic order. The Indexers decide how many terms to include and how specific the terms should be.

14.2 KINDS OF INDEXING

Depending upon the terms used to represent the subject content of the document being indexed there are two kinds of indexing. If the terms to represent the subject of the document are taken from the document itself, it is known as derived indexing. If the terms are taken from controlled vocabularies and assigned to represent the subject of the document being indexed, it is called assigned indexing.

14.2.1 Derived indexing

Derived indexing involves taking words directly from the document. It uses natural language. Derived indexing lends itself well to automated techniques where word frequencies are calculated and those with a frequency over a pre-determined value are used as index terms. A stop-list containing common words such as ‘the’, ‘introduction’, ‘an’, ‘and’ etc. would be referred to and such stop words would be excluded as index terms. It is also called Extraction indexing since it involves extraction of terms from the natural language terms used to describe the subject of the document being indexed.

Automated derived indexing may lead to loss of meaning of terms by indexing single words as opposed to phrases. Although it is possible to extract commonly occurring phrases, it becomes more difficult if key concepts are inconsistently worded in phrases. Automated derived indexing also has the problem that even with use of a stop-list to remove common words such as “the,” some frequent words may not be useful for allowing discrimination between documents. For example, the term glucose is likely to occur frequently in any document related to diabetes. Therefore use of this term would likely return most or all the documents in the database. Post-co-ordinated indexing where terms are combined at the time of searching would reduce this effect but the responsibility would be on the searcher to link appropriate terms as opposed to the information professional.

In addition terms that occur infrequently may be highly significant and included as index terms. One method for allowing rarer terms to be included and common words to be excluded by automated techniques would be a relative frequency approach where frequency of a word in a document is compared to frequency in the database as a whole. Therefore a term that occurs more often in a document than might be expected based on the rest of the database could then be used as an index term, and terms that occur equally frequently throughout will be excluded. Another problem with automated extraction is that it does not recognise when a concept is discussed but is not identified in the text by an indexable keyword.

14.2.2 Assigned indexing

An assigned indexing is where index terms are taken from a controlled vocabulary. This has the advantage of controlling for synonyms as the preferred term is indexed and synonyms or related terms direct the user to the preferred term. This means the user can find records regardless of the specific term used by the author and saves the user from having to know and check all possible synonyms. It also removes any confusion caused by homographs by inclusion of a qualifying term. A third advantage is that it allows the linking of related terms whether they are linked by hierarchy or association, e.g. an index entry for an oral medication may list other oral medications as related terms on the same level of the hierarchy but would also link to broader terms such as treatment. Assigned indexing is used in manual indexing to improve inter-indexer consistency as different indexers will have a controlled set of terms to choose from. Controlled vocabularies do not completely remove inconsistencies as two indexers may still interpret the subject differently.

14.3 INDEXING TECHNIQUES

As you know the process of constructing document surrogates or document representations is called as Subject Indexing. Indexing has to specify exactly the content of documents. This needs a language by which the contents could be described precisely (i.e. Indexing Language).

14.3.1 Types of Indexing Techniques: There are broadly two types of Indexing Techniques or Indexing Methods. They are:

a) Pre-Coordinate Indexing, and

b) Post-Coordinate Indexing.

a) Pre-Coordinate Indexing Systems: A Compound or composite subject is analyzed into its constituent concepts according to a plan and these constituent concepts are then represented by symbols (when using a Classification Scheme) or Words of the Indexing Language in use. The next step is to synthesize or combine the components in order, according to the rules of the language (i.e. coordination of concepts).

- That is the Coordination of concepts contained in a documents is done in-anticipation of readers approach.
- In Pre-coordinate Indexing systems, the Subject Heading consists of two or many index terms and these terms are arranged in a pre-determined order using its own syntax or the citation order.
- The Characteristics of Indexing language are the syntax, the method of rotation or cycling of components, the significant word order, use of relationship symbols, system of references, etc.

The examples are: Chain Indexing, PRECIS, POPSI, etc.

b) Post-Coordinate Indexing Systems: A compound or composite subject is analyzed into its constituent components, but they will be kept separately uncoordinated by the indexer. Coordination of concepts is done at the time of search by the user.These indexing systems could bypass / overcome the

most difficult question of deciding on an order of significance (which eventually controls the approaches that could be provided by a pre-coordinate Indexing system).

E.g. Uniterm Indexing; Zatocoding system; Optical coincidence methods (Peek-a-Boosystem)

14.4 CHAIN INDEXING

14.4.1 Definition and Use

Chain Indexing is the first systematic procedure laid down for derived subject indexing. It is a semi-automatic method of deriving alphabetical subject entries from the chain of successive subdivisions from the classification scheme leading from general to specific level needed to be indexed. This method systematizes the compilation of an index which is intended to show the hierarchical relationship of a specific topic to its broader topics. Thus, the indexer works from bottom to top, that is, from the specific to the general. Though originally developed for the classified catalogue it may also be applied to all other systematically organized indexes including an alphabetical one.

The chain procedure provides indexers with a rational technique in place of 'hit or miss' methods of subject indexing. The mechanism of the system provides general as well as specific approaches for information. It is easier to apply this procedure with a fully faceted scheme but it is independent of any one scheme and can be carried out with any system.

14.4.2 Arrays, Chains and Links

An array is a set of coordinate or collateral classes. A chain is, in fact, a hierarchy of terms in a classification scheme; each term includes all those which follow it. Ranganathan could distinguish between divisions of a basic class which are coordinate and those divisions which are in a hierarchical order and hung like a chain from the basic class. Thus in any one basic class we have as many chains as there are sub-divisions. Each successive step in the chain serves as a link. Links may be of different types such as False link, Unsought link, Sought link, Hidden link and Missing link. A clear idea about these links is necessary in the construction of chains for ultimate formulation of subjects.

False links may assume a number of different forms. It may be one which carries no concept. A false link may represent time concept or phase relation. Finally a false link may represent a class which is not strictly super-ordinate to one below. This is the most difficult type of false link to identify. For example, 600 Technology is not strictly super-ordinate to 610 Medical Science.

An Unsought link represents a concept for which users are not likely to search when looking for the specific subject denoted by the final digit of the class number.

A sought link is neither a false nor an unsought one and is therefore an essential constituent in the construction of chain.

A hidden link holds no specific class number. It is usually represented by a block of numbers rather than a single number.

A missing link corresponds to the missing isolate in the chain with gaps. This is to be inserted in the appropriate place of the chain for proper formulation of subjects.

14.4.3 Establishing the chain

The class number of a subject formulated according to a classification scheme is taken as the base for use in chain indexing. This class number represents the chain of subordinated classes or steps of division from the most general class to the particular subject classified.

Thus 769.56 is made up as follows (DDC 19)

700	The arts. Fine and decorative arts
760	Graphic arts. Print making and prints
769	Prints
769.5	Various specific forms of prints
769.56	Postage stamps and related devices

Similarly, J381 : 7 is made up as follows (CC 6)

J	Agriculture
J3	Food
J38	Seed
J381	Rice
J381 : 7	Harvesting

14.4.4 Setting out the Chain

The class number is written in the left hand column with each digit separate to see what step each represents. But with non-expressive notation it is not feasible. Even Dewey has ‘hidden links’ which should be included.

Starting with the first link and working down step by step each level should be analysed according to schedules. Sometimes word or words representing the subject may have to be altered or supplied.

Single words should be used, but nouns are preferred in plural forms. Long phrases should be edited to make short ones. A word used higher in the chain should not be repeated in lower level to avoid redundant entries. Any vague or unhelpful terms should be omitted. Synonyms should be provided at the appropriate level.

Chain for 769.56 : Document on ‘Postage Stamps’.

700	Fine arts. Arts. Decorative arts
760	Graphic
769	Prints
769.5	(Speciic forms)
769.56	Postage stamps. Philately. Stamp Collecting.

14.4.5 Creating the Index

Index entries are prepared starting with the bottom term, which is the most specific link in the chain, and proceeding upward step by step through the chain, qualifying, where necessary, by a more general term or terms to show the context :

Index entries for the document on Postage stamps

Postal stamps : Philately	769.56
Philately	769.56
Stamp collecting	769.56
Prints : Graphic arts	769
Graphic arts	760
Fine arts	700
Decorative arts	700
Arts, Fine	700

Index entries for the document on harvesting of rice : J381 : 7

Harvesting : Rice : Agriculture	J381 : 7
Rice : Agriculture	J381
Seed : Agriculture	J38
Food : Agriculture	J3
Agriculture	J

14.4.6 Alphabetization

The individual index entries prepared for the document are sorted into alphabetical order and interfiled with other such entries in the subject index of the catalogue.

14.4.7 Choice of Qualifiers

Only those qualifiers which keep the index entries short and clear are used to classify the subject context. This process of qualification ultimately leads to the production of relative index. The subject 'Food', for example, will not only appear in class J 'Agriculture' but also in other classes such as 'Cookery' and 'Medicine'. When items dealing with these aspects of the subject are added to a collection and indexed, the alphabetical index will assume the following appearance:

Food : Agriculture	J3
Food : Animal husbandry	KZ3
Food : Cookery	M31
Food : Medicine	L573

Thus the various aspects of the subject which are separated by the classification are brought together in the index. Repetitious qualifiers and those that indicate a permanent 'generic' relation obvious to the user should be avoided.

Ex.: Physics: Pure Sciences 530

Misleading qualifiers should be avoided because these are not the usual generic associations made by users and might be interpreted as special relationships.

Ex.: Agriculture : Technology 630

(Agricultural technology is really at 631.3)

As the first qualifier dictates sub arrangement in the alphabetical index only 'preferred terms' should be used as qualifiers.

14.4.8 Authority file

It is useful to maintain an authority file in classified order, giving each class number that has been indexed with all the index entries made for it. The file can be checked to see if newly coined synonyms need to be indexed or if any defunct entries should be withdrawn as documents are withdrawn from a class number.

4.9 From reverse to forward rendering

At the initial stage of chain indexing reverse rendering of the subject formulation was suggested. The subject formulation obtained by transforming class number into verbal plane was rendered in an inverted manner in the index to provide a different approach to the file. Ranganathan agreed to the system of forward rendering of the chain in the year1964. Due to this modification subject formulation of a document may be rendered directly as

MEDICINE, CHILD, HEART, DISEASE, BIHAR instead of

BIHAR, DISEASE, HEART, CHILD, MEDICINE

According to reverse rendering, successive entries are obtained as usual by deletion of components.

14.4.10 Rotation of component terms

In order to overcome the problem of disappearing chain, rotation of component terms in the subject heading was suggested by Bhattacharyya and Neelameghan without disturbing syntactical relations between components of the chain. This modification was introduced for dictionary catalogue only. Other than the problem of disappearing chain the modification was also put forward to minimize the problem of physical collocation in alphabetic file.

Efforts to achieve this objective were made by limiting the area of search in the file. The nature of entries derived from this modified procedure will be evident from the following examples:

1 Bihar,/ MEDICINE, Child, Heart, Disease
2 Child, Heart, Disease, Bihar./ MEDICINE
3 Disease, Bihar./ MEDICINE, Child, Heart
4 Heart, Disease, Bihar./ MEDICINE, Child
5 MEDICINE, Child, Heart, Disease, Bihar

From the above examples it becomes clear that rotation of component terms has been made without disturbing syntactical relations of components and thus the original syntax of the chain has been maintained in all other entries. The main chain has been derived with the help of a classification scheme. All the component terms have been

brought to the leading position in turn. Forward rendering has been followed and the main chain is represented here by the example 5. Other entries serve as a reference to the main entry. End of the chain in reference entries is indicated by a stop (.) and a stroke (/) signs. Thus context is provided for the approach terms. A user can go to the specific subject entry, if required, for understanding the context.

14.4.11 Advantages

The importance of chain indexing lies in the fact that it is the first systematic procedure laid down for subject indexing. The method influenced successive systems of subject indexing. It is a systematic and exhaustive procedure based on the hierarchy of steps of division. Therefore all general approaches are provided at the same time as specific approaches. It scores heavily over earlier systems on grounds of economy and speed.

1. **Help from Classification**: You already know that, there is a symbiotic relation between cataloguing and classification. Practitioners in the both fields have to formulate the subject of a document in their respective methods. This similarity can be utilized effectively in a classified file to provide an alphabetic approach to the file. When a structured formulation of the subject is obtained by the classifier with the help of an analytic-synthetic scheme, subject indexer starts from this point to retranslate the structured notation and provides an alphabetical approach. The process thus saves much time of the indexer.

2. Alternative approach through reverse rendering: A classified file reflects a particular arrangement out of different possibilities and thus provides approach from a particular point of view. Any mechanism ensuring an alternative approach is definitely helpful to users. This procedure provides this alternative approach to classified file by changing the order of constituents of subject formulation resulting in reverse rendering.

3. Adaptability with different notational schemes: The system has the potentiality to provide good results if subject headings are derived from a classification scheme having expressive notation. But it has successfully made use of almost all the classification schemes such as Bibliographic Classification, Dewey Decimal Classification, Universal Decimal Classification and of course Colon Classification.

4. Economy: This system does not repeat the 'general to particular' structure of the Classification scheme, but complements it with a 'particular to general' approach. This is done by dropping each index term after it has been indexed and thus the system avoids the permutation of component terms. In a chain of four components only four subject headings are made according to this system. Permutation of four terms would have provided 4 or 24 headings. Thus economy achieved is significant one.

5. Almost mechanical: Because the method is based on the structure of the classification scheme and on the terminology found in the schedules, it is speedy and semi-mechanical operation. Moreover, the successive subject headings are derived by deleting components one by one.

14.4.12 Problems of Chain Indexing

Some of the disadvantages of chain indexing are, disappearing chain, its citation order, its link with scheme of classification etc.

1. **Disappearing Chain:** Out of the subject entries generated for a document only one is specific and others stand for broader classes of successive stages. Thus the specific subject entry will be available to a user only with a particular search formulation. Another aspect of the problem is that the method may generate a number of entries for empty links in the chain in case of a document of a highly specialized field. Here the subject of a document will come at the end of a long chain. It is argued that these empty links create noise in the file. The problem of empty links has been solved by converting them into 'see also' references.

2. Link with Classification scheme: The chain index is bound up with the classification scheme. An index linked with a classification scheme naturally has to share the defects and rigidity of the scheme. Where the scheme does not provide a specific class number for a subject it is necessary to continue to analyse the subject beyond the explicit class number. The extra subject words added are called 'verbal extensions'. After all, the index should compensate for deficiencies in the structure of the classification scheme.

3. Missing links: Sometimes a step of division may go unrepresented by a further digit of the class-number. If the original allocation of subjects was faulty a subject may be represented by some of the coordinate classes in an array, instead of all of them. The schedules in DC 18 indicated this by 'Centred Headings'. These should be treated as separate steps in indexing. For example, 'Field crops':

600	Applied sciences, Technology
630	Agriculture, Farming
633-5	Crops
633	Field

Field crops:	Agriculture	633
Crops :	Agriculture	633 – 635 (and so on)

4. False links: Some steps or digits in the notation do not represent subjects, but are meant for structural devices such as facet indicators or synthetic devices. These should be avoided in indexing.

300	Social sciences
330	Economics
333	Land, Natural resources
0	(Scheduled indicator)
0	(form division)
9	(indicator for area table)
4	Europe

Europe : Natural resources : Economics	333.0094
Natural resources : Economics	333 (and so on)

5. Unsought link: Some steps in the notation may be given verbal equivalents which are unsuitable as index approaches. "Specific types" may be used for synthesis in the schedules. "Others" may be used to fill out the last class of an array. "General principles" and "General special" are often represented by Dewey class numbers. All these should be ignored in the index:

5	Science. Pure sciences
7	Life sciences
2	Human races. Races. Ethnology
8	(Specific races.Notation added from Table 5
9	(other racial groups)
9	(other peoples)
9	(others)
4	Etruscan

Etruscan race : Ethnology	572.89994
Races : Ethnology	572 (and so on)

6. Concept of main class: The basic concept of main class is useful in the method to establish the general context of the subject as well as the contexts of component concepts.

14.5 SUMMARY

The present day inter disciplinary research lead to the production of documents dealing with complex and compound subjects. In order to provide access points to the documents by all concepts that were discussed in the document, the indexer has to use either derived indexing or assigned indexing depending up on the objective of Information storage and retrieval system developed. There are two techniques of indexing such as pre-coordinate and post-coordination of indexing terms. Chain indexing is one of the pre-coordinate indexing techniques.

Dr. S. R. Ranganathan developed a method of indexing, called chain procedure of subject indexing or simply Chain Indexing. Chain Indexing or chain procedure is a mechanical method to derive subject index entries or subject headings from the Class Number of the document. In Chain Procedure the indexer or cataloguer is supposed to start from where the classifier has left. No duplication of work is to be done. He/she has to derive subject headings or class index entries from the digit by digit interpretation of the class number of the document in the reverse direction, to provide alphabetical approach to the subject of the document. This method was distinctly different from the enumerated subject heading systems like LCSH or SLSH. Chain Indexing was originally intended for use with Colon Classification. However, it may be applied to any scheme of classification whose notation follows hierarchical pattern.

14.6 Technical Terms

PRECIS: Preserved Context Index System

POPSI : Postulate-Based Permuted Subject Indexing

14.7 Suggested Readings

1. Rajan, T.N. Indexing Systems. Bangalore, IASLIC, 1981
2. Ranganathan, S.R. Prolegomena to Library Classification, 3rd ed. Bangalore:SRELS, 1989
3. Sinha, M.P. Subject indexing by chain procedure. New Delhi : Academic, 1975

LESSON - 15

DATABASES AND SEARCH STRATEGIES

AIMS AND OBJECTIVES

The objective of this lesson is to explain in detail about database and its search strategies.

After studying this lesson you will understand:

- What is database?
- Types of databases and
- Searching of databases
- Techniques of searching

Structure

15.1 INTRODUCTION

In the last three decades bibliographic search system have become extremely sophisticated. The number of databases available, there sixe and the differences in indexing systems have become considerably simpler and more comprehensive today when compared to their pioneer services in the 1970s. There were very few databases and their size was relatively small. The users were having very limited options. Users need to enter strings of keywords, or sets of single keywords, and combined them with Boolean operators. The ability to search one database and execute it in another database was unknown. Users had no option but to re-enter the search in each database to be searched.

Today, the scenario of databases has completely changed with regard to the number of databases available, their size and the search systems they offer. The search system software capabilities have grown to facilitate efficient database selection and searching with a minimum effort, say bb a click of the mouse. The same search can be carried in a number of databases without re-entry of the search options,. Users can even store their searches offline must spend online to retrieve the desired information. However, there is one thing that the users cannot escape the database searching systems – building the search strategy.

15.2 DATABASES

All databases consist of data (records) described in fields and a mechanism called search engine to search these fields. Databases may look different to appear on screen but the underlying principles for creating, searching and formulating search strategies are common to all.

15.3 TYPES OF DATABASES

15.3.1 Bibliographic databases:

Bibliographic database provide publication details of a document such as book or serial etc, but the document itself is not provided in the database. Document publication details such as author(s), title, subject(s) and publisher etc. are provided. The information provided is called a reference or citation and with this information one should be able to find the document required and locate the same within the Library.

15.3.2 Bibliographic with some text content:

There are some popular databases which indexes journal articles with abstracts. Chemical abstracts and Physics abstracts are good examples of Bibliographic databases with content along with citation. Some such databases often, but not always, include the full text of an article.

15.3.3 Bibliographic databases with full text content:

These databases include the entire full text for all articles and other documents indexed.

15.4 DATABASE SEARCH

To make the records in a database searchable, the information contained within it must be indexed. There are variations in the way in which different database producers index their particular product, but the underlying principles are similar. In general terms, the procedure involves taking all the useful words from a field or part of a record and storing them in an index belonging to that field. Usually, individual words form the article title; author's name, subject terms (also known as keywords/subject headings), abstract and the full text of the article are indexed. These fields may be searched individually or in a 'keyword' search, across more than one 'field'. Phrases that combine two or three words may also be used

e.g. : heart attack/myocardial infarction,
in vitro fertilization,
Asperger's syndrome etc.

15.4.1 Creating a 'Search Strategy'

Before conducting search for any information, one should first develop a search strategy. Think about the concepts that form the basic issues of topic of search. Think about the keywords one can use to search the database for required information. One should consider possible synonyms e.g. car/automobile; alternative spellings e.g. sulphur/3ulphur, labour/labor, organization/organisation and pediatric/paediatric; plurals and other endings of the words using to search the database. The following steps are to be observed to create a search strategy:

1. Define the search topic(s) and break it down into its component parts.
2. The terms, words or phrases that describe the topic are to be identified.
3. The other terms that might be used to describe this topic are to be identified.

To identify the terms that describe the topic of search Encyclopaedias or handbooks of the subject of the topic can be used for background information and terminology. Catalogues, list of subject headings and thesaurus are also useful to select the keywords to search the database. When looking for older materials there is possibility of change in terminology due to the development of a more technical vocabulary

e.g. fire fighter rather than fireman
e.g. visually impaired rather than blind.

15.4.2 Boolean Logic

Boolean logic takes its name from the British mathematician George Boole (1815-1864), who wrote about a system of logic designed to produce better search results by formulating more precise queries. He called it the 'calculus of thought.' From his writings, we have derived Boolean logic and its operators: AND, OR and NOT, which we use to link words or phrases for more precise searches. Using Boolean operators (AND, OR, NOT) will help focus and define our search. They can help broaden (increase) and narrow (decrease) search results. Boolean searching is an important skill to learn these operators are needed to effectively search the library catalogue, electronic databases and the Internet.

AND: The operator 'AND' between the words narrows a search. Because it insists All Terms must be present in each hit. It limits a search by requiring that the search terms before and after AND must both appear in the article for the article to be retrieved by the search process.

e.g. when we search the database with the following string

eating disorders AND children

the results will be all records that contain both words 'eating disorders' and 'children'.

OR: The operator 'OR' widens a search because each hit will contain either term in the search string. The Boolean operator "OR" expands a search by requiring that either search terms before and after OR must appear in the article for the article to be retrieved by the search process.

e.g. **epinephrine OR adrenaline** retrieve articles that contain either epinephrine or adrenaline.

This operator is used when a term can be described in more than one way (a synonym) and will also find results that contain both terms. It may also be used to find variants on words that are hyphenated.

e.g. **x- ray OR xray**

NOT: This operator narrows a search by excluding records containing specified words. The Boolean operator "NOT" also limits a search by requiring that the search term after NOT must not appear in the article when retrieved by the search process

e.g. The strings **bulimia NOT anorexia**

hypertension NOT obesity will yield records that contain only 'bulimia' and hypertension respectively.

Mountain AND (bike or bicycle) yield records which contain words mountain and either bike or bicycle.

Common cold AND (vitamin c OR zinc) yield records which contain the words common cold and either the words vitamin c or zinc

15.5 THESAURUS SEARCH

If a database includes a thesaurus, it is advisable to make use of it. Running a search on a database's own thesaurus can help in the selection of terms that have been indexed in that database. It also serves to suggest related, broader and narrower terms as well as indicating preferred terms. Terms may have a scope (or explanatory) note giving details of their meaning in the context of the database. As an example Medline/PubMED has a medical subject headings 'MESH thesaurus' included.

15.6 OTHER WAYS TO FOCUS OR LIMIT SEARCH

Sometimes the problem in conducting a search is limiting the number of records retrieved to make them more relevant. Another way of looking at this is that in addition to defining what is required from a search, it is equally important to define what is not required and how the search may be limited. A good well-focused search should retrieve manageable records of articles. Retrieval of 30 -50 articles for a query is an ideal situation. Remember, that from this number only a few articles may be relevant to the needs. However from these articles by consulting the list of references or bibliography one can source other articles relevant to the search topic. This process is known as hand-searching.

A useful limiter is date, all databases will allow to focus searches within specified date ranges e.g. articles from the last two or three years. This feature enables to limit the searches to within the most up-to-date articles and as a result the most up-to-date information available. Other methods for limiting searches include limiting the search to one or more specific fields, such as title or abstract, to eliminate items where the search term(s) occur only in the full text.

15.7 SEARCH STRATEGY

A search strategy is the planned and structured organisation of terms used to search a database. The search strategy will also indicate how these terms are combined in order to retrieve optimal results. Reasons for searching the literature will have a significant impact on the strategy to be developed. For example, if only looking for background reading, a shorter and more basic search will suffice. However, carrying out more detailed research (for a thesis or particularly for a systematic review), will need to construct a much more sensitive and comprehensive search strategy, incorporating all relevant search techniques.

Sometimes need may arise to search multiple databases for different aspects of research. Each database works in its own way. So, separate search strategies specific to the database using need to be constructed.

This process can be broken down into 5 common steps known as **SKILL**:

- **S**ummarisation of topic in one or two sentences
- **K**eywords and phrases need to be selected
- **I**dentification of synonyms/ alternate terms and variant spellings
- **L**inking of keywords and phrases
- **L**ocating and evaluating results

a). **Summarisation of topic**

A great way to test the understanding of the assignment topic is to summarise it in the own words of the user. This requires thorough reading of assignment topic or question. One has to concentrate on the aim of the assignment and about the tasks involved. The main concepts are to written in a sentence or paragraph.

b). Selection of keywords

Too many search terms may retrieve no references, or very few references. Because the database is trying to find references that contain all those words. Too few search terms may retrieve too many references. Use of terms that are not commonly used to describe the subject may retrieve irrelevant references. Incorrect spelling of the words used for searching may result in retrieval of no references or irrelevant references.

From the summarised topic written in sentence, keywords and phrases that are relevant to the topic are to be circled, highlighted or underline the keywords. The following example illustrates the procedure:

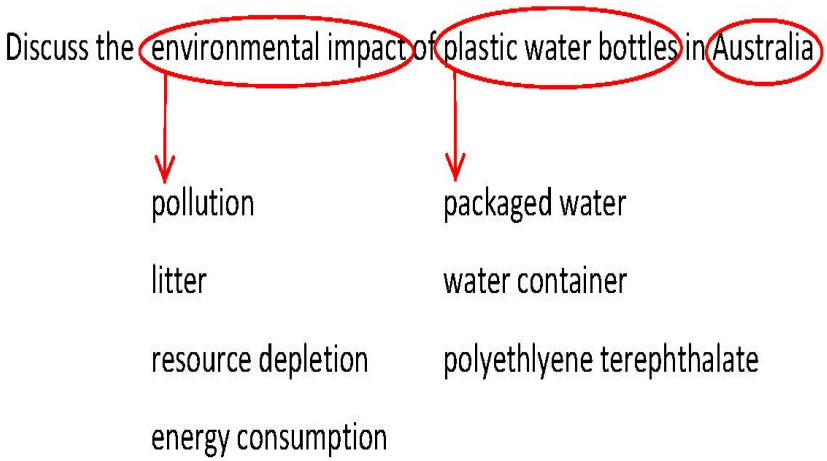
c). Identification of Synonyms / alternative keywords and variant spellings

Researchers or producers of information may not always use exactly the same keyword to represent relevant literature. The possible synonyms or alternate keywords used in literature to represent the same concept are to be identified. Keywords with

variant spellings are to be noted down. Encyclopaedias, Subject dictionaries are helpful tools to identify alternative keywords.

Synonyms

Using some of the keywords from our sample topic taken as example, let us identify some words or phrases that mean the same thing.



Alternative words endings

When searching on some databases, one needs to consider alternative word endings and variant spellings. For instance, if our keyword was '*recycle'* we should consider some derivatives of *recycle* to conduct effective search to retrieve all the relevant: *recycled*, *recycling*, *recyclable* are some other keywords with different word endings to be used in searching the database. All databases will provide a technique to cope up with this problem. The technique is known as truncation. We can represent all derivatives of *recycle* by using an asterix '*recyc*'*

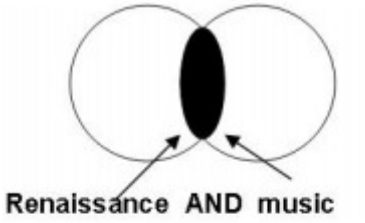
Variant spellings

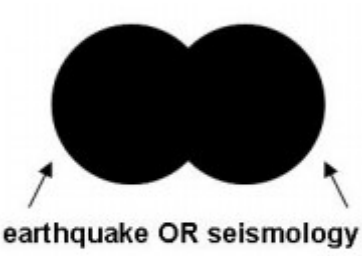
There are often differences in British and American spelling. For instance, if we are looking for information on *Organisation*, we may notice that the American spelling is *organization* while the Australian/British spelling is *organisation*. To solve this problem **wildcard** feature is used in all database packages. For instance, by putting a question mark in *organi?ation*, you will find both the British spelling (organisation) and the American spelling (organization). Symbols used to represent wild card may vary across different databases.

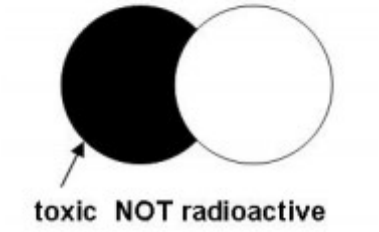
d). Linking of keywords and phrases

The keywords or phrases are to be linked to construct search string for retrieval of desired records from the database. For instance, we may use a combination of words:

Boolean operators: These are connectors used to combine search terms. There are three connectors: AND, OR, NOT. These are described below:







AND: If this operator is placed between words, both words must appear in each reference. This will narrow the search.

For example, *renaissance **AND** music* will retrieve all references which contain both terms.

OR: If this operator is placed between words, either, or all, word/s may appear in each reference. This will **broaden** the search.

For example, *earthquake **OR** seismology* will retrieve all references with *earthquake or seismology*, as well as references with both terms.

NOT: If this operator is placed between words, it means that the second word must not appear in any reference. This will **narrow** the search.

For example, *toxic **NOT** radioactive* will retrieve all references with *toxic*, **except** references which include *radioactive*.

NOT operator is to be used carefully as it may exclude useful references.

Phrase searching: Some databases will assume that a string of words should be searched for as a **phrase**. In other words, it will only retrieve references in

which the words occur side by side or in very close proximity. This works well if you have typed *information technology*, but it will be a problem if you have typed *depression teenagers* (instead of *depression in teenagers*).

If you are searching a database that does **not** automatically search the terms as a phrase, you may find it useful to **force** the database to search them as a phrase. Often you can do this by enclosing the terms in **double quotation marks**, e.g. *"information technology"*. In some databases, there will be a separate search box for phrase searching.

Proximity searching: Some databases allow searching for words within a specified distance of one another. This is particularly important when searching large full-text databases. If one of the required search terms appears on page 3 of an article, and the other search term appears on page 7, the article is probably not very relevant. Proximity searches limit the number of words between the search terms.

For example:

(television) within 5 (violence) retrieves references that contain television and violence in any order, but not more than five words apart.

(television) near (violence) retrieves references that contain television and violence in any order, but within a certain proximity, which is defined by the database (perhaps in the same sentence, or in the same paragraph).

Search syntax for a proximity search will vary from database to database. Database's *Help* screens or *Searching Tips* shows whether the database allows proximity searching and, if so, how to construct search statement. Generally the syntax is expressed variously as w/n, near/n, n/n, or NEAR (as in *television* w/5 *violence*).

e). Locating and evaluating results

Search strategy need to be revised when the results of the search are too many or too few.

If the results are too many we have to:

- Use different keywords or phrases
- Limiting search results by document type, date, subject etc.
- Conduct search in a particular field (title or abstract fields).

If there are few results for a search we have to:

- check spelling
- remove some of the keywords
- try alternate keywords and phrases
- try alternate databases

15.8 SUMMARY

A Database is an organized collection of data. It is the collection of schemes, tables, queries, reports, views and other objects. Smallest unit of discrete data is stored in the Field. A group of related fields constitute a Record. A group of related records constitute Database. Database may store information itself or the bibliographic data of information containers.

A database management system (DBMS) is a computer software application that interacts with the user. A general-purpose DBMS is designed to allow the definition, creation, querying, update, and administration of databases. Database management systems are often classified according to the database model that they support, The most popular database systems since the 1980s have all supported the relational model.

Search strategy is a structured organisation of terms used to search a database. The search strategy shows how these terms combine in order to retrieve the best results. Different databases work in different ways. Adaption of search strategy depends upon the database selected for search. Sometimes need arises to develop separate search strategies for different aspects of research. One need to test the strategies adopted several times and refining them based on the results retrieved from the database.

15.9 TECHNICAL TERMS

DBMS: Database Management System

15.10 SUGGESTED READINGS

1. Date, C.J. An introduction to Database systems. Reading, Massachusetts : Addison-wesley, 1985.
2. Elmasri, Ramez and Navathe, Shamkant B. Fundamentals of Database Systems 6th Edition. New Delhi : Prentice Hall, 2006
3. Fenichel, Carol H. And Hagan, Thomas H. Online searching: a primer. New Jersey: Learned Information Inc., 1981
4. Gupta, S.K. Data Base Management Systems (DBMS) by IIT Video Lectures
5. Henry, W.M. Online Searching : an introduction. London : Butterworth, 1982
6. Kashyap, M.M. Database Systems : design and development. New Delhi : Sterling publishers, 1993
7. Martin, James. Computer database organization. 2nd edition. New delhi : PHI, 1984
8. Ramakrishnan, Raghu and Gehrke, Johannes.Database Management Systems. 3rd Edition. New York : McGraw-Hill, 2003.
9. Ullman, Jeffrey D. Principles of Database systems. New Delhi : PHI, 2003

LESSON 16

EVALUATION OF INFORMATION RETRIEVAL SYSTEMS

AIMS AND OBJECTIVES

The objective of this lesson is to explain various criteria used to evaluate Information Storage and Retrieval Systems. The lesson also discussed the history of development of Information Retrieval Systems.

After studying this lesson you will understand the

- meaning of Information Retrieval
- history of development of Information Retrieval Systems
- measures employed in evaluation of Information Retrieval Systems and
- some Information Retrieval Models.

Structure

16.1 INTRODUCTION

Information retrieval is the activity of obtaining information directly or resources containing information relevant to need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing.

Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible examples of Information Retrieval applications.

16.2 HISTORY

The idea of using computers to search for relevant pieces of information was popularized in the article *As We May Think* by Vannevar Bush in 1945. The first automated information retrieval systems were introduced in the 1950s and 1960s. By 1970 several different techniques had been shown to perform well on small text searches such as the Cranfield collection (several thousand documents). Large-scale retrieval systems, such as the Lockheed Dialog system, came into use early in the 1970s.

In 1992, the US Department of Defence along with the National Institute of Standards and Technology (NIST), cosponsored the Text Retrieval Conference (TREC) as part of the TIPSTER text program. The aim of this was to look into the information retrieval community by supplying the infrastructure that was needed for evaluation of text retrieval methodologies on a very large text collection. This catalyzed research on methods that scale to huge corpora. The introduction of web search engines has boosted the need for very large scale retrieval systems even further.

16.3 TIMELINE

- Before the **1900s**
  - **1801**: Joseph Marie Jacquard invents the Jacquard loom, the first machine to use punched cards to control a sequence of operations.
  - **1880s**: Herman Hollerith invents an electro-mechanical data tabulator using punch cards as a machine readable medium.
  - **1890** Hollerith cards, keypunches and tabulators used to process the 1890 US Census data.
- **1920s-1930s**
  - Emanuel Goldberg submits patents for his "Statistical Machine" a document search engine that used photoelectric cells and pattern recognition to search the metadata on rolls of microfilmed documents.
- **1940s–1950s**
  - **Late 1940s**: The US military confronted problems of indexing and retrieval of wartime scientific research documents captured from Germans.
  - **1945**: Vannevar Bush's *As We May Think* appeared in *Atlantic Monthly*.
  - **1947**: Hans Peter Luhn (research engineer at IBM since 1941) began work on a mechanized punch card-based system for searching chemical compounds.
  - **1950s**: Growing concern in the US for a "science gap" with the USSR motivated, encouraged funding and provided a backdrop for

mechanized literature searching systems (Allen Kent *et al.*) and the invention of citation indexing (Eugene Garfield).

1950: The term "information retrieval" appears to have been coined by Calvin Mooers.

1951: Philip Bagley conducted the earliest experiment in computerized document retrieval in a master thesis at MIT.[3]

1955: Allen Kent joined Case Western Reserve University, and eventually became associate director of the Centre for Documentation and Communications Research. That same year, Kent and colleagues published a paper in American Documentation describing the precision and recall measures as well as detailing a proposed "framework" for evaluating an IR system which included statistical sampling methods for determining the number of relevant documents not retrieved.

1958: International Conference on Scientific Information Washington DC included consideration of IR systems as a solution to problems identified. See: *Proceedings of the International Conference on Scientific Information, 1958* (National Academy of Sciences, Washington, DC, 1959)

1959: Hans Peter Luhn published "Auto-encoding of documents for information retrieval."

- **1960s**:

Early 1960s: Gerard Salton began work on IR at Harvard, later moved to Cornell.

1960: Melvin Earl Maron and John Lary Kuhns[4] published "On relevance, probabilistic indexing, and information retrieval" in the Journal of the ACM 7(3):216–244, July 1960.

1962: Cyril W. Cleverdon published early findings of the Cranfield studies, developing a model for IR system evaluation. See: Cyril W. Cleverdon, "Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems". Cranfield Collection of Aeronautics, Cranfield, England, 1962.
Kent published *Information Analysis and Retrieval*.

1963:Weinberg report "Science, Government and Information" gave a full articulation of the idea of a "crisis of scientific information." The report was named after Dr. Alvin Weinberg.
Joseph Becker and Robert M. Hayes published text on information retrieval. Becker, Joseph; Hayes, Robert Mayo.*Information storage and retrieval: tools, elements, theories*. New York, Wiley (1963).

1964:Karen Spärck Jones finished her thesis at Cambridge, *Synonymy and Semantic Classification*, and continued work oncomputational linguistics as it applies to IR.
The National Bureau of Standards sponsored a symposium titled "Statistical Association Methods for Mechanized Documentation." Several highly significant papers, including G. Salton's first published reference (we believe) to the SMARTsystem.

Mid-1960s:National Library of Medicine developed MEDLARS Medical Literature Analysis and Retrieval System, the first major machine-readable database and batch-retrieval system.
Project Intrex at MIT.

1965: J. C. R. Licklider published *Libraries of the Future*.

1966: Don Swanson was involved in studies at University of Chicago on Requirements for Future Catalogs.

Late 1960s: F. Wilfrid Lancaster completed evaluation studies of the MEDLARS system and published the first edition of his text on information retrieval.

1968:Gerard Salton published *Automatic Information Organization and Retrieval*.

John W. Sammon, Jr.'s RADC Tech report "Some Mathematics of Information Storage and Retrieval..." outlined the vector model.

1969: Sammon's "A nonlinear mapping for data structure analysis" (IEEE Transactions on Computers) was the first proposal for visualization interface to an IR system.

- **1970s**

Early 1970s: First online systems—NLM's AIM-TWX, MEDLINE; Lockheed's Dialog; SDC's ORBIT.

Theodor Nelson promoting concept of hypertext, published *Computer Lib/Dream Machines*.

1971: Nicholas Jardine and Cornelis J. van Rijsbergen published "The use of hierarchic clustering in information retrieval", which articulated the "cluster hypothesis."

1975: Three highly influential publications by Salton fully articulated his vector processing framework and term discrimination model:

A Theory of Indexing (Society for Industrial and Applied Mathematics)

A Theory of Term Importance in Automatic Text Analysis

A Vector Space Model for Automatic Indexing

1978: The First ACM SIGIR conference.

1979:C.J. van Rijsbergen published *Information Retrieval* (Butterworths). Heavy emphasis on probabilistic models.

- **1980s**

1980: First international ACM SIGIR conference, joint with British Computer Society IR group in Cambridge.

1982: Nicholas J. Belkin, Robert N. Oddy, and Helen M. Brooks proposed the ASK (Anomalous State of Knowledge) viewpoint for information retrieval. This was an important concept, though their automated analysis tool proved ultimately disappointing.

1983: Salton (and Michael J. McGill) published *Introduction to Modern Information Retrieval* (McGraw-Hill), with heavy emphasis on vector models.

1985: David Blair and Bill Maron publish: An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System

mid-1980s: Efforts to develop end-user versions of commercial IR systems.

1985–1993: Key papers on and experimental systems for visualization interfaces.

Work by Donald B. Crouch, Robert R. Korfhage, Matthew Chalmers, Anselm Spoerri and others.

1989: First World Wide Web proposals by Tim Berners-Lee at CERN.

- **1990s**

1992: First TREC conference.
1997: Publication of Korfhage's *Information Storage and Retrieval*[6] with emphasis on visualization and multi-reference point systems.
Late 1990s: Web search engines implementation of many features formerly found only in experimental IR systems. Search engines become the most common and maybe best instantiation of IR models, research, and implementation.

16.4 OVERVIEW

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

An object is an entity that is represented by information in a database. User queries are matched against the database information. Depending on the application the data objects may be, for example, text documents, images, audio, mind maps or videos. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.

16.5 PERFORMANCE AND CORRECTNESS MEASURES

Many different measures for evaluating the performance of information retrieval systems have been proposed. The measures require a collection of documents and a query. All common measures described here assume a ground truth notion of relevancy: every document is known to be either relevant or non-relevant to a particular query. In practice queries may be ill-posed and there may be different shades of relevancy.

16.5.1 Precision

Precision is the fraction of the documents retrieved that are relevant to the user's information need.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

In binary classification, precision is analogous to positive predictive value. Precision takes all retrieved documents into account. It can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n or P@n.

Note that the meaning and usage of "precision" in the field of Information Retrieval differs from the definition of accuracy and precision within other branches of science and technology.

16.5.2 Recall

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

In binary classification, recall is often called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query.

It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision.

16.5.3 Fall-out

The proportion of non-relevant documents that are retrieved, out of all non-relevant documents available:

$$\text{fall-out} = \frac{|\{\text{non-relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{non-relevant documents}\}|}$$

In binary classification, fall-out is closely related to specificity and is equal to (1 – specificity). It can be looked at as the probability that a non-relevant document is retrieved by the query.

It is trivial to achieve fall-out of 0% by returning zero documents in response to any query.

F-measure

The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is:

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}.$$

This is also known as the F_1 measure, because recall and precision are evenly weighted.

The general formula for non-negative real β is:

$$F_\beta = \frac{(1 + \beta^2) \cdot (\text{precision} \cdot \text{recall})}{(\beta^2 \cdot \text{precision} + \text{recall})}.$$

Two other commonly used F measures are the F_2 measure, which weights recall twice as much as precision, and the $F_{0.5}$ measure, which weights precision twice as much as recall.

The F-measure was derived by Van Rijsbergen (1979) so that F_β "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision". It is based on van Rijsbergen's effectiveness measure $E = 1 - \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}$. Their relationship is $F_\beta = 1 - E$ where $\alpha = \frac{1}{1+\beta^2}$.

16.5.4 Average precision

Precision and recall are single-value metrics based on the whole list of documents returned by the system. For systems that return a ranked sequence of documents, it is desirable to also consider the order in which the returned documents are presented. By computing a precision and recall at every position in the ranked sequence of documents, one can plot a precision-recall curve, plotting precision $p(r)$ as a function of recall r. Average precision computes the average value of $p(r)$ over the interval from $r = 0$ to $r = 1$:

$$\mathrm{AveP} = \int_0^1 p(r)dr$$

That is the area under the precision-recall curve. This integral is in practice replaced with a finite sum over every position in the ranked sequence of documents:

$$\mathrm{AveP} = \sum_{k=1}^{n} P(k)\Delta r(k)$$

where k is the rank in the sequence of retrieved documents, n is the number of retrieved documents, $P(k)$ is the precision at cut-off k in the list, and $\Delta r(k)$ is the change in recall from items $k-1$ to k.[11]

This finite sum is equivalent to:

$$\mathrm{AveP} = \frac{\sum_{k=1}^{n}(P(k) \times \mathrm{rel}(k))}{\text{number of relevant documents}}$$

where $\mathrm{rel}(k)$ is an indicator function equaling 1 if the item at rank k is a relevant document, zero otherwise.[12] Note that the average is over all relevant documents and the relevant documents not retrieved get a precision score of zero.

Some authors choose to interpolate the $p(r)$ function to reduce the impact of "wiggles" in the curve.[13][14] For example, the PASCAL Visual Object Classes challenge (a benchmark for computer vision object detection) computes average precision by averaging the precision over a set of evenly spaced recall levels {0, 0.1, 0.2, ... 1.0}:

$$\text{AveP} = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1.0\}} p_{\text{interp}}(r)$$

where $p_{\text{interp}}(r)$ is an interpolated precision that takes the maximum precision over all recalls greater than r:

$$p_{\text{interp}}(r) = \max_{\tilde{r}:\tilde{r} \geq r} p(\tilde{r})$$

An alternative is to derive an analytical $p(r)$ function by assuming a particular parametric distribution for the underlying decision values. For example, a binormal precision-recall curve can be obtained by assuming decision values in both classes to follow a Gaussian distribution.

16.5.5 R-Precision

Precision at **R**-th position in the ranking of results for a query that has **R** relevant documents. This measure is highly correlated to Average Precision. Also, Precision is equal to Recall at the **R**-th position.

16.5.6 Mean average precision

Mean average precision for a set of queries is the mean of the average precision scores for each query.

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP(q)}}{Q}$$

where Q is the number of queries.

16.5.7 Discounted cumulative gain

DCG uses a graded relevance scale of documents from the result set to evaluate the usefulness, or gain, of a document based on its position in the result list. The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result.

The DCG accumulated at a particular rank position p is defined as:

$$\text{DCG}_\text{p} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}.$$

Since result set may vary in size among different queries or systems, to compare performances the normalised version of DCG uses an ideal DCG. To this end, it sorts documents of a result list by relevance, producing an ideal DCG at position p ($IDCG_p$), which normalizes the score:

$$\text{nDCG}_\text{p} = \frac{DCG_p}{IDCGp}.$$

The nDCG values for all queries can be averaged to obtain a measure of the average performance of a ranking algorithm. Note that in a perfect ranking

algorithm, the DCG_p will be the same as the $IDCG_p$ producing an nDCG of 1.0. All nDCG calculations are then relative values on the interval 0.0 to 1.

16.6 MODEL TYPES

For effectively retrieving relevant documents by IR strategies, the documents are typically transformed into a suitable representation. Each retrieval strategy incorporates a specific model for its document representation purposes. The picture on the right illustrates the relationship of some common models. In the picture, the models are categorized according to two dimensions: the mathematical basis and the properties of the model.

16.6.1 First dimension: mathematical basis

Set-theoretic models represent documents as sets of words or phrases. Similarities are usually derived from set-theoretic operations on those sets. Common models are:

- Standard Boolean model
- Extended Boolean model
- Fuzzy retrieval
- Algebraic models represent documents and queries usually as vectors, matrices, or tuples. The similarity of the query vector and document vector is represented as a scalar value.
- Vector space model
- Generalized vector space model
- (Enhanced) Topic-based Vector Space Model
- Extended Boolean model
- Latent semantic indexing aka latent semantic analysis
- Probabilistic models treat the process of document retrieval as a probabilistic inference. Similarities are computed as probabilities that a document is relevant for a given query. Probabilistic theorems like the Bayes' theorem are often used in these models.
- Binary Independence Model
- Probabilistic relevance model on which is based the okapi (BM25) relevance function
- Uncertain inference
- Language models
- Divergence-from-randomness model
- Latent Dirichlet allocation
- Feature-based retrieval models view documents as vectors of values of feature functions (or just features) and seek the best way to combine these features into a single relevance score, typically by learning to rank methods. Feature functions are arbitrary functions of document and query, and as such can easily incorporate almost any other retrieval model as just a yet another feature.

16.6.2 Second dimension: properties of the model

Models without term-interdependencies treat different terms/words as independent. This fact is usually represented in vector space models by the orthogonality assumption of term vectors or in probabilistic models by an independency assumption for term variables.

- Models with immanent term interdependencies allow a representation of interdependencies between terms. However the degree of the

interdependency between two terms is defined by the model itself. It is usually directly or indirectly derived (e.g. by dimensional reduction) from the co-occurrence of those terms in the whole set of documents.

- Models with transcendent term interdependencies allow a representation of interdependencies between terms, but they do not allege how the interdependency between two terms is defined. They relay an external source for the degree of interdependency between two terms. (For example a human or sophisticated algorithms.)

16.7 SUMMARY

Information retrieval (IR) has experienced huge growth in the past decade as increasing numbers and types of information systems are being developed for end-users. Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Several measures employed in evaluation of Information Retrieval Systems. Precision, Recall and Fall out Ratio are some of the measures adopted for evaluating Information Storage and Retrieval systems. As the precision and Recall are inversely proportional the best ISRAS must provide optimum Recall and precision.

16.8 SUGGESTED READINGS

1. Baeza-Yates, Ricardo. Modern Information Retrival. New Delhi : Pearson Education, 2007.
2. Grossman, David A. And and Ophir Frieder. Information Retrieval: Algorithms and Heuristics.2nd Edition.New York : Springer International Edition, 2004.
3. Frakes, William B and Ricardo Baeza-Yates. Information Retrieval Data Structures and Algorithms. New Delhi : Pearson Education, 1992.
4. Korfhage, Robert. Information Storage & Retieval. New York : John Wiley & Sons. 2001
5. Kochtanek, Thomas R. And Joseph R. Matthews. Library Information Systems: from library automation to distributed information access solution. Westport : Library unlimited Inc. 2004
6. Manning, Christopher D. and Prabhakar Raghavan. Introduction to Information Retrieval. Cambridge : Cambridge University Press, 2008.

LESSON - 17
EVALUATION OF INFORMATION RETRIEVAL SYSTEMS CRANFIELD PROJECT STUDIES

AIMS AND OBJECTIVES

The objective of this lesson is to explain various criteria used to evaluate Information Storage and Retrieval Systems. The lesson also discussed the Cranfield Project studies, which are the landmarks in the history of development of evaluation of Information Retrieval Systems.

After studying this lesson you will understand the

- meaning of Information Retrieval
- criteria used for evaluation of Information Retrieval Systems (IRS)
- Cranfield projects of evaluation

Structure

17.1 INTRODUCTION

Information is an all pervasive resource in every human activity and helps in establishing a continuum from the past and ultimately the future. The natural consequences of all activities and stored s tend to generate large amount of information. This information is recorded and stored only when it is expected to have potential importance. Thus information is accumulated. This accumulated information is to be organised to facilitate access when needed by the users. It is to meet this situation Information storage and retrieval systems are designed and used. Consequently the librarians and information scientists were placed in dilemma regarding the choice of a particular method of information storage and retrieval. In other words evaluation in regard to the workability and efficiency of the systems became a necessity. Evaluation in this context means measuring the performance of the system in terms of retrieval efficiency. The Cranfield indexing experiments in the 1960s are often cited as the beginning of the modern era of information retrieval system evaluation.

17.2 INFORMATION RETRIEVAL SYSTEM (IRS)

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing. The term IR was introduced by Calvin Mooers in 1951, who defined it as "Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him…"

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs. In information retrieval, a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy. An object is an entity that is represented by information in a database. User's queries are matched against the database information. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata. Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.

The whole work of information storage and retrieval can be broadly divided into three areas, viz. Content Analysis of documents, Representation of the content in a suitable form of records and creation of file and actual retrieval of information or document surrogates. Information retrieval system on a whole utilises various tools and techniques such as indexing languages and searching techniques etc.

The users are basically interested in ease of use, speed of search, accuracy of the results of search. The librarians are concerned with cost-effectiveness and cost-benefit of the system. An overall assessment of the information system takes care of needs of users and librarians.

17.3 CRITERIA OF EVALUATION OF INFORMATION RETRIEVAL SYSTEM

Perry and Kent are credited for bringing the concept of evaluation into information retrieval systems in the mid-fifties. The evaluation measures they suggested were:

L / N = Resolution factor	(N – L / N) = Elimination factor
R / L = Pertinence factor	(L – R) / L = Noise factor
R / C = Recall factor	(C – R) / C = Omission factor

Where N = Total number of documents
L = Number of retrieved documents
C = Number of relevant documents
R = Number of documents that are both retrieved and relevant

However only two measures – namely, Precision (new name for pertinence), and Recall are presently used in evaluation studies.

An evaluation programme may be subjective, which categorises into very useful, useful, and not useful without qualifying the levels of performance. On the other hand evaluation may be quantitative and this implies use of some yardsticks to express the degree of success or failure of system. To conduct a quantitative evaluation, criteria

must be established to measure the success or failure of search. Following is the list of significant criteria identified for evaluation of information retrieval system:

1. Recall
2. Precision
3. Response time
4. User effort
5. Coverage
6. Form of output

The criteria mentioned above have been discussed in detail in the previous lesson. The first two points are of prime importance in respect of user's requirements and these are inherent in the qualities of indexing languages. The third and fourth criteria are marginally concerned with the nature of the indexing language. The last two criteria are independent of any indexing language. They depend on the nature of information system and thus are of librarian's concern.

17.4 CRANFIELD STUDIES

ASLIB has carried out a project of comparison of four information retrieval systems at the College of Aeronautics, Cranfield. This work was named CRANFIELD PROJECT I. The four systems of study were UDC, Faceted Classification Scheme, Alphabetical Subject Headings, and Uniterm System of Indexing. Three indexers who had varied backgrounds in relation to proficiency in indexing and subject knowledge of aeronautics were appointed for the project. They indexed materials which consisted of batches of 100 technical reports and articles on various aspects of aeronautics. Thus with three indexers, four systems, five time periods, and three runs, a total of 18,000 items were indexed.

To carry out the test, 1400 questions were initially suggested by users from different organisations and these were scrutinised by a panel of three experts and finally, 400 were selected. These questions were then put to the four systems for retrieval, and a search was counted as successful if the "source document" was retrieved.

17.4.1 Findings of Cranfield I:

The following were the findings:

1. All four systems were of approximately equal effectiveness
2. No significant difference was found between the indexers or three runs.
3. It was sufficient to spend four minutes for indexing a particular document, since additional time gave no improvement in results.
4. Largest single factor leading to error was human factor. Failure to use the systems correctly or to search correctly was the cause for 50% of errors.
5. Indexing beyond the optimum level of exhaustivity did not help to increase the recall ratio, but adversely affected the precision.

17.4.2 Criticism of Cranfield I:

1. One of the main criticisms against this study was its artificiality without much relation to real life situation. Searching the system for known document is very much artificial. In a natural environment queries are not based on previous knowledge of availability of the documents in the store.

2. Since the test was carried out by questions largely, based on titles, efficiency of four minute indexing time holds good and cannot be generalised.
3. Another conclusion that the subject knowledge was not of much help for better indexing was also subjected to criticism. Indexers of the project were also subjected to criticism since they performed two roles as indexers and searchers.
4. One of the highlighted results of the study that all the four systems operated at the same level of performance was also confusing. Swanson argued that the condition of the test was not sufficient to distinguish the performance of the systems.

17.5 CRANFIELD II

Cranfield II study was conducted with 1400 research papers in the field of aerodynamics. The documents were indexed in three different ways:

1. Important concepts were isolated and recorded in natural language;
2. Single words in the concepts were listed; and
3. The concepts with a weighting (ranging from 1 to 3) were combined to represent subject contents of documents.

Five different types indexing systems with variations were used in this study. To conduct searches in all 33 indexing systems and 221 questions were used. The questions were generated by authors of research papers. Relevancy of each document was ascertained in respect of questions. Relevancy thus obtained was graded from 1 to 4 in the following manner:

1. complete answer to question;
2. high degree of relevance;
3. useful, providing general background of the work or dealing with a specific area; and
4. minimum interest, providing information like historical view point.

For the assessment, a single performance measure, called *normalised recall* was introduced. This is a ratio of cumulated recall ratio and number of search stages involving document output cut-off groups.

17.5.1 Findings of Cranfield II:

1. Inverse relation of Recall and Precision established
2. It was shown that best performance was obtained by the use of single term language.
3. Use of precision devices like portioning or intermixing was not as effective as the basic precision of coordination.
4. The test has shown that natural language, with the slight modifications of confounding synonyms and word forms, combined with single coordination, can give a reasonable performance.

17.5.2 Criticism of Cranfield II:

Unusual and unexpected results were found in these studies. Commenting on the results, Vickery observed that the indexes were made for the document set vocabulary and, naturally they did not reflect an ordinary operational situation. He also observed discrepancies in absolute number of postings of vocabulary, search term and search broadening, which did not reflect a real life situation. He also

pointed the lack of statistical tests, but agreed that findings of Cranfield II are valuable exploration of the retrieval process.

17.6 SUMMARY

Evaluation measures by themselves do not always provide sufficient information for operational decision making. Information retrieval system in general and their evaluation in particular have been influenced to such an extent by Cranfield studies, that the impact will be felt in years to come. The techniques such as Recall and Precision used in evaluating information retrieval today are introduced by Cleverdon's Cranfield projects more than fifty years ago. To conclude, evaluation is a form of enquiry where the end product is information. While information is power, evaluation is powerful.

17.7 TECHNICAL TERMS

IRS: Information Semantic Index

17.8 SUGGESTED READINGS

1. Cleverdon, C. W. The effect of variations in relevance assessments in comparative experimental tests of index languages. Cranfield, UK: Cranfield Institute of Technology. (Cranfield Library Report No. 3), 1970
2. Cleverdon, C. W., Mills, J. & Keen, E. M. Factors determining the performance of indexing systems. Cranfield, UK: Aslib Cranfield Research Project, College of Aeronautics. (Volume 1:Design; Volume 2: Results), 1966
3. Doyle, Lauren B. Information retrieval and processing. Los Angeles : John Wiley & Sons Inc., 1975.
4. Foskett, A.C. The subject approach to information, 5th ed. London : Clive Bingley, 1995.
5. Harter, S. P. & Hert, C. l. A. Evaluation of information retrieval systems: approaches, issues, and methods. *Annual Review of Information Science and Technology,* Vol. 32, No.3, p94-97, 1997.
6. Hildreth, C. R. Accounting for users' inflated assessments of on-line catalogue search performance and usefulness: an experimental study. *Information Research,* vol. 6 no.2, 2001.
7. Lancaster, F.W. Guidelines for the evaluation of information systems and services, Paris : UNESCO, 1978.

LESSON - 18

INFORMATION RETRIEVAL MODELS

AIMS AND OBJECTIVES

As an information user you need to keep up with the explosive growth of networked information. Information Retrieval Models like Boolean model, Statistical model and Linguistic and Knowledge –Based models are designed to support searching and retrieval of information from computer –to – computer communication systems.

The first model is often referred to as the "exact match" model; the latter ones as the "best match" models. After studying the lesson, you should in a position to:

- Explain what is information Retrieval Models with specific reference
- Discuss the meaning of data mining, its process, techniques and tools
- Describe the data warehousing and its implementation
- Explain web mining.

Structure

18.1 INTRODUCTION

A quick overview of the major textual retrieval methods were discussed in this lesson. First a general model of the information retrieval process was discussed. Then major retrieval methods were described in terms of their strengths and shortcomings.

An information retrieval system is a software programme that stores and manages information on documents. The system assists users in finding the information they need. It does not explicitly return information or answer questions. Instead, it informs on the existence and location of documents that might contain the desired information. Some suggested documents will, hopefully, satisfy the user's information need. These documents are called relevant documents. A perfect retrieval system would retrieve only the relevant documents and no irrelevant documents. However, perfect retrieval systems do not exist and will not exist, because search statements are necessarily incomplete and relevance depends on the subjective opinion of the user. In practice, two users may pose the same query to an information retrieval system and judge the relevance of the retrieved documents differently: Some users will like the results, others will not.

There are three basic processes an information retrieval system has to support:
- the representation of the content of the documents,
- the representation of the user's information need, and
- the comparison of the two representations.

The processes are visualised in the following figure. In the figure, squared boxes represent data and rounded boxes represent processes.

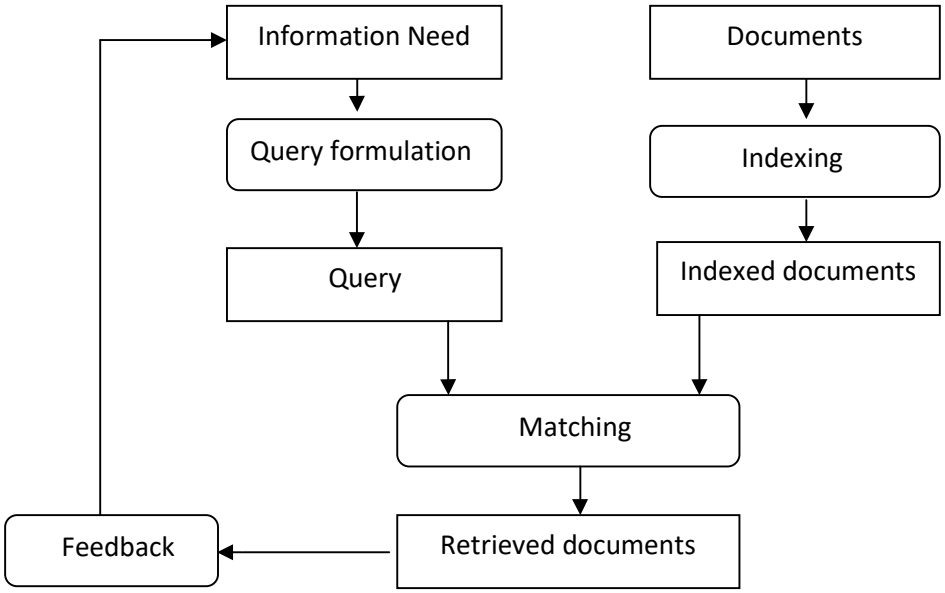


Figure 1: Information retrieval processes

Representing the documents is usually called the indexing process. The process takes place off-line, that is, the end user of the information retrieval system is not

directly involved. The indexing process results in a representation of the document. Often, full text retrieval systems use a rather trivial algorithm to derive the index representations, for instance an algorithm that identifies words in an English text and puts them to lower case. The indexing process may include the actual storage of the document in the system, but often documents are only stored partly, for instance only the title and the abstract, plus information about the actual location of the document. Users do not search just for fun; they have a need for information. The process of representing their information need is often referred to as the query formulation process. The resulting representation is the query. In a broad sense, query formulation might denote the complete interactive dialogue between system and user, leading not only to a suitable query but possibly also to the user better understanding his/her information need: This is denoted by the feedback process in Figure 1.

The comparison of the query against the document representations is called the matching process. The matching process usually results in a ranked list of documents. Users will walk down this document list in search of the information they need. Ranked retrieval will hopefully put the relevant documents towards the top of the ranked list, minimising the time the user has to invest in reading the documents. Simple but effective ranking algorithms use the frequency distribution of terms over documents, but also statistics over other information, such as the number of hyperlinks that point to the document. Ranking algorithms based on statistical approaches easily halve the time the user has to spend on reading documents. The theory behind ranking algorithms is a crucial part of information retrieval.

There are two good reasons for having models of information retrieval. The first is that models guide research and provide the means for academic discussion. The second reason is that models can serve as a blueprint to implement an actual retrieval system. Mathematical models are used in many scientific areas with the objective to understand and reason about some behaviour or phenomenon in the real world. A model of information retrieval predicts and explains what a user will find relevant given the user query. The correctness of the model's predictions can be tested in a controlled experiment. In order to do predictions and reach a better understanding of information retrieval, models should be firmly grounded in intuitions, metaphors and some branch of mathematics. Intuitions are important because they help to get a model accepted as reasonable by the research community. Metaphors are important because they help to explain the implications of a model to a bigger audience. Mathematics is essential to formalise a model, to ensure consistency, and to make sure that it can be implemented in a real system. As such, a model of information retrieval serves as a blueprint which is used to implement an actual information retrieval system. The following sections will describe different models of information retrieval.

18.2 GENERAL MODEL OF INFORMATION RETRIEVAL

The goal of **information retrieval** (IR) is to provide users with those documents that will satisfy their information need. We use the word "document" as a general term that could also include non-textual information, such as multimedia objects. Figure 4.1 provides a general overview of the information retrieval process, which has been adapted from Lancaster and Warner. Users have to formulate their information need in a form that can be understood by the retrieval mechanism. There are several steps

involved in this translation process that we will briefly discuss below. Likewise, the contents of large document collections need to be described in a form that allows the retrieval mechanism to identify the potentially relevant documents quickly. In both cases, information may be lost in the transformation process leading to a computer-usable representation. Hence, the matching process is inherently imperfect.

Information seeking is a form of problem solving. It proceeds according to the interaction among eight sub processes: problem recognition and acceptance, problem definition, search system selection, query formulation, query execution, examination of results (including relevance feedback), information extraction, and reflection/iteration/termination. To be able to perform effective searches, users have to develop the following expertise: knowledge about various sources of information, skills in defining search problems and applying search strategies, and competence in using electronic search tools.

Marchionini contends that some sort of spreadsheet is needed that supports users in the problem definition as well as other information seeking tasks. The ‘InfoCrystal’ is such a spreadsheet because it assists users in the formulation of their information needs and the exploration of the retrieved documents, using the visual interface that supports a "what-if" functionality. He further predicts that advances in computing power and speed, together with improved information retrieval procedures, will continue to blur the distinctions between problem articulation and examination of results. The ‘InfoCrystal’ is both a visual query language and a tool for visualizing retrieval results.

The information need can be understood as forming a pyramid, where only its peak is made visible by users in the form of a conceptual query. The conceptual query captures the key concepts and the relationships among them. It is the result of a conceptual analysis that operates on the information need, which may be well or vaguely defined in the user's mind. This analysis can be challenging, because users are faced with the general "vocabulary problem" as they are trying to translate their information need into a conceptual query. This problem refers to the fact that a single word can have more than one meaning, and, conversely, the same concept can be described by surprisingly many different words. Furnas, Landauer, Gomez and Dumais have shown that two people use the same main word to describe an object only 10 to 20% of the time. Further, the concepts used to represent the documents can be different from the concepts used by the user. The conceptual query can take the form of a natural language statement, a list of concepts that can have degrees of importance assigned to them, or it can be statement that coordinates the concepts using Boolean operators. Finally, the conceptual query has to be translated into a query surrogate that can be understood by the retrieval system.

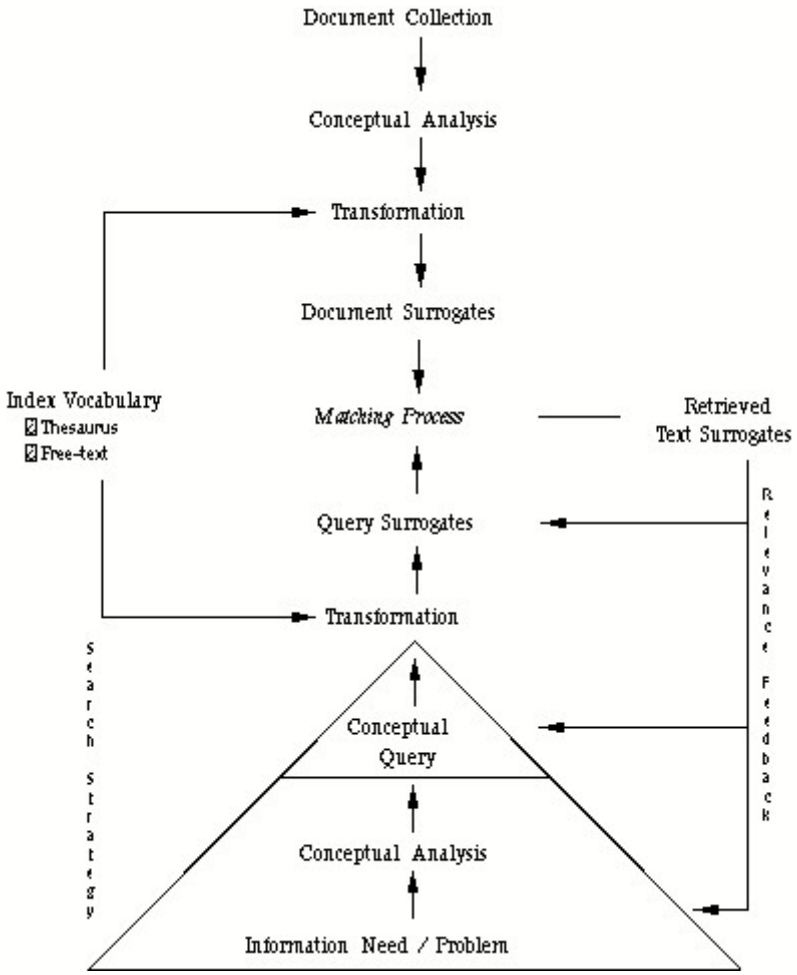


Figure 2

The above figure represents a general model of the information retrieval process, where both the user's information need and the document collection have to be translated into the form of surrogates to enable the matching process to be performed. This figure has been adapted from Lancaster and Warner.

Similarly, the meanings of documents need to be represented in the form of text surrogates that can be processed by computer. A typical surrogate can consist of a set of index terms or descriptors. The text surrogate can consist of multiple fields, such as the title, abstract, descriptor fields to capture the meaning of a document at different levels of resolution or focusing on different characteristic aspects of a document. Once the specified query has been executed by IR system, a user is presented with the retrieved document surrogates. Either the user is satisfied by the retrieved information or he will evaluate the retrieved documents and modify the query to initiate a further search. The process of query modification based on user evaluation of the retrieved documents is known as relevance feedback. Information retrieval is an inherently interactive process, and the users can change direction by modifying the query surrogate, the conceptual query or their understanding of their information need.

It is worth noting here the results, which have been obtained in studies investigating the information-seeking process, that describe information retrieval in terms of the cognitive and affective symptoms commonly experienced by a library user. The findings by Kuhlthau et al. indicates that thoughts about the information need become clearer and more focused as users move through the search process. Similarly, uncertainty, confusion, and frustration are nearly universal experiences in the early stages of the search process, and they decrease as the search process progresses and feelings of being confident, satisfied, sure and relieved increase. The studies also indicate that cognitive attributes may affect the search process. User's expectations of the information system and the search process may influence the way they approach searching and therefore affect the intellectual access to information.

Analytical search strategies require the formulation of specific, well-structured queries and a systematic, iterative search for information, whereas browsing involves the generation of broad query terms and a scanning of much larger sets of information in a relatively unstructured fashion. Campagnoni et al. have found in information retrieval studies in hypertext systems that the predominant search strategy is "browsing" rather than "analytical search". Many users, especially novices, are unwilling or unable to precisely formulate their search objectives, and browsing places less cognitive load on them. Furthermore, their research showed that search strategy is only one dimension of effective information retrieval; individual differences in visual skill appear to play an equally important role.

These two studies argue for information displays that provide a spatial overview of the data elements and that simultaneously provide rich visual cues about the content of the individual data elements. Such a representation is less likely to increase the anxiety that is a natural part of the early stages of the search process and it caters for a browsing interaction style, which is appropriate especially in the beginning, when many users are unable to precisely formulate their search objectives.

18.3 MAJOR INFORMATION RETRIEVAL MODELS

The following major models have been developed to retrieve information: the **Boolean** model, the **Statistical** model, which includes the vector space and the probabilistic retrieval model, and the **Linguistic and Knowledge-based** models. The first model is often referred to as the "exact match" model; the latter ones as the "best match" models.

Queries generally are less than perfect in two respects: First, they retrieve some irrelevant documents. Second, they do not retrieve all the relevant documents. The following two measures are usually used to evaluate the effectiveness of a retrieval method. The first one, called the **precision rate**, is equal to the proportion of the retrieved documents that are actually relevant. The second one, called the **recall rate**, is equal to the proportion of all relevant documents that are actually retrieved. If searchers want to raise precision, then they have to narrow their queries. If searchers want to raise recall, then they broaden their query. In general, there is an inverse relationship between precision and recall. Users need help to become knowledgeable in how to manage the precision and recall trade-off for their particular information need.

18.3.1. Standard Boolean

In the following Table summary of characteristics of the standard Boolean approach and its key advantages and disadvantages is given.

It has the following strengths:

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

The standard Boolean approach has the following shortcomings:

1) Users find it difficult to construct effective Boolean queries for several reasons. Users are using the natural language terms AND, OR, NOT that have a different meaning when used in a query. Thus, users will make errors when they form a Boolean query, because they resort to their knowledge of English. For example, in ordinary conversation a noun phrase of the form "A and B" usually refers to more entities than "A" alone would. But when used in the context of information retrieval it refers to fewer documents than would be retrieved by "A" alone. Hence, one of the common mistakes made by users is to substitute the AND logical operator for the OR logical operator when translating an English sentence to a Boolean query. Furthermore, to form complex queries, users must be familiar with the rules of precedence and the use of parentheses. Novice users have difficulty using parentheses, especially nested parentheses. Finally, users are overwhelmed by the multitude of ways a query can be structured or modified, because of the combinatorial explosion of feasible queries as the number of concepts increases. In particular, users have difficulty identifying and applying the different strategies that are available for narrowing or broadening a Boolean query.

	Standard Boolean
Goal	• Capture conceptual structure and contextual information
Methods	• Coordination: AND, OR, NOT • Proximity • Fields • Stemming / Truncation
(+)	• Easy to implement • Computationally efficient => all the major on-line databases use it • Expressiveness and Clarity Synonym specifications (OR-clauses) and phrases (AND-clauses).
(-)	• Difficult to construct Boolean queries. • All or nothing AND too severe, and OR does not differentiate enough. • Difficult to control output: Null output <-> Overload. • No ranking • No weighting of index or query terms • No uncertainty measure

Table 1

2) Only documents that satisfy a query exactly are retrieved. On the one hand, the AND operator is too severe because it does not distinguish between the case when none of the concepts are satisfied and the case where all except one are satisfied. Hence, no or very few documents are retrieved when more than three and four criteria are combined with the Boolean operator AND (referred to as the Null Output problem). On the other hand, the OR operator does not reflect how many concepts have been satisfied. Hence, often too many documents are retrieved (the Output Overload problem).

3) It is difficult to control the number of retrieved documents. Users are often faced with the null-output or the information overload problem and they are at loss of how to modify the query to retrieve the reasonable number documents.

4) The traditional Boolean approach does not provide a relevance ranking of the retrieved documents, although modern Boolean approaches can make use of the degree of coordination, field level and degree of stemming present to rank them.

5) It does not represent the degree of uncertainty or error due the vocabulary problem.

18.3.1.1 Narrowing and Broadening Techniques

As mentioned earlier, a Boolean query can be described in terms of the following four operations: degree and type of coordination, proximity constraints, field specifications and degree of stemming as expressed in terms of word/string specifications. If users want to (re)formulate a Boolean query then they need to make informed choices along these four dimensions to create a query that is sufficiently broad or narrow depending on their information needs. Most narrowing techniques lower recall as well as raise

precision and most broadening techniques lower precision as well as raise recall. Any query can be reformulated to achieve the desired precision or recall characteristics, but generally it is difficult to achieve both. Each of the four kinds of operations in the query formulation has particular operators, some of which tend to have a narrowing or broadening effect. For each operator with a narrowing effect, there are one or more inverse operators with a broadening effect. Hence, users require help to gain an understanding of how changes along these four dimensions will affect the broadness or narrowness of a query.

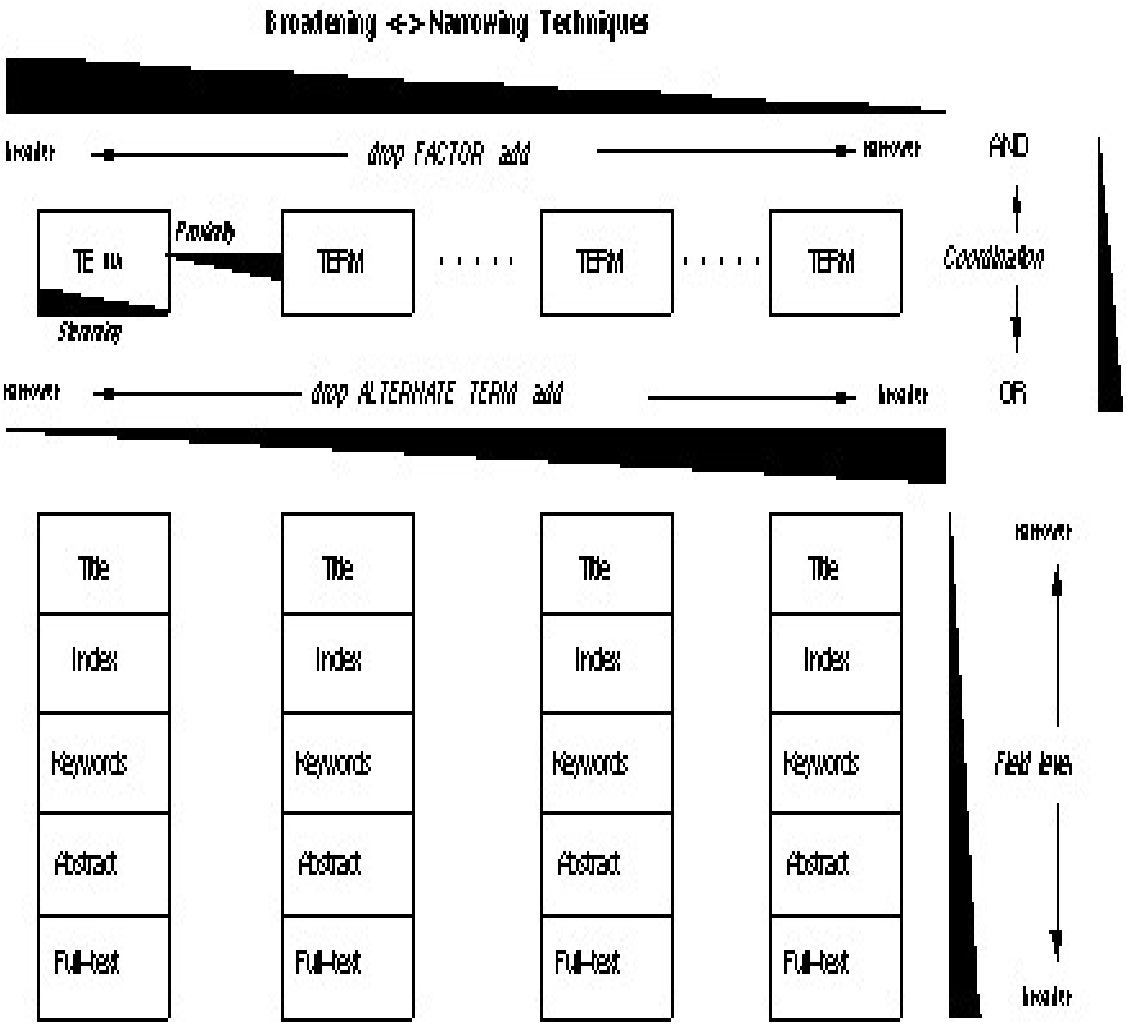


Figure 2:

The above figure captures how coordination, proximity, field level and stemming affect the broadness or narrowness of a Boolean query. By moving in the direction in which the wedges are expanding the query is broadened.

It shows how the four dimensions affect the broadness or narrowness of a query:

1) *Coordination*: the different Boolean operators AND, OR and NOT have the following effects when used to add a further concept to a query:
   a) the AND operator narrows a query;

b) the OR broadens it;
c) the effect of the NOT depends on whether it is combined with an AND or OR operator. Typically, in searching textual databases, the NOT is connected to the AND, in which case it has a narrowing effect like the AND operator.

2) *Proximity*: The closer together two terms have to appear in a document, the more narrow and precise the query. The most stringent proximity constraint requires the two terms to be adjacent.
3) *Field level*: current document records have fields associated with them, such as the "Title", "Index", "Abstract" or "Full-text" field:
a) the more fields that are searched, the broader the query;
b) the individual fields have varying degrees of precision associated with them, where the "title" field is the most specific and the "full-text" field is the most general.
4) *Stemming*: The shorter the prefix that is used in truncation-based searching, the broader the query. By reducing a term to its morphological stem and using it as a prefix, users can retrieve many terms that are conceptually related to the original term.

Using the above figure, one can easily read off how to broaden query. It is just need to move in the direction in which the wedges are expanding: One can use the OR operator (rather than the AND), impose no proximity constraints, search over all fields and apply a great deal of stemming. Similarly, one can formulate a very narrow query by moving in the direction in which the wedges are contracting: one can use the AND operator (rather than the OR), impose proximity constraints, restrict the search to the title field and perform exact rather than truncated word matches.

18.3.1.2 Smart Boolean

There have been attempts to help users overcome some of the disadvantages of the traditional Boolean discussed above. One such method is called *Smart Boolean*, developed by Marcus. It tries to help users construct and modify a Boolean query as well as make better choices along the four dimensions that characterize a Boolean query. Smart Boolean method is a good example that illustrates some of the possible ways to make Boolean retrieval more user-friendly and effective. Following Table provides a summary of the key features of the Smart Boolean approach and its advantages and disadvantages.

	Smart Boolean
Goal	- structure search (re-)formulation process - Use structural and contextual knowledge-bases and clarity of Boolean expressions
Methods	- Natural language statement is automatically translated into Boolean topic representation - Boolean topic Representation: ANDs of ORs of concepts keyword/stem, all fields Conceptual info - Coordination and add/drop factor Contextual info - Proximity Structural info - Field levels Synonym or word relationships - Stemming/Truncation overlap = All this information can be used to rank documents - Techniques to Broaden and Narrow query

(+)	- No need for Boolean operators = convert operator–free statement into ANDs of ORs - Assist user in query (Re)formulation: by asking users targeted questions to automatically modify the query - "Why irrelevant?" - Activities narrowing methods - "Broaden by Dropping factors" to estimate recall.
(-)	- How to visualize - Conceptual query representation (BTR) - Query modification techniques and their effects - Structured relevance feedback

Users start by specifying a natural language statement that is automatically translated into a Boolean Topic representation that consists of a list of factors or concepts, which are automatically coordinated using the AND operator. If the user at the initial stage can or wants to include synonyms, then they are coordinated using the OR operator. Hence, the Boolean Topic representation connects the different factors using the AND operator, where the factors can consist of single terms or several synonyms connected by the OR operator. One of the goals of the Smart Boolean approach is to make use of the structural knowledge contained in the text surrogates, where the different fields represent contexts of useful information. Further, the Smart Boolean approach wants to use the fact that related concepts can share a common stem. For example, the concepts "computers" and "computing" have the common stem comput*.

The initial strategy of the Smart Boolean approach is to start out with the broadest possible query within the constraints of how the factors and their synonyms have been coordinated. Hence, it modifies the Boolean Topic representation into the query surrogate by using only the stems of the concepts and searches for them over all the fields. Once the query surrogate has been performed, users are guided in the process of evaluating the retrieved document surrogates. They choose from a list of reasons to indicate why they consider certain documents as relevant. Similarly, they can indicate why other documents are not relevant by interacting with a list of possible reasons. This user feedback is used by the Smart Boolean system to automatically modify the Boolean Topic representation or the query surrogate, whatever is more appropriate. The Smart Boolean approach offers a rich set of strategies for modifying a query based on the received relevance feedback or the expressed need to narrow or broaden the query.

18.3.1.3 Extended Boolean Models

Several methods have been developed to extend the Boolean model to address the following issues:

1) The Boolean operators are too strict and ways need to be found to soften them.
2) The standard Boolean approach has no provision for ranking. The Smart Boolean approach and the methods described in this section provide users with relevance ranking.
3) The Boolean model does not support the assignment of weights to the query or document terms.

The *P-norm* and the *Fuzzy Logic* approaches that extend the Boolean model to address the above issues are discussed below:

	Extended Boolean Models
Goal	- Less strict Boolean operators - Ranked output
Methods	- Fuzzy logic [OR = max], [AND = min] and [NOT = 1- max] (-) Lack of sensitivity of min and max; min (0.2, 0.8) = min (0.2, 0.3)

Table 3:

The above table summarizes the defining characteristics of the Extended Boolean approach and enumerate its key advantages and disadvantages.

The **P-norm** method developed by Fox allows query and document terms to have weights, which have been computed by using term frequency statistics with the proper normalization procedures. These normalized weights can be used to rank the documents in the order of decreasing distance from the point (0, 0, ..., 0) for an OR query, and in order of increasing distance from the point (1, 1, ... , 1) for an AND query. Further, the Boolean operators have a coefficient 'P' associated with them to indicate the degree of strictness of the operator (from 1for least strict to infinity for most strict, i.e., the Boolean case). The P-norm uses a distance-based measure and the coefficient 'P' determines the degree of exponentiation to be used. The exponentiation is an expensive computation, especially for P-values greater than one.

In **Fuzzy Set theory**, an element has a varying degree of membership to a set instead of the traditional binary membership choice. The weight of an index term for a given document reflects the degree to which this term describes the content of a document. Hence, this weight reflects the degree of membership of the document in the fuzzy set associated with the term in question. The degree of membership for union and intersection of two fuzzy sets is equal to the maximum and minimum, respectively, of the degrees of membership of the elements of the two sets. In the "Mixed Min and Max" model developed by Fox and Sharat, the Boolean operators are softened by considering the query-document similarity to be a linear combination of the min and max weights of the documents.

18.3.2 Statistical Model

The *vector space* and *probabilistic* models are the two major examples of the statistical retrieval approach. Both models use statistical information in the form of term frequencies to determine the relevance of documents with respect to a query. Although they differ in the way they use the term frequencies, both produce as their output a list of documents ranked by their estimated relevance. The statistical retrieval models address some of the problems of Boolean retrieval methods, but they have disadvantages of their own. Following Table provides summary of the key features of the vector space and probabilistic approaches. This lesson also describe *Latent*

Semantic Indexing and *clustering* approaches that are based on statistical retrieval approaches, but their objective is to respond to what the user's query did not say, could not say, but somehow made manifest.

18.3.2.1 Vector Space Model

The **vector space model** represents the documents and queries as vectors in a multidimensional space, whose dimensions are the terms used to build an index to represent the documents. The creation of an index involves lexical scanning to identify the significant terms, where morphological analysis reduces different word forms to common "stems", and the occurrence of those stems is computed. Query and document surrogates are compared by comparing their vectors, using, for example, the cosine similarity measure. In this model, the terms of a query surrogate can be weighted to take into account their importance, and they are computed by using the statistical distributions of the terms in the collection and in the documents. The vector space model can assign a high ranking score to a document that contains only a few of the query terms if these terms occur infrequently in the collection but frequently in the document. The vector space model makes the following 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. While it is a reasonable first approximation, the assumption that words that are pair wise independent is not realistic.

18.3.2.2 Probabilistic Model

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

Turtle and Croft developed a state-of-art system that uses Bayesian inference networks to rank documents by using multiple sources of evidence to compute the conditional probability P (Information need / document) that information need is satisfied by a given document. An inference network consists of a directed acyclic dependency graph, where edges represent conditional dependency or causal relations between propositions represented by the nodes. The inference network consists of a document network, a concept representation network that represents indexing vocabulary, and a query network representing the information need. The concept representation network is the interface between documents and queries. To compute the rank of a document, the inference network is instantiated and the resulting probabilities are propagated through the network to derive a probability associated with the node representing the information need. These probabilities are used to rank documents.

The statistical approaches have the following advantages:

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

However, the statistical approaches have the following disadvantages:

1. They have a limited expressive power. For example, the NOT operation can't be represented because only positive weights are used.
2. The statistical approach lacks the structure to express important linguistic features such as phrases. Proximity constraints are also difficult to express, a feature that is of great use for experienced searchers.
3. The computation of the relevance scores can be computationally expensive.
4. A ranked linear list provides users with a limited view of the information space and it does not directly suggest how to modify a query if the need arises.
5. The queries have to contain a large number of words to improve the retrieval performance. As is the case for the Boolean approach, users are faced with the problem of having to choose the appropriate words that are also used in the relevant documents.

Following Table summarizes the advantages and disadvantages that are specific to the vector space and probabilistic model, respectively. This table also shows the formulas that are commonly used to compute the term weights. The two central quantities used are the inverse term frequency in a collection (*idf*), and the frequencies of a term 'i' in a document 'j' (*freq(i,j)*). In the probabilistic model, the weight computation also considers how often a term appears in the relevant and irrelevant documents, but this presupposes that the relevant documents are known or that these frequencies can be reliably estimated.

Statistical	Vector Space	Probabilistic
Motivation	Simplify query formulation Ability to control output	Address uncertainty in query representations
Goal	Rank the output based on Similarity	Probability of Relevance
Methods	Cosine measure	Use of different models
Source	**Query Term Statistics** Vector-Space: • similarity(Q,D) = Σ (w_{iq} x w_{ij}) / "normalizer" where w_{iq} = (0.5 + 0.5 $freq_{iq}$ / $maxfreq_q$) x **idf**(i) w_{ij} = $freq_{ij}$ x **idf**(i) • inverse term freq. in collection **idf**(i) = $\log_2$ (N-n(i)) / n(i). Probabilistic: • term weight = log [(r_t / R-r_t) / ((n_t -r_t) / ((N-n_t) - (R-r_t)))] ="(hits / misses) / (false alarms/correct misses)" • similarity $_{jk}$ = Σ (C + **idf**(i)) x **tf**(i,j) where **tf**(i,j) = K + (1-K) (freq(i,j) / maxfreq(j)).	
Issues	• How to express NOT ? Proximity searches ? • Limited expressive power • Computationally intensive • Assumes that terms are independent.	• Estimation of needed probabilities • Prior knowledge needed. • Independence assumption • Boolean relations lost. • Which model is best ?
	• Lack of structure to represent important linguistic features • How to better visualize the retrieved set ?	

18.3.2.3 Latent Semantic Indexing

Several statistical and Artificial Intelligence techniques have been used in association with domain semantics to extend the vector space model to help overcome some of the retrieval problems described above, such as the "dependence problem" or the "vocabulary problem". One such method is **Latent Semantic Indexing** (LSI). In LSI the associations among terms and documents are calculated and exploited in the retrieval process. The assumption is that there is some "latent" structure in the pattern of word usage across documents and that statistical techniques can be used to estimate this latent structure. An advantage of this approach is that queries can retrieve documents even if they have no words in common. The LSI technique captures deeper associative structure than simple term-to-term correlations and is completely automatic. The only difference between LSI and vector space methods is that LSI represents terms and documents in a reduced dimensional space of the derived indexing dimensions. As with

the vector space method, differential term weighting and relevance feedback can improve LSI performance substantially.

Foltz and Dumais compared four retrieval methods that are based on the vector-space model. The four methods were the result of crossing two factors, the first factor being whether the retrieval method used Latent Semantic Indexing or keyword matching, and the second factor being whether the profile was based on words or phrases provided by the user (Word profile), or documents that the user had previously rated as relevant (Document profile). The LSI match-document profile method proved to be the most successful of the four methods. This method combines the advantages of both LSI and the document profile. The document profile provides a simple, but effective, representation of the user's interests. Indicating just a few documents that are of interest is as effective as generating a long list of words and phrases that describe one's interest. Document profiles have an added advantage over word profiles: users can just indicate documents they find relevant without having to generate a description of their interests.

18.3.3 Linguistic and Knowledge-based Approaches

In the simplest form of automatic text retrieval, users enter a string of keywords that are used to search the inverted indexes of the document keywords. This approach retrieves documents based solely on the presence or absence of exact single word strings as specified by the logical representation of the query. Clearly this approach will miss many relevant documents because it does not capture the complete or deep meaning of the user's query. The Smart Boolean approach and the statistical retrieval approaches, each in their specific way, try to address this problem. Linguistic and knowledge-based approaches have also been developed to address this problem by performing a morphological, syntactic and semantic analysis to retrieve documents more effectively. In a morphological analysis, roots and affixes are analyzed to determine the part of speech (noun, verb, adjective etc.) of the words. Next complete phrases have to be parsed using some form of syntactic analysis. Finally, the linguistic methods have to resolve word ambiguities and/or generate relevant synonyms or quasi-synonyms based on the semantic relationships between words. The development of a sophisticated linguistic retrieval system is difficult and it requires complex knowledge bases of semantic information and retrieval heuristics. Hence these systems often require techniques that are commonly referred to as artificial intelligence or expert systems techniques.

18.3.3.1 DR-LINK Retrieval System

DR-LINK system developed by Liddy et al., is an exemplary linguistic retrieval system. DR-LINK is based on the principle that retrieval should take place at the conceptual level and not at the word level. Liddy et al. attempt to retrieve documents on the basis of what people mean in their query and not just what they say in their query. DR-LINK system employs sophisticated, linguistic text processing techniques to capture the conceptual information in documents. Liddy et al. have developed a modular system that represents and matches text at the lexical, syntactic, semantic, and the discourse levels of language. Some of the modules that have been incorporated are: The Text Structure is based on discourse linguistic theory that suggests that texts of a particular type have a predictable structure which serves as an indication where certain

information can be found. The Subject Field Coder uses an established semantic coding scheme from a machine-readable dictionary to tag each word with its disambiguated subject code (e.g., computer science, economics) and to then produce a fixed-length, subject-based vector representation of the document and the query. The Proper Noun Interpreter uses a variety of processing heuristics and knowledge bases to produce: a canonical representation of each proper noun; a classification of each proper noun into thirty-seven categories; and an expansion of group nouns into their constituent proper noun members. The Complex Nominal Phrase provides means for precise matching of complex semantic constructs when expressed as either adjacent nouns or a non-predicating adjective and noun pair. Finally, The Natural Language Query Constructor takes as input a natural language query and produces a formal query that reflects the appropriate logical combination of text structure, proper noun, and complex nominal requirements of the user's information need. This module interprets a query into pattern-action rules that translate each sentence into a first-order logic assertion, reflecting the Boolean-like requirements of queries.

Linguistic Level	Boolean Retrieval	Statistical	Linguistic and Knowledge Based
Lexical	Stop word list	Stop word list	Lexical
Morphological	Truncation symbol	Stemming	Morphological analysis
Syntactic	Proximity operators	Statistical Phrases	Grammatical Phrases
Semantic	Thesaurus	Clusters of Co-occurring words	Network of words/phrases in semantic relationships

Table 5:

The above table characterizes the major retrieval methods in terms of how deal with lexical, morphological, syntactic and semantic issues.

To summarize, the DR-LINK retrieval system represents content at the conceptual level rather than at the word level to reflect the multiple levels of human language comprehension. The text representation combines the lexical, syntactic, semantic, and discourse levels of understanding to predict the relevance of a document. DR-LINK accepts natural language statements, which it translates into a precise Boolean representation of the user's relevance requirements. It also produces a summary-level, semantic vector representations of queries and documents to provide a ranking of the documents.

18.4 SUMMARY

There is a growing discrepancy between the retrieval approach used by existing commercial retrieval systems and the approaches investigated and promoted by a large segment of the information retrieval research community. The former is based on the Boolean or Exact Matching retrieval model, whereas the latter ones subscribe to statistical and linguistic approaches, also referred to as the Partial Matching approaches. First, the major criticism levelled against the Boolean approach is that its queries are difficult to formulate. Second, the Boolean approach makes it possible to represent structural and contextual information that would be very difficult to

represent using the statistical approaches. Third, the Partial Matching approaches provide users with a ranked output, but these ranked lists obscure valuable information. Fourth, recent retrieval experiments have shown that the Exact and partial matching approaches are complementary and should therefore be combined.

Table 6:

Key Problems	Possible Solutions
Selection of search vocabulary	- Thesaurus - Latent Semantic Indexing
Search strategy (re)formulation	- Smart Boolean - Statistical & Linguistic Approaches - Thesaurus - Graphical Interfaces
Information Overload	- Ranking - Clustering - Visualization

The above Table summarizes some of the key problems in the field of information retrieval and possible solutions to them. This lesson explains: 1) how visualization can offer ways to address these problems; 2) how to formulate and modify a query; 3) how to deal with large sets of retrieved documents, commonly referred to as the information overload problem.

There is no such thing as a dominating model or theory of information retrieval, unlike the situation in, for instance, the area of databases where the relational model is the dominating database model. In information retrieval some models work for some applications, where as others work for other applications. For example vector space model is well suited for similarity search and relevance feedback. The probabilistic retrieval model is a good choice if relevant and non-relevant documents are available.

18.5 Technical Terms

LRS: Latent Semantic Index

18.6 Suggested Readings

1) 1. Fuhr, N. Models for retrieval with probabilistic indexing. *Information processing and management* 25 (1), 1989, p55–72.
2) 2. Robertson, S. E. The Probability Ranking Principle in IR. *Journal of Documentation* Vol.33, 1977. p294-304.
3) 3. Robertson, S. E., C. J. van Rijsbergen, and M. F. Porter. Probabilistic models of indexing and searching. In R. Oddy et al. (Ed.), Information Retrieval Research, pp. 35–56. New York : Butterworths, 1981
4) 4. Salton, G.; Buckley, C. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science* 41(4), 1990.p288–297.
6. Salton, G. (ed). The SMART Retrieval System - Experiments in Automatic Document Processing. Englewood, Cliffs, New Jersey: Prentice Hall, 1971.
7. Salton, G. and M. McGill. Introduction to Modern Information Retrieval. New Jersey : McGraw-Hill, 1983.
8. Ullman, Jeffrey D. Principles of Database and Knowledge-Base Systems, volume I. Rockville (Md.) : Computer Science Press, 1988.