# NETAJI SUBHAS OPEN UNIVERSITY

STUDY MATERIAL

## MLIS

Paper - 03 (Eng.)

Information, Processing and Retrieval

## Library and Information Science

# PREFACE

In the curricular structure introduced by this University for students of Post-Graduate degree programme, the opportunity to pursue Post-Graduate course in any subject introduced by this University is equally available to all learners. Instead of being guided by any presumption about ability level, it would perhaps stand to reason if receptivity of a learner is judged in the course of the learning process. That would be entirely in keeping with the objectives of open education which does not believe in artificial differentiation.

Keeping this in view, study materials of the Post-Graduate level in different subjects are being prepared on the basis of a well laid-out syllabus. The course structure combines the best elements in the approved syllabi of Central and State Universities in respective subjects. It has been so designed as to be upgradable with the addition of new information as well as results of fresh thinking and analyses.

The accepted methodology of distance education has been followed in the preparation of these study materials. Co-operation in every form of experienced scholars is indispensable for a work of this kind. We, therefore, owe an enormous debt of gratitude to everyone whose tireless efforts went into the writing, editing and devising of a proper lay-out of the materials. Practically speaking, their role amounts to an involvement in invisible teaching. For, whoever makes use of these study materials would virtually derive the benefit of learning under their collective care without each being seen by the other.

The more a learner would seriously pursue these study materials the easier it will be for him or her to reach out to larger horizons of a subject. Care has also been taken to make the language lucid and presentation attractive so that they may be rated as quality self-learning materials. If anything remains still obscure or difficult to follow, arrangements are there to come to terms with them through the counselling sessions regularly available at the network of study centres set up by the University.

Needless to add, a great part of these efforts is still experimental—in fact, pioneering in certain areas. Naturally, there is every possibility of some lapse or deficiency here and there. However, these do admit of rectification and further improvement in due course. On the whole, therefore, these study materials are expected to evoke wider appreciation the more they receive serious attention of all concerned.

**Professor (Dr.) Subha Sankar Sarkar**
Vice-Chancellor

First Revised Edition : April, 2016

Printed in accordance with the regulations and financial assistance of the Distance Education Bureau of the University Grants Commission.

# Post-Graduate : Library and Information Science [MLIS]

## Paper - III

## Information processing and Retrieval

**Course Writing**

Prof. Bhubaneswar Chakrabarti
Dr. Swapna Banerjee (Unit 7)

**Editing**

Dr. R. Ramachandran

Netaji Subhas Open University

MLIS - 03
Information processing and Retrieval

## Module 1: Intellectual Organisation of Information



## Module 2 : Bibliographic Description and Subject Indexing



## Module 3 : Indexing Languages and Vocabulary Control



## Module 4 : Information Retrieval

# Unit 1 ❐ Intellectual Organisation : An Overview

**Structure**



## 1.0 Objectives

After reading this unit you will be able to—

* understand the basic drive in mankind to organise
* why do we need to organise ?
* who do we need to organise information?
* how is the organisation of information approached in different environments,

## 1.1 Introduction

Human intellect produces information which is used by others. Transfer of information and absorption of the same by others depend on the intellectual organisation of information. There seems to be a basic drive in us to organise images into categories such as "faces' or 'foods'. Children do a lot of organizing during play. With some of us the need is stronger than the others. Those who act on the maxim, states, "A place for everything and everthing in its place" will not be able to operate until the work surface is cleared and every stray object has been put in its place. Such a person has to be "organised" before beginning a new project. "But even those whose world spaces appear to be chaotic have some organisation in their heads". Such persons usually have some idea or perhaps certain knowledge, of what is in the various collections of "stuff Regardless of one's personal style, however human learning is based on-the ability to observe, to organise to assimilate, to judge correlation application in contexts to data, information, knowledge and wisdom respectively.

## 1.2 Need to organise

We need to organise because we need to retrieve. If kitchens are well organised cooking equipment will be easily accessible and foodstuffs and spices can used as needed. If workplaces are organised so that appropriate records are retrievable. Learning processes are organised so, that relationships among ideas can be used to the learner in recalling the learned material.

## 1.3 Organisation of Recorded Information

This study material addresses the history, theory and practice of the organisation of recorded information. Other means are required to organise information that has only spoken, heard or thought about. A certain amount of disagreement exists among us as to whether we organise, retrieve, and make use of information or knowledge. According to several dictionaris, knowledge exists in the mind of an individual who has studied a matter, understands it and perhaps has added to it through research and other

means. Dictionaries also indicate that information is the communication or reception of knowledge. That communication occurs in great part through the recording of the knowledge in some fashion. People write, speak, compose, scudpt and in many other ways attempt to communicate their knowledge to others. Recorded information includes much more than text. Therefore, terms such as information bearing activity or information package are used. Ronald Haggler has indicated six functions of bibliographic control. His listing emphasizes upon the work of librarians. However, the list with altered wording, reflects the major activities involved in all of organisation of information—

(a) Identifying the existence of all types of information-bearing entities as they are made available—

Most publishers create red catalogues listing of their products along with abstracts for them. Reference tools such as **books in Print** are products of their activity.

(b) Identifying the works contained within those information-bearing entities or as parts of them—

A collection of short stories or a grouping of artistic works may be considered to be an information-bearing entity as a whole.

(c) Systematically pulling together these information-bearing entities into collections in libraries, archives, museums, Internet communication files, and other such depositories. The activity of creating collections traditionally has been thought of as the province of institutions such as libraries, archives and museums. There are also personal collections, office collections, university departmental collection.

Now that it is easy to bring these collections known publicly, lists arc being provided at web sites.

(d) Producing lists of these information-bearing entities prepared according to standard rules for citation—

Lists created in this activity include bibliographies, indexes, library

catalogues, archival finding aids, and museum registers. Such lists may be in print or electronic form.

(e) Providing name, title, subject and other useful access to these information-bearing entities—

This is the activity that adds the most value to the usefulness and retrieval potential of a collection. Keyword access can be provided more or less automatically. More satisfactory retrieval comes from being able to search for names, titles and controlled vocabulary that have been created under authority control If a system uses controlled vocabulary, a search for a word with more than one meaning will allow differentiation among different meanings and will direct one to broader, narrower and related terms. It will also bring together under one term all the synonymous terms that may be used to express a concept. Therefore, a major part of organizing information is designing systems for searching that will allow information seekers to find easily what they need and want.

(f) Providing the means of locating each information-bearing entity or a copy of it—

The catalogues or other lists created in century old institutions give information on the physical location of the entity. Bibliographic utilities (for example, OCLC, RLIN, WLN) allow one to find out which locations physically own a particular item. Many library, museum and archival catalogues are available on the internet. Traditionally, bibliographies and indexes have not given location information. Bibliographies list what exists somewhere but seldom tells where. Indexes give larger work in which smaller work being listed can be found, (for example, in which journal an article can be found), but do not indicate the physical location of the larger work. But electronic resources found on the internet is becoming more common to give the location (for example URL) in any listing that includes the electronic resource. However, the stability of URLs always raises the question.

## 1.4 Organisation of Information : Approaches in different Environments

There are many environments in which there is a desire to organise information so that it will be retrievable for various purposes. There are libraries of all types archives, museums, art galleries and the Internet

### 1.4.1 Libraries

Collections in libraries are created through the process called collection development. Collections are developed most often in three ways : (a) librarians learn about existence of new works through reviews, publisher's announcements, requests from users of the library, etc. and then order appropriate materials, (b) gifts are given to the library;.(c) approval plans, worked out with one or more vendors, bring in new items according to preselected profiles.

When new materials arrive for addition to the collection, physical entities have to be arranged in some fashion. They may be placed in alphabetical order (lor example, fiction and biography). Most, however, are arranged by classification.

Classification of materials is part of the process of cataloguing which is the first activity following the receipt of the items. Cataloguing of individual items involves creating a description of the physical item: choosing certain names and titles to serve as access points for getting to the description in the catalogue; doing authority work on these names and titles; doing subject analysis of the content of the work in the item; choosing subject heading and classification numbers to represent the subject analysis; and creating call numbers. All records thus created are coded with the MARC so that can be displayed in online systems.

Another major pan of the organisation process in libraries is found in the reference process. In the reference process the success of the organisation is tested.

### 1.4.2 Archives

Archives usually consist of unique items. Archival materials are arranged and described in groups. Each archive choses its own to organise the

information, particularly regarding level of contol and depth of description.

Descriptions of archival materials can take one or more different forms. An accession record summarizes information about the source of the collection, gives the circumstances of its acquisition and briefly describes the physical data and the contents of a collection. A finding aid gives a detailed conterits note of the historical and organizational context of the collection and continues by describing its context, perhaps providing an inventory outlining what is in each box. Archival materials generally are in closed stacks, accessible only to staff.

### 1.4.3 Museums and Art Galleries

In museums other than natural history, items are registered after being accessioned. Registration is a process much like cataloguing in libraries and archives. The register establishes the organizational control over the art works and artifacts. In museums, as in archives, provenance is important information and is essential in determining the name of the object.

As with archival materials the museum/art gallery collections are accessible only to staff. Much of the collection is stored behind scenes which only source of it is on display at any one time. Behind the scenes, items are numbered in a way so that it can be retrieved as needed. Persons responsible for the exhibits must make heavy use of the system of organizational control.

### 1.4.4 Internet

Several different approaches are being taken in the attempt to organise the Internet. Librarians have attempted to use traditional means for the organisation. Librarians are compiling bibliographies of websites. Librarians have been part of the team of people who have been working on a metadata standard called Dublin Core. Digital libraries are also being developed.

In fact, much work on organizing internet has been done by persons other than librarians. Search engines are being developed by computer and programming specialists. There is software that automatically classifies and indexes electronic documents, but automated tools categorize information differently than people do. The search site called yahoo! classifies by broad subject are as using human indexers. This approach has been popular, although

not completely successful as a classification. Also a research project at OCLC is improving an approach to automatic classification using the Dewey Decimal Classification. Although some believe that organizing the Internet will be impossible, the parts of it that are important for retrieval and for posterity will be brought under the organizational control. It is human nature, and the principles learned over centuries of organizing print information can be used to speed the process of organizing electronic resources.

## 1.5 Summary

In this unit we have discussed the basic needs to organise, have defined organisation of information and have looked at an overview of the different organizing environments.

## 1.6 Exercise

1. Why do we need to organise information?
2. What is organisation of recorded information?
3. Give an overview of the organisation of information approached in different environments.

## 1.7 References and Further Reading


1. Higgler, Ronald : The Bibliographic record and information technology. 3rd ed. American Library Association, 1997.
2. Stoll, Clifford : Silicon snake oil : second thoughts on the information highway. Doubleday, 1995.
3. Taylor, Arline G : The organisation of information. Libraries Unlimited, 1999.
4. Taylor, Arlene G : "The Information universe : will we have chaos or control?" **American Libraries** 1994, 26(7), 629—32.

# Unit 2 ❐ Classification Systems 1 : General Systems

**Structure**

## 2.0 Objectives

This unit will give you an idea about the major general classification schemes being used today in libraries all over the globe for the organisation of knowledge. You will understand the characteristic features and limitations of each scheme and will be in a position to choose a suitable scheme for your library.

## 2.1 Introduction

IR cannot afford to ignore the concept of a general classification scheme. The general scheme alone can provide a bird's eye view of the whole field of knowledge, offering a comprehensive context within which searches in a very large store can be framed. If the first step in establishing what are loosely called its main classes were to be the division of the whole field of knowledge by applying explicit characteristics of division, the only passible contenders would be of the nature of fundamental categories. The earliest and best—known set of such categories is seen in those advanced by Aristotle. Some of these are ostensibly feasible as constituting the initial divisions of the whole field of knowledge, e.g., substance, quantity, quality place, time, and action. Such a first step has not been attempted by any of the general classification schemes produced since Dewey's annus mirabilis in 1876,

although-something,Like it was attempted by James Duff Brown (Subject classification) with its quadruple division into Matter, Life, Mind, Record. Brown's scheme was notorious in its day for its subordination of music to sonics in physics—an example of its attempt to ignore disciplines as a primary level of division.

The particular notion of the fundamental forms of knowledge that underpin main classes has received significant attention by Language in 1976, who has drawn extensible on the work of number of philosophers. Of particular significance is the distinction Langride draws between the forms of knowledge on the one hand and the objects of knowledge on the other. The order in which main classes might appear became a particular focus of attention in the work of Bliss. The Colon classification pioneered acted classification but its main class in quite conventional. $Bc_2$ is now virtually a new general classification and constitutes the fully faceted general classification in existent is quite conventional. Bc2 is now virtually a new general classification and constitutes the fully faceted general classification in existence. Here we intend to study the major living general classification schemes.

## 2.2 Dewey Decimal Classification (DDC)

Of modern library classification schemes, the Dewey Decimal Classification (DDC) is both the oldest and the most widely used in libraries all over the world. Such widespread use is a tribute to Melvil (Le Louis Kossuth] Dewey, whose original plan was adaptable enough to accommodate new subjects as they emerged and flexible enough to withstand changes imposed by the passage of time. Born on December 10, 1851, and graduated from Amherst in 1874, Dewey was appointed as an assistant College librarian. He drew the first draft of his system for arranging books on the shelves at that time.

He soon became a leader in American librarianship and founded both the American Library Association (ALA) and the first American Library school at Columbia University. He was also an advocate of spelling reform.

The first edition of Dewey's scheme was published anonymously in 1876. The second edition was issued under Dewey's name in 1885. Since that time 22 more full editions and 13 abridgments have appeared. The present

twenty-second edition (DDC-22) was published in 2003. The twenty-first edition (DDC-2I) appeared in 1996. The associated thirteenth abridged edition was published in 1997. DDC notations arc assigned the 082 in MARC 21 (the Current U.S./Canadian Machine-Readable Cataloging format) when they have been created for a particular item by the Library of Congress (LC). A DDC notation created by a local library participating in a network is placed in MARC.

### 2.2.1 Changes in the Schedules of DDC-21

In DDC 21 three significant areas have undergone complete remodelling : 350—354—Public administration, 370—Education, and 560—590—Life sciences. For public administration and life sciences facets and facet indicators are a basic part of the design. Faceting in this form was introduced in the music schedule and that was completely revised for the twentieth edition. It allows building of members through use ofthe indicators 0 and 1. Faceting makes possible the identification of meaningful components in a number both in the classification process and in the retrieval process.

In addition to those complete revisions, the 21st edition relocates the standard subdivisions of Christianity from 201—209 to 230—270. In future editions more reduction of bias toward Christianity is expected. Also dealt with in the 21st edition are political and social changes such as revision of the geographic area members for the countries of the former Soviet Union. New topies such as rap music and snowboarding have been added, terminology has been updated, and attention has been given to international needs.

### 2.2.2 Difficulties

Among the difficulties built into the DDC system are its long numbers, which increase as the system grows, nullifying much of the mnemonic character of the basic system. Thus the number 636.08969897, which was cited on page 286 as coming from the relative index entry for radiation injury in veterinary medicine, is so long that any mnemonic associations between it and the number 616.9897 (from which it was built) are obscured. Librarians who intend to retain these long numbers because of extensive

holdings in one or more fields should write them on cards and items to be shelved in several lines. The above number could be written in short segments as follows :

636
.089
698
97

field 092. DDC complete call numbers are also placed in the field 092.

The DDC system has many advantages. Its content is compact, consisting in DDC 21 of a volume for a introductory matter, anxiliary tables, and a list of relocations and schedule reductions, two volumes for schedule summaries and schedule development and a fourth volume for the index and the manual. A basic premise of the Dewey approach is that there is no one class for any given subject, The primary arrangement is by discipline. Any specific topic may appear in any number of disciplines. Various aspects of such a topic are usually brought together in the relative index. For example, a work on 'families' may be classed in one of several places depending on its emphasis thus :

| | |
|---|---|
| 173 | Ethics of family relationships |
| 241.63 | Christian family ethics |
| 296.4 | Religious family rites, celebrations, services |
| 304.666 | Family planning |
| 306.8 | Marriage and family |
| 362.82 | Families with specific problems |
| 392.3 | Dwelling places [including those for families] |
| 6.16.89156 | Family psychotherapy |
| 796 | Sparts for families |
| 929.2 | Family histories. |

The basic concepts of the system are found in two places in DDC 21 : the introduction in volume 1 and the manual in volume 4. In addition to the manual in DDC 21, two outside sources are available for guidance : **Dewey Decimal Classification : A Practical Guide,** by Lois Mai Chan, John P. Comaromi, Joan S. MitchelJ, and Mohinder, P. Satija, and **Dewey Decimal**

**Classification, 21st Edition : A study Manual and Number Building Guide,** by Mona L. Scott.

### 2.2.3 Changes in DDC 22nd Edition

The revisions in the 22nd edition are not as sweeping as those undertaken in the 21st edition. Some major changes in specific areas are :

Data processing in computer science 004—006

Completion of the relocations and expansions in 200 Religion.

Relocation of Culture and institutions of indigenous racial, ethnic, national groups from 306.08 to 305.8.

Updates in 510 Mathematics.

Updates and expansions in 610 Medicine and health updated historical periods throughout 930-990.

The more significant of these changes are outlined below :

Table 5 **Ethnic and National Groups**

The name of the Table 5 has been changed from **Racial, Ethnic and National Groups** to **Ethnic and National Groups** to reflect the de-emphasis on race in current trends.

Table 7 **Groups of persons.**

This table has been removed and replaced with direct use of notation already available in the schedules and in notation 08 from Table 1. A complete list of revised add instructions is provided in the introduction to Relocations immediately following Tables 1—6 in both print and web editions.

**Web Dewey and Abridged Web Dewey.**

DDC 22 and ADDC 13 are available by subscription on the Internet.

## 2.3 Universal Decimal Classification (UDC)

The UDC was adapted from the DDC in 1895 by two Belgian lawyers, Paul otlet and Henri La Fontaine, for the classification of a huge catalogue

of the World's literature in all fields of knowledge. It was based on the fifth edition of DDC but was with Dewey's permission expanded by the addition of many more detailed subdivisions and the use of typographical signs to indicate complex subjects and what we know today as facets. DDC's decimal notation was retained except for final zeros, but class 4(i.e., DDC 400) has amalgamated with class 8 and is currently vacant. Many major and almost all minor subdivisions are now quite different from those in DDC The main difference lies, however, in the synthetic structure of" UDC. Thus a work dealing with two or more subjects can be classed by two or more UDC class notations linked by a colon sign, for example,

362.1 : 658.3 : 681.34 Hospital : Personnel management : Computers.

For a work on the use of computers in the management of hospital personnel. Such a class notation is; however, not a 'call number' but is intended for a classified catalogue in which each of the three class notations may serve as an access point, while the other two are shown in notation. For example.

658.3 : 681.31 : 362.1 and 681.31 : 362.1 : 658.3.

If UDC is to be used for shelf classification, one of three class notations may be chosen as a call number for a work on this complex subject.

UDC's faceted structure has its roots in DDCs device for indication of place, namely the intercalation of —09 followed by the class notation for a country of region for example,—0954 for India. UDC uses largely the same place notation as DDC but encloses them in parenthesis. Thus, "plant cultivation in India" is 631.5(54). In addition to place facet UDC has also specific symbols and notations for the language of a work, its physical form, races and people, time periods, materials, persons, specific points of view, and recurring subdivisions in certain classes.

### 2.3.1 UDC Medium Edition

The Medium edition was published in 1985—1988 in 2 Parts. Part I : 1985 consists of Systematic Tables and Part-2 : 1988 is the alphabetical subject index. Part I contains the schedules of about 70,000 concepts arranged in systematic order of main classes 0/9 preceded by preface, introduction (operation manual) and common auxiliary tables. The common auxiliary

tables are marked 1(a) to 1(k). These common auxiliary tables are applicable to any class number in schedules.

### 2.3.2 New Changes in the Medium Edition

Class 4 Languages has been merged with 8 Literature. The schedule for computers has been updated at 681.3 Data processing equipment. The class 38 Commerce (except 389) has been merged with 33 Economics. The whole of Sociology has been placed at 301. The class 34 Law has been expanded. The classes 5 and 6 Science and Technology are classed with full details. The schedule for Space Sciences has been extended and incorporated at 629.7 In the signs of additions and relations two new signs have been added, for example, Double Colon—the order of components being irreversible. The square brackets [] has been assigned the function algebraic subgrouping.

Two new common auxiliaries have been added. These are :

Table 1(k)— 03 hyphen nought nought three for materials.

Table 1(k)— 05 hyphen nonght nonght five for persons.

The medium edition widens the use of apostrophe 7'9 the part 2 contains 1,20,000 entries generated by computer and are arranged word by word.

### 2.3.3 Management, Revision and Use

Until 1992 UDC was managed by the International Federation of Information and Documentation (FID) in the Hague (Netherlands). When it became apparent in the 1990s that a more broadly based organisation was needed to administer UDC. FID and publishers of the Dutch, English, French, Japanese and Spanish editions combined to found a new body, the UDC Consortium (UDCC). An early action of the UDCC was to create an international database that would serve as a master file the database, called the Master Reference File (MRF) and containing about 62,000 entries, is held at the Royal Library in the Hague and is updated once a year. An Editor-in-Chief and an Editional Board of international membership oversee the continuous revision and expansion.

Since 1992, UDCC has been maintaining the scheme by reviewing its content and initiating revisions and extensions. The results are brought out in **Extensions and Corrections to the UDC.** A two-volume, easy to use edition

of UDC was published in its second edition by the British Standards Institution in 1993. It is referred to as BS 1000M. Supplements are issued each year, each one cumulating all previous ones so that one has only to look in two places for the latest notations. The newest edition is a compact one volume "pocket edition" published in 1999. It contains about 4000 entries and is referred to as PD 1000.

UDC has been published in whole or in part in 23 languages. It is widely used in special libraries and information centres of Europe, Canada, Australia, New Zealand, India. The UDC is now being geared for use in on-line catalogues. In the USA UDC is used mainly in some scientific and technical libraries and by one abstracting database.

## 2.4 Library of Congress Classification (LCC)

The library of Congress was founded in 1800. Its earliest classification was by size, subdivided by accession numbers. May significant changes occurred at LC near the turn of the century. Dr. Herbert Putnam, the Librarian decided to reorganise and reclassify his rapidly growing collection in 1899. The debt especially to Cutter is implicit in the basic structure of the system. While the outline and notation of main classes of LCC are very similar to those of the **Expansive Classification,** there are no main classes I, O, W.X, Y as they are in Cutter system. To keep the system functionally up-to-date, individual schedule volumes are frequently reviewed in committee. Revisions, reallocations, and additions keep it flexible and hospitable to new subjects or points of view. For example, in.the 1960s interest in Eastern regions and the increase in materials from Asia occasioned a reallocation in 1972 of the topic "Buddhism" from the span BL—1495 into a whole new subclass, BQ. Revisions were likewise made in subclass PL, particularly in the sections for Chinese, Japans and Korean literatures.

In MARC 21 Call numbers based on LCC that are assigned by the LC or the British Library are placed in the field 050. Those assigned by the National Library of Canada, the National Library of Medicine, or the National Agricultural Library are placed in fields 055, 060, or 070 respectively. In the OCLC system, members are asked to enter locally assigned call numbers based on LCC in field 090.

### 2.4.1 The System

The working schedules are contained in over 40 separate volumes. Besides the basic schedules, there are a separately published partial index, for P-PM subcategories in the Language and Literature class and a short-general Outline, now in the sixth edition.

The main classes area :—

| | | | |
|---|---|---|---|
| A | General work, 1998 Edition | QAT | Science, Technology. Medicine |
| B | Philosophy, Psychology, Religion | 07V | Military Science, Navat Science |
| C-F/G | History and Geography | Z | Bibliography and Library Science |
| H/L | Social Sciences | A-Z | Outline (6th ed. 1990) |
| M/P | Humanities | | |

Because LCC was developed as utilitarian scheme for books at LC. it is an enumerative system. Among the basic features borrowed from Cutter are its order of main classes, its use of capital letters for main and subclass notation, its use of Indo-Arabic numerals for further subdivision and its modifications of the Cutter author-mark idea to achieve alphabetic subarrangements*of various kinds. All the LC schedules have similar, but not identical sequencing arrangements and physical appearance. Within each sequence of class numbers the order proceeds, as a rule, from general aspects of the topic or discipline to its particular divisions and subtopics. Chronological sequences may trace historical events, publication dates or other useful time frames. Geopraphical arrangements are frequently alphabetical but just as frequently given in a "preferred order", starting with Western Hemisphere and United states. LCC provides broad subclasses in class P for various national literatures, subdividing next by chronology and then by individual author. A recurring pattern of organisation within each literature affords the shelf or shelflist browser a useful guide :

1. History and Citicism, subdivided
   a. Chronologically
   b. Then by form
2. Collections or anthologies, subdivided by form
3. Individual authors, subdivided
   a. Chronologically
   b. Then alphabetically by author

(1) Collective works
(2) Individual works
(3) Biography and criticism.

LCC breaks the "Generaiia" into two classes at opposite ends of the alphabet. General encyclopedias are placed in subclass AE, general indexes in AI and so on.

The typical LCC notation contains a mixed notation of the one to three letters, followed by one to four integers, and possible a short decimal. Decimal numbers are not used much until it becomes necessary to expand certain sections where no further integers are available. Decimals do not usually indicate subordination but allow a new topic or aspect, to be inserted into an established context.

### 2.4.2 The Use and Revision and Future

The scheme is enumerative, rather than a deductive system and in it even form and space auxiliaries are enumerated. It is particularly useful for large university and research collections because of its hospitality and inherent flexibility. It has been used in USA effectively in smaller academic and public libraries, although its adaptability for broad classification is limited. Even special libraries in USA frequently base their own more technical contracts on it, extending its schedules or parts of schedules to cover their unique materials. Some libraries outside USA also use the system, although, in spite of LC's large foreign holdings, it is primarily designed from an American perspective.

It is constantly revised. Updating is accomplished by a variety of publications, most of which are available directly from LC.

**Revised editions of individual schedules** : The various schedules differ widely in the number and kinds of revisions made. All schedules and the **Outline** are sold individually by LC's Cataloguing Distribution Service (CDS) at nominal prices.

**Library of Congress Classification—Additions and Changes.** This publication reports quarterly on the latest adjustments in all schedules and schedules indexes of LCC. Subscriptions may be placed with CDS. The additions and changes are also available on the Web. Substantial changes,

such as the revision of subclass HM 1— 299, are posted on the web in addition to appearing in Library of Congress Classification : Additions and Changes.

LCC is loosely coordinated and essentially pragmatic. It aims firsl to class closely and then to identify uniquely, particular works, using the most economical notation available within its broad parameters of theory and practice, ll has an assured future because of its institutional patronage. It is highly successful classification scheme.

## 2.5 Colon Classification (CC)

Shiyali Ramamrita Ranganathan is considered to be the foremost theorist in the field of classification. Colon classification is the manifestation of his theory.Although Ranganathan's scheme of classification has not been widely used his theory has influenced in one or another, all currently used classification schemes. Ranganathan wrote his Prolegomena to Library Classification in 1937 (3rd edition, 1967) in which he presented the theoretical basis for his classification. Over the years Ranganathan refined and redefined his thinking about classification. The first edition was published in 1933. The sixth edition came out in 1960 and the seventh appeared in 1987. Each edition reflected the progress of his thinking. Sometimes drastic changes took place between editions. The stability was sacrificed for the sake of keeping up with knowledge.

### 2.5.1 Seventh Edition

In the seventh edition the following 10 types of basic subjects have been indicated.

1. Main Basic Subject
   - 1.1 Traditional
   - 1.2 Newly emerging
   - 1.3 Fused
   - 1.4 Distilled
   - 1.5 Cluster
   - 1.6 Agglomerates
2. Non-Main-Subjects.
   - 2.1 Canonical classes
   - 2.2 System constituents
   - 2.3 Special constituents
   - 2.4 Environmental constituents.

In the seventh edition, the number of basic subjects has been increased to 700. However, the traditional basic subjects remain the same.

A few new basic subjects remain the same.

| | | | |
|---|---|---|---|
| 1 | Universe of subjects | 4 | Mass communication |
| BT | Statistical Calculus | LT | Physical exercise and sports |
| M4 | Smithy | PW6 | Typewriting |

In the new edition, the mainfestations of 'Matter' are of three kinds—Matter Material, Matter Property and Matter Method. Upto the sixth edition, matter was present in a few classes. Matter Material usually implies mateiial used for construction. In the subject, Sculpture stone, marble are considered Matter material. In Library Science books, periodicals, maps, etc. are considered Matter material. In sixth edition, in a few cases, what was considered energy now forms part of matter facet. In Medicine, anatomy, physology, diseases are now viewed as Matter Properly. Similarly, in Agriculture, soil, manure, propagation are treated as manifestations of Matter Property. Magnetic method, chemical method are Matter Method isolates to be found in Analytical Chemistry.

In the 7th edition, the Common Isolates have advanced both in kind and number, mostly in Posteriorising Common Isolates. The new schedules of Common Isolates are :

Common Matter Property (Chapter DL)

Common Energy Isolates (Chapter DK)

**The Anteriorising Common Isolates are now added by the indicator digit "double inverted comma." For example**

**Encyclopedia of Physics**

**C''$_k$**

**The most important addition is the enumeration of Environmental Divisions (Chapter ED) (Chapter DD). These can be added to any subject using **- 9" as the indicator digit. Another important change is the use of Speciators of kind 1 and kind 2 to be denoted by "—" hyphen and "=" equal signs.**

In CC7 the insertion of connecting symbol, (comma) is made obligatory

even in case of first level of Personality. Moreover, the facet structure of different compound subjects has been changed in CC7 (vide chapter DR, Page 108) In Phase relation digit O(zero) has been replaced by & (Ampersand) and a new kind of phase relation "tool" has been reintroduced.

Emptying digits and Empty-Emptying digits have been reorganised.
Emptying Digit

| | | | |
|---|---|---|---|
| | | 44 | India |
| L | Medicine | 44T | Nepal |
| LX | Pharmacognosy | 44V | Ceylon |
| M | Useful Arts | 44X | Pakistan |
| | | 45 | Iran |

Empty-Emptying digits : infinite interpolations are possible

| | |
|---|---|
| LU5 | Public health |
| LU6 | Hospital |
| LU7 | Sanitarium |
| LUD | Medical technology |
| M | Useful Arts |

Systems and Specials have found place in the schedule of basic subjects in 7th edition.

| | | | |
|---|---|---|---|
| L-K | Allopathy | K | stands for 1600 to 1699 A.D. |
| L-L | Homeopathy | L | stands for 1700 to 1799 A.D. |
| X-J | Capitalism | J | stands for 1500 to 1599 A.D. |
| X-Nl | Communisim | Nl | stands for 1910's. |

### 2.5.2 Notation

Colon classification adopts mixed base. The notation consists of :—

1. Indo—Arabic numerals I—9
2. Roman alphabet, both capitals and lower case, A to Z and a to z
3. Parenthesis ( )
4. Indicator digits
5. A Greek letter (delta)

The capacity of base is 60, the total number of indicator digti-s is 14. The total number of digits is 74. Indicator Digits arc divided into three groups.

**Group A** : Anteriorising Indicators

Asterisk indicates agglomeration and interpolation
Backward arrow indicates backward times range, "Double inverted comma : ACI

**Group B** : Posteriorising Indicators

| | |
|---|---|
| & | Ampersand indicates phase relation |
| ' ' | single inverted comma indicates time facet |
| . | dot indicates space facet |
| : | colon indicates energy facet |
| , | comma indicates presonality facet |
| ; | semi-colon indicates matter facet |
| – | Hyphen indicates speciator of kind 1 |
| = | equal sign indicates speciator of kind 2 |
| + | plus sign indicates addition |
| → | Forward arrow indicates time range |

**Group C** : Indicators for levels of phase relation

### 2.5.3 Evaluation

This is the great universal classification scheme to be designed by one person— mathematician-turned-librarian. The Colon classification stands in a class of its own for coherence and system making it the easiest of the general classification schemes to use properly. It has coustituted an almost complete break with the traditional method of classifying. In fact, Ranganathan's Colon Classification manifests the threory of facet analysis and synthesis. The systematic order and the degree of detail due to analysis

and synthesis are the two merits of CC, Although CC is used relatively in few libraries its underlying theory has had a major impact on all classification schemes. Revision of DDC, particularly in recent editions,reflects (he influence of facet analysis and synthesis.

It the implementation of the colon is limited, its impact still provides guidelines as to what is missing in the techniques of other classification systems, in spite of its "grand unintelligibility' as pointed out by Shera quoting from Carlyle, it has provided a refreshingly stimulating contribution to our subject. It has given an enormous impetus to ciassificatory research in systematizing the organisation of law. canons, postulates and principles devices, facets. Remembering the personality of this illuminated classificationist Cavalcanti once remarked that Colon was felt like rhythm of poetry. When others provide frameworks for warehousing information, Colon assures enriched material to be used in the classification 'Cylotrons' of the future.

The revision policy of the scheme has been criticized. Major Changes have been incorporated in the seventh edition for obvious reasons. The changes are so enormous that many libraries may not be in a position to change over to the current edition. But still the Colon Classification will certainly make the strongest appeal to the seekers of perfection.

## 2.6 Bibliographic Classification (BC)

The classification (known as BC) was originally devised by Henry Evelyn Bliss and was first published in four volumes in the USE between 1940 and 1953. Bliss stated that one of the purposes of the classification was to "demonstrate that a coherent and comprehensive system, based on logical princiles of classification and consistent with the systems of science and education, may be available to services in libraries, "to aid revision ... of long established ... classications" and to provide an "adaptable, efficient and economical classification, notation and index" A fundamental principle is the idea of subordination—each specific subject is subordinated to the appropriate general one. This version is now known as BCI.

On the formation of the Bilss Classification Association (BCA) in 1967, it was suggested that a new and completely revised edition of the full BC

should be made avaliable. However, the revision has been so radical that it is more accurately described as a completely new system, using only the broad outline developed by H. E. Bliss. Its main features are outlined below, but in addition to these the vocabulary is.very much greater than that of BCL An annual Bulletin had been providing revised schedules, moslty in science and technology. The new, revised edition was initiated by Jack Mills and was to be produced in 22 parts, comprising one or two subjects per volume. The first volume was published in 1977 (Bliss Bibliographic Classification, edited by Jack Mills and Vanda Broughton, London, Butterworth, (1977—). Publication is now Undertaken by K. G. Saur. Further revisions have been mader to some of the BC2 volumes in order to retain subject currency and updates continue to be published in the BCA Bulletin.

### 2.6.1 The Second Edition of ($BC_2$)

The aim of BC2A was to revise and revive the system and promote its use. The second edition known as BC2 is edited by Jack Mills. The BC2 is so radically revised that it is accurately described as a new system. It derives techniques of facet analysis, as will as explicit citation and filing orders from Ranganathan's contributions to classification theory. The outline of BC2 is as follows :

Introduction & Auxiliary schedules. 1977

| | |
|---|---|
| 2/9 | Generalia, Phenomena, Knowledge, Information science & Technology |
| A/AL | Philosophy & Logic, 1991 |
| AM/AX | Mathematics, Probability, statistics. 1993 |
| AY | General Science, 1999. |
| C | Chemistry, Chemical Engineering, 2000 |
| D | Space & Earth Sciences |
| E/GQ | Biological Sciences |
| F | Botany |
| G | Zoology |

| | |
|---|---|
| GR | Agriculture |
| GU | Veterinary Science |
| GY | Ecology |
| H | Physical Anthropology. Human biology Health Sciences. 1980. |
| I | Psychology & Psychiatry, 1978. |
| J | Education 1990. |
| K | Society, 1984. |
| L/O | History (includes Archaeology, biology and travel) |
| P | Religions, Occult, Morals and ethics 1977 |
| Q | Social welfare & criminology. Rev ed, 1994 |
| R | Politics & Public administration 1996 |
| S | Law 1996. |
| T | Economics & Management of economic enterprises, 1987. |
| u/v | Technology, Engineering, 2000 |
| w | Recreation, Arts. Music |
| x/y | Language, Literature. |

### 2.6.2 Commendable Features

The Bibliographic classification has many commendable features :

1. The order of main classes is based on the Educational and Scientific consensus.
2. Alternative locations are provided in the scheme.
3. It provides short notation consisting of Roman capitals and Indo-Aratic numerals.
4. It uses retroactive notation which can be combined backward.
5. It is last general classification scheme based on facet analysis expounded by Ranganathan. BC-2 has avoided many complexities of Colon Classification. BC-2 will be in 19 parts of which 15 have been published.Each part has its own index prepared according to chain

indexing. A cumulative index will be brought out on the completion of the whole scheme.

## 2.7 Broad System of Ordering (BSO)

For many years the members of the scientific community have been feeling after powerful tools for arrangement and retrieval of subject information. However, slight be the achievement of the Royal Society and Scientific Information Conference in 1948, its intention was significant. In fact, it marked the beginning of a new ear and concentrated on the concern of the scientific community to cope with the ever increasing flood of scientific and technical publication. There was a thrill of suggestions in the Washington Conference in 1958. In spite of the significant role of the computer, developments tended to be along rather conventional lines. A new vista opened up during the 1960's. The developed countries were on the whole managing to keep the problems in the handling of information under control. But the third world countries were increasingly at a disadvantage in not having the same access to the free flow of scientific information as the developed world. In this context a joint control committee was set up by UNESCO and ICSU (International Council of Scientific Unions) in 1967 to carry out a feasibility study of a World Science Information System, to be called UNISIST. A detailed report of the committee was largely approved at an International Congerence in October 1971.

The UNESCO programme was aimed at the development of a network of information systems. In the UNISIST study (UNESCO, 1971) several topics were discussed. It was also suggested that a standard list of broad subject headings might prove useful to locate and transfer large blocks of information. Wysocki presented a summary of what should be expected of a Broad System of Ordering at the FID conference in Budapest in September 1972 and The FID was entrusted with the task of developing such a system. It may be mentioned that prior to a contract with UNESCO on the elaboration of such a Broad System the FID/CCC had been investigating whether UDC could be transformed into a 'rood classification' for other classification schemes or not.

### 2.7.1 Facet structures

The BSO is a faceted classification inspite of the fact that the facet structure is not explicitly set out in the schedules. In majority of cases the citation order within subject fields is regularly the reverse of the sequence of the separate elements as found in the schedules. Thus implication of this for schedule sequence is that within each subject field, the further detail is laid out in the schedules in sequence corresponding to an underlying series of facets and the pattern is recurred as we pass from one major domain to another, for example, natural Sciences to Life Sciences, Life Science to Human Sciences; and so on.

The basic facet pattern embodied in particular subject fields is as follows:

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

In addition to the above fact structures there are common Time and Place facets. For the use in classing organisations and information sources by BSO, an optional facet is also provided to designate the type of organisation for information sources. Thus the scheme has three general facets roughly corresponding to form anxiliaries in other schemes.

### 2.7.2 Notation

The scheme provides an entirely new system of notation based on Indo-arabic numerals. The numerals are supplemented by two punctuation, sings—the hyphen and, the comma and occasionally roman alphabet is used where only individualisation but not grouping is required. In ordinal values the punctuation signs precede numerals and numerals precede the alphabets.

(b) an abstracting or indexing service
(c) an information collection (as a library or a data bank) and
(d) a department of university teaching.

Thus the criterion for cut-off actual in BSO was in a form of "institutional warrant". It may be noted that the category (d) was later dropped.

### 2.7.3 Main Classes

The basic order of main classes resembles closely that Bibliographic Classification (BC), identifying the following main areas into which main classes can be categorised.

Second Outline of BSO

088 Phenomena and Entities from a multi-non-disciplinary point of view

100 Knowledge generally

200 Science and Technology (Together)

300 Life Sciences

359 Applications of Life Sciences

460 Education

470 Human Needs

500 Humanities and Social studies

600 Technology

710 Construction Technology

740 Transport Technology and Services

760 Military Science and Technology

780 Mining

800 Process Industries

860 Metal Technology

910 Language and Literature

940 Arts

970 Religion and Atheism

992 Esoteric practices and movements.

BSO becomes a 'switching language' for the transfer of 'large blocks of information' or whether a Standard Reference Code (SRC) could be developed.

But before completion of this investigation a new combined group was formed, following the Budapest Conference. This group included representatives from both the FID/CCC and FID/CR.

In may 1975 this group presented to this Advisory Board of UNISIST at its meeting in Paris, a scheme to be called the Broad System of Ordering (BSO). Thus a paper presented on behalf or UNESCO at the FID meeting bore its first visible fruit in 1978 in the form of published "Broad System of Ordering : Schedule and Index." The work has been delegated to three—member committee consisting of Eric Coates. Geofrey Lloyed and Dusan Simandl.

### 2.7.4 Definition and Scope

The scheme has been devised to function as—

(a) A tool for interconnection of information systems, services and centres.
(b) A tool for tagging (that is, shallow indexing)
(c) a referral tool for identification and location of all kinds of information sources and services.

The proposed switching language for UNISIST was rightly described as a broad classification. But at the outset there were considerable difficulties in arriving at an operationally satisfactory and practicable definition *of* what was to be understood by "broad'. On the one hand the scheme was not to be too detailed, which on the plea of 'broad' might have been overlooked. Finally, the criterion adopted for what was to be included in the scheme was dependent on the extent to which organisational activity and institutionalization were revealed in relation to subject field or area of study. Thus there was an inquiry about the existence of an "organised information source' in relation to the subject. Organised information was detailed as the following :

(a) an organisation supporting or sponsoring the regular issue of specialist information.

All codes begin with a number of the millesimal array 000 to 999. Further subjects may be interposed between any two subjects between consecutive numbers of centesimal array be adding a comma followed by two digits from a further 00 to 99 array.

The notation tor primary topics consists of three digits for main headings followed by a comma and two more digits for subdivision :

| | |
|---|---|
| 716 | Building construction and services |
| 10.30 | Building materials |
| .30 | Building construction work |
| .37 | Timber Construction |
| .40 | Parts of buildings |
| ,45 | Walls |

Combination within the same subject field is achieved by the use of. 0, as a linking symbol.

Complex subjects can be expressed by combining two notations, separated by hypen and the citation order is governed by the following relational formula :

"Cite first : the notation for the element denoting application area, mission, purpose, and 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 : more generally the subject element which "contributes" an aspect approach or action".

For example, the citation order for environmental aspects of building construction is, Building construction 716) followed by Environment (290), giving composite notation 716—390.

Similarly, Chemistry in glass technology is 856—230 and Medical gurisprudence is 420—560.

### 2.7.5 Index

The index has definitely enhanced the usefulness of the scheme. But this has been criticized in connection with the inadequate size if its entry vocabulary. It contains about twenty five per cent more entries than the schedules, which indicate that synonyms are not granted generous treatment. The index exhibits cross-references and see references. Entries are more specific

than the heading they load to. The index entry is indicated by a '0' preceding it.

### 2.7.6 Observations

The overall order of main subject fields offers an interesting structure because it is quite different from that in most other schemes. In fact, it has resulted in .some coveted collocations, but also in one or two undersirable separations. The most prominent perhaps, is the placement of philosophy as the first discipline to be treated while Religion is the last. But there are several instances of welcome arrangements. It is helpful to have subjects such as communication and management at the early part of the schedule. The physical Sciences stand near the beginning and modern areas of study such as plasma and fluids in physics and polymer chemistry are paid proper attention. But closely related applications such as Materials and Energy are treated with Technology. Geography fares well here, being placed among earth Sciences, immediately following Astronomy. The Enviroment has achieved a disciplinary status through institutional warrant, which in turn, has brought Human Needs class earlier to offer a useful umbrella for topics such as housing, clothing, etc. The placement of History under combined reading Humanities and Social Sciences evades the issue of whether or not it is a social science.

The succeession of topics reflects the theory of integrative level as explored *by* the Classification Research Group (CRG).

A. C. Foskett points out that the scheme suffers from two weaknesses. The low intensity of indexing necessitates improvement. The second defect is the absence of an established organisation to maintain the vigour of scheme.

The FID relinquished responsibility for the scheme in 1990, and the copyright is in the hand of the BSO Panel. Although the BSO was not designed for classification of books in libraries, it could very well be employed for broad goal as an international switching language.

## 2.8 Summary

The DDC was brought out in 1876. The 22nd edition came out in 2003.

*DDC* notations are assigned the 082 in MARC 21 Despite many defects DDC is still the most popular scheme. DDC 22 and Abridged edition 13 are available on the Internet. The UDC based on DDC 5th edition has elaborated its syntheitc structure. UDC Medium Edition was published in 1985—1988 in two parts, To look after *UDC*-revision and maintenance UDCC came into being in 1991. UDC has been published 23 laugnages. The LCC based on the Expansive Classification consists of more than 40 separate volumes, In MARC 21 call numbers based on LCC are assigned by the Library of Congress or the British Library in the field 050. Those assigned by the National Library of Canada, the National Library of Medicine and the National Agricultural Library are placed in the fields 055, 060 or 070 respectively. The CC first published in 1933 in not a popular scheme but is theory has influenced all the existing schemes. The seventh edition published in 1987 has accommodated drastic changes. The changes are so enormous that many libraries may not switch over to this new edition. The Biblographic Classification originally devised by Bliss has been undergoing thorough revision. Jack Mills and Vanda Broughton are editing the fully acted general scheme. The BSO, a switching language is the product of three classificationists namely E. J. Coats, Geofrey Lloyed and Dusan Simandl. The proposed switching language for UNISIST was rightly described as a broad classification. It is now in its third edition published in 1978.

## 2.9 Exercise

1. Describe the salient features of DDC 21 and DDC 22
2. What are the major changes in UDC International Medium Edition?
3. What are major changes in CC7?
4. Discuss the unique features of BC 2
5. Why was BSO designed? Discuss its salient features.

## 2.10 References and Further Reading

1. Chakruborty, A. R. and Chakrabarti. B. : Indexing : principles, processes and products. World Press, 1984.

2. Chan, Lois M : A Guide to Library ol Congress Classification. 5th ed.. Libraries Unlimited. 1999.
3. Foskett, A. C. : The Subject approach to information. 5th ed.. Library association Publiching, 1996.
4. Satija, M. P. : Colon Classification 7th ed. A practical introduction. Ess Ess, 1989.
5. Taylor, Airline G : Wynar's introduction to Cataloging and classification. 9th ed.. Libraries Unlimited, 2000.

# Unit 3 ❐ Classification Systems 2 : Special Systems

**Structure**

## 3.0 Objectives

After reading this unit you will be able to :—

1. understand the need for special schemes
2. identify the various types of special schemes
3. acquaint with a few special schemes in The fields of Science, Social sciences and Humanities

## 3.1 Introduction

Classification schemes comprising the entire universe of knowledge, as noted in the earlier unit, must of necessity be general and cannot deal with finer details. They are also largely inflexible, presenting the particular viewpoint of their designers. DDC, UDC and LCC represent the late nineteenth century. They often disperse subjects that, for the purposes of specialists, ought to be dealt with in close proximity. For example, DDC has chemistry in 540 but chemical technology in 660. which is very unhelpful for a chemistry library. Even UDC although providing for very fine detail, suffers from many of the faults inherited from DDC and often has long notations.

Librarians of collections devoted to one specific field to knowledge, a discipline, or one of its subfields have therefore often found it necessary to design special classification schemes. Special schemes are designed for a variety purposes. They are constructed to bring out classified indexing and abstracting services, compilation of bibliographies to satisfy the requirements of specialised groups of user community.

Ranganathan defined specialist classification scheme as "scheme designed for depth classification of micro subjects, going with one and only one specified subject field." Jack Mills made a distinction between general and special schemes. According to Mills general schemes embrace all knowledge in the subject classes, special schemes are restricted to varying degress to conventional subject field (e.g. physics, chemistry, etc.), to certain physical forms (e.g. pictures, gramophone records, etc.), to certain form of publications (e.g. Patents, standards, etc.) to certain type of readers (e.g. children, visually handicapped, etc.) and so on.

## 3.2 Need for Special Schemes

Special classification schemes are needed with more limited aims of covering just one main subject field. Since the last Second World War the emphasis has been primarily on the development of science and technology, secondly on social problems produced by the so-called progress of technology. The increasing output of science and technology during the 20th Century concentrated the attention of Americans and Europeans on the needs of classification of special subjects. Americans tried to do without classification. In England the more systematic European tradition is exemplified by the classification Research Group. Its origins lay in the demands of scientists for better classification of their literature. However, there are many reasons claimed for the need of a special classification scheme. Some of them are given below :

(a) Lack of co-extensiveness : most general schemes *do* not provide sufficient details required for dealing with micro-documents in documentation.

(b) Special requirements or special point of view : general schemes take into consideration the majority point of view. Therefore, these are able to cope with special requirements of a particular special library :

(c) Lack of flexibility : provision for new subjects without disturbing the preferred sequence. Very often, general schemes lack flexibility to certain extent.

## 3.3 Types of Special Schemes

Special classification schemes are mainly of two types : "general-special" schemes which classify a special field or subject exhaustively while providing only very general class marks to peripheral or extraneous subjects, and "special" schemes, dealing only with a particular specialty in sometimes very fine detail bit learning the classifying of other subjects to one of the general schemes. Hybrids of the two types are those schemes that expand an existing class of a general scheme or make use of an unused notation under which a special subject field is developed in detail.

Two examples of latter type may be cited here. **The National Library of Medicine Classification** uses class W which remains vacant and the medical parts of LCC's class Q for the entire biomedical field. Subdivision of W is by all letters of the alphabet, including even I and O which are avoided by LCC, followed by up to three digits used enumeratively and an occasional fourth digit used decimally, followed by more detailed subdivision using cuttering. For example.'

WD Metabolic diseases
WD 205.3 A.5 Amino acid metabolism

Since NLM's W class appear on MARC records for medical books, most medical libraries use the W special classification. **A classification Scheme for Law Books,** which was derived by expanding class K of LCC was published in 1968 and found wide acceptance. But the LCC has completed many parts of class K in recent years thereby reducing the need for a separate scheme.

## 3.4 Design of Special Scheme

There has been the major development in classification for IR in libraries in the past fifty years, although its first formulation was in the work of Ranganathan. Although curiously enough, Rangananthan never referred explicitly to the fact, the fundamental feature of his colon classification is that it divides any given subject in accordance with the rules of logical division. But logical division is not the whole story. The work of BC2. covering every field of knowledge, clearly has shown that the design of a special classification requires recognition of six fundamental steps. Only the first two use logical division; the other four use extralogical procedure. The steps are summarised below :—

1. Division of the subject into broad facets (Categories)
2. Division of each facet into specific subfacets (usually called arrays, following Ranganathan)
3. Deciding the citation order between facets and between arrys.
4. Deciding the filing order between facets and between arrays and the order of classes within array.

5. Adding a notation.
6. Adding an A/Z index.

## 3.5 The Role of Logical Division

Before considering each of these steps in detail, the general role of logical division, which governs the crucial first two steps, must be noted. The rules of logical division, developed more than two millennia ago, are admirably brief :

1. Only one characteristic of division should be applied at a time;
2. Division should not make a leap; steps should be proximate;
3. Division should be exhaustive.

The first and crucial rule is purely one of conceptural analysis and does not depend on practical considerations. The second and third rules involve to some extent subjective practical considerations as to the size of vocabulary to be accommodated and the degree of specificity with which compound classes are to be described. They are manifested only at the level of arrays. Observance of the first rule is the hallmark of faceted classification.

## 3.6 Division of Facets

The first step is to assign all the terms constituting the vocabulary of the subject into a limited number of broad categories. The use of the term "Category" requires some explanation here. The outcome of the classification is an almost infinite number of possible subject descriptions of documents or parts of documents, nearly all of which will be compound classes that is, requiring two or more terms to summarise their content. For example, a document on 'radiographic diagnosis of bone cancer' reflects four different categories of concepts in medicine; if the human body' is seen to be the entity with which all medicine is concerned, bone is seen to be a Part, cancer a Process (action internal to the body), diagnosis an operation (an action performed on the body), and radiograph an Agent of the operation. But notion of Part is not a category in traditional sense of the term, since it

implies being a part of something—i.e., it is a relation, not a unique and independent category. Similarly, Agent is relative to the action it assists—it is a relation. So facet analysis might be said to be the assignment of terms to true categories (Time, Space, Matter, etc.) and to relational categories (kind, Part, Agent, etc.).

### 3.6.1 Categories in subject fields

Ranganathan was the first to see the need for initial categories. He provided five and called them Fundamental Categories—Personality, Matter, Energy, Space, Time. He claimed that this order represented one of decreasing concreteness. The (British) Classification Research Group (CRG), formed in 1952, developed a detailed set of Categories, entirely consistent with PMEST in outcome but aiming to be more explicit—particularly in its interpretation of Personality; the set may be summarized as defining system or entity, its Kinds, its Parts, its Materials, its Properties, its Processes, Operations on it, Agents of the Processes and Operations, Place, Time, Forms of Presentation.

Assigning terms to categories is a deductive approach to concept organisation, and it may be noted that one member of the CRG advocated and developed an inductive approach Farradone, 1950). This he appropriately called relational analysis, since it is the relations between concepts that are at the heart of retrieval and categories are really a first step recognising those relations.

### 3.6.2 Facet analysis

This assignment of categories is simply another way of expressing how a particular characteristic of division is applied to obtain classes that share that characteristic, although in different ways (as division of objects by colour will produce classes of different colours). The process is best explained by considering one example :

**Classification of "Medicine"**

"Medicine" may be defined as the technology concerned with the actions taken by the human person to maintain their health and treat their sickness. The definition of the subject leads directly to the primary category (the

defining entity, the person) and all other categories are realised in their relationship to this.

Kinds-of human persons (females, males, young, old ...)

Parts of the person (anatomical and regional, and physiologically functional subsystems—trunk, circulatory, neurological ...) Processes in the person (normal physiology, pathology).

Operations acting on the person (health maintaining or preventive, diagnostic, therapeutic)

Agents of operations (medical personnel, instruments, institutions—hospitals, health services).

So a particular document entitled "Rehabilitation following Fracture of the Femoral Neck [in old persons]" would get index description : Old persons (geriatrics)—Bone—Femur—Neck of Femur—Fracture—Therapy—Rehabilitation,

## 3.7 Division of a Facet into Arrays

The classes constituting each facet are now organised into more specific subfacets (called arrays by Rangauathan). At the facet level, classes are undifferentiated and in most cases will not be mutually exclusive. An array consists of mutually exclusive classes. To achieve this condition, these classes now must be differentiated by applying specified characteristic of division. The process of subdivision continues until characteristics are so specific that they generate mutually exclusive classes in an array, e.g., Persons by age, Persons by sex.

### 3.7.1 Division must be exhaustive

The constituent species collectively must be coextensive with the extension of the genus. In practice all significant kinds of characteristics would be enumerated with a possible residual class for "others".

### 3.7.2 Each step of division should be proximate

Division should not make a leap.

## 3.8 Extralogical Steps in Classification Design

### 3.8.1 Citation Order (combination order)

It may be defined as the order in which the characteristics governing division of a class into its facets and arrays are applied. This in turn is reflected in the order in which the constituent terms/concepts (which together summarise the content of a document) appear in an index-description.

### 3.8.2 Citation Order of Facets

The primary facet in a subject represents a summum genus and other categories at the facet level clearly reflect the different relationships that concepts may have to it. For example, in the class Building technology, the primary facet is that of Buildings. Terms in other facets always imply the relationship of the concept represented to buildings e.g. weather resistance in the properties facet means weather resistance in buildings. This relationship provides a clear and powerful basis for the citation order.

## 3.9 Filing Order of Classes

This is the sequence in which the individual classes, simple or compound, file one after other in a linear order.

## 3.10 Notation

This assigns to each and every class in the system a symbol (class mark) that possesses or is given an ordinal value.

## 3.11 The Alphabetical Index

The A/Z index performs two essential functions : it provides the user of the classification with a key, linking the natural language terms-for the classes to the classmarks that locate them; it complements the systematic display of relations in the hierarchy by showing under any term the distributed relatives.

## 3.12 Examples of Special Classification Schemes

The following examples represent the special type that combine a faceted structure and .high flexibility with brief notations.

1. **The Physics and Astronomy Classification Scheme** : The American Institute of Physics (AIP) created this scheme which uses a mixed notation of the two decimal groups followed by a letter.

84 Electromagnetic technology

84.30 Electronic circuits

84.30L Amplifiers

Except for two sections, the AIP scheme closely resembles the **International Classification for Physics** and parts of it have been adopted by the Institution of Electrical Engineers in London for its Classification **for the INSPEC Database.**

2. **The London Education Classification** (LEC) : After ten years of experiment with the first edition as a Classification scheme the second edition published in 1974 is in thesaurofacet form. It fulfils a double role as a classification and thesaurus and uses a letter notation, with uppercase letters as facet indicators, e.g.

Town USE Urban area

Urban area Bedd

UF Town

BT Community

NT Suburbs

RT City.

Thus the entry Urban area, Bedd in the alphabetical authority list refers to classification number Bedd in the classifiation schedule. This new breed of thesaurus consists of an alphabetical authority list together with the faceted classification from which the list has been derived.

It should be noted that unlike Thesaurofacet it does not require the user to turn to the schedules to find out BT/NT relationships if he does not wish to.

3. **The London Classification for Business Studies (LCBS)** : Like the LEC thesaurus and the thesaurofacet the LCBS was constructed by means of facet analysis. From LCBS a new thesaurus prepared by KGB Bakcwell and DA Cotton was published in 1979. The revised edition (LCBS-2) serves as a thesaurus well as a classification scheme.

Three of the "conventional" thesaurus abbreviations have been used : SN (Scope Notes), UF (Used For) and RT (Related Term). The other two commonly used abbreviations (BT and NT) have not been used as these are obvious from the lay-out of the schedules.

| | |
|---|---|
| Class C : | PRODUCTION |
| | (CEQ Quality Control) |
| | CEQG Reliability |
| UF | Dependability |
| | Durability |
| RT | Production Development CD |
| CEQM | Zero defects |
| CER. | Control of replacements |

Narrower terms are indented two spaces. When part of a class is continued on another page, the broader terms for the first term on the new pages is always given in parenthesis, so that the hieranchy is apparent. Here the 'Quality Control' is the broader term for 'Reliability' and 'Zero defects! 'Reliability' is preferred term rather than 'Dependability' and 'Durability'. Related to 'Reliability' is the term 'Production development' classed at CD.

4. **Classification of Library & Information Science** : The scheme was designed by the Classification Research Group (CRG). The scheme too, makes use of a letter notation with capitals as facet indicators. The CRG published this scheme in 1975.

5. **Classification for London Literature** : Special classification schemes are sometimes developed for extremely narrow subjects. This scheme is concerned with works only on the city of London. It uses decimal subdivisions for every place and event in its long history, e.g.,

10 Religion

12 St Paul's Cathedral

12.4 Dome and roof.

The Guildhall Library has been using this scheme. Its third edition was published in 1966.

A bibliography of classification schemes and thesauri for hundred of different subjects lists 2,250 items in several languages. Classification systems and Thesauri, (1950—1982). Frankfurt, West Germany, Indexs Verlag, 1982 (International Classification and Indexing Bibliography 1).

## 3.13 Exercise

1. Discuss the factors that emphasise the need for special classification schemes.

2. What are the steps to be followed in designing a special classification scheme?

3. Discuss the following classification schemes

   (a) LEC, (b) The Physics and Astronomy classification Scheme, (c) Classification for London Literature.

## 3.14 References and Further Study

1. Farradane, J. E. L : A scientific theory of classification and indexing. **Journal of Documentation** 1950, 6, 83—99.

2. Langridge, D. W. : Classification and indexing in the humanities. Butterworths, 1976.

3. Mills, J., & Broughton, V : Bliss bibliographic Classification. 2nd ed. Butterworths, 1977—

4. Ranganathan, S. R. : Prolegomena to library classification. 3rd ed. Asia Publishing House, 1967.

# Unit 4 ❐ Thesaurus : Origin, Growth and Development

**Structure**



## 4.0 Objectives

Language in IR assumes the form of either natural language or controlled vocabulary. Whenever there is a choice, a question arises as to which language one should use for retrieving information. This unit will give you an idea about the different types of thesaurus right from Roget to Root along with their characteristic features.

## 4.1 Introduction

The word "thesaurus' has come from the Greek 'thesauros' meaning a storehouse or treasury. The Oxford English Dictionary defines 'thesaurus' as an archaeological term, 'a treasury of a temple, etc' and quotes its use in 1736. as a "treasury or storehouse of knowledge., as a dictionary, encyclopedia or the like". An interesting account has been given by Sparck Jones who regards the treatment of synonym as relevant to the history of thesauri. Jones maintains that organisation of a vocabulary on a conceptual basis is an old idea, which has appeared in different cases, in different contexts and for different purposes and modern arguments about the nature and construction of thesauri have their historical analogues.

A historical survey of vocabulary classifications may therefore, be useful. In fact, the oldest and the most persistent form of vocabulary classification is the thesaurus meant for language description and aid to language use. The subject classification of vocabularies grouping related words, goes back to the Sanskrit **Amar Kosha,** the classical **Onomastikon of** Pollax and **Aelfric's Anglo-Saxen Latin Glossary.** Jones cites a sequence of entries from Crabb's work to illustrate his point:

To know, be acquainted with

Knowledge, Science, Learning, Erudition

..... ..... .....

School, Academy

Education, Instruction, Breeding

Ignorant, Illiterate, Unlearned, Unlettered.

..... ..... .....

Though Roget did not think of his thesaurus as a dictionary, but rather as a literary aid, his work can be regarded as the successor of Crabb's work.

## 4.2 Roget

Spark Jones traces the origins of 'synonymy' in dictionaries and identifies the basic difference from natural language : the thesaurus involves

'vocabulary normalisation' Modern usage may be said to date from 1852 when Peter Mark Roget published his first edition of the **Thesaurus of English Words and Phrases.** Roget thought of his thesaurus as a classification of ideas. His subtitle reads "classed and arranged so as to facilitate the expression of ideas and assist in literary composition" The main part of the Thesaurus consists of 'classed catalogue of words', arranged under 'topics or heads of signification, which are grouped into categories that are enclosed by six primary classes. The first three classes cover the external world : class one, Abstract Relations, class two, Space, is concerned with movement, shapes and sizes, while class three Matter. The last three classes deal with the internal world of human beings : the human mind (class four, Intellect), the human will (class five, Volition) and the human heart and soul (class six : Emotion, Religion and Morality). There is a logical progression from abstract concepts, through the material universe, to mankind itself, culminating in what Roget saw as mankind's highest achievement : morality and religion. Roget borrowed his scheme from natural history, with its hierarchy of Phyla, Classes, Orders and Families.

Roget's thesaurus has two characteristics—its purpose and rts form. Its purpose is to help the user move from an idea to the word which the user may use in expressing that idea in a written text. The ideas are, of course, symbolized by words, so that in fact the thesaurus links a certain limited number of idea-words with a much limited number of text-words. The form of this linkage is two fold. First* the idea words are listed in a classified order and against each idea-word are set out all text words, which express the idea. A word which is unwanted but approximates to a required idea, is looked up in an alphabetical index which refers the **user** to one or more groups of words conceptually related, amongst which the **user** may find a preferred word.

Roget's thesaurus is the progenitor of a whole range of thesauri and synonym dictionaries on various plans, topical or alphabetical.

## 4.3 Early Explorer in IR System

According to Lancaster the first thesaurus actually developed for controlling the vocabulary of an information retrieval system was constructed

by the Du Point Company, USA and appears to date from about 1959. The Du Point Thesaurus became the immediate predecessor of the Chemical Engineering Thesaurus brought out by the American Institute of Chemical Engineers in 1961. According to Masterman, Joyce and Needham triggered off the idea of using thesaurus for classing and indexing through their paper titled "The thesaurus approach to information retrieval" in 1958.

The Thesaurus of ASTIA (Armed Services Technical Information Agency) Descriptors was published in 1960. Thus from the experiences at Du Point and ASTIA sixties saw the surge of thesauri development. In fact, development swelled up in succession and secured the status of the thesaurus as the major tool of vocabulary control in modern information retrieval systems.

## 4.4 MeSH (Medical Subject Headings)

Its first edition appeared in 1960, but the second edition in 1963 was first designed specifically for use in the post-co-ordinate machine-based system, MEDLARS. The thesaurus of 16,000 descriptors used in indexing the biomedial literature is thus published as part 2 of Index Medicus in January each year and is revised each year to incorporate new headings as well as any deletions or changes in the form of exhisting headings. MeSH is divided into two sections. The first section lists headings in alphabetical order and includes cross-references from related terms as well as synonyms.

Headings are all in bold : preferred terms are in large capitals. A preferred heading is followed by one or more class numbers from the Categorised list. The plus sign indicates that there are indented descriptors in Tree Structures at this number.

A typical entry from MeSH is as follows :—

JURISPRUDENCE

I 1.880.604.583+ N 3.706.535+

See related

FORENSIC DENTISTRY

FORENSIC MEDICINE

X DUE PROCESS

XU CRIMINAL LAW

XR FORENSIC DENTISTRY

XR FORENSIC MEDICINE

The second section is a classified arrangement of the headings into hierarchies. called the MeSH Tree structures. In the Tree Structures, the headings are allocated to one or more of 15 Categories ; Category A-Anatomy, Category B-Organisms, etc Most categories are further divided into subcategories, each of which is identified by alphanumeric designation, Bl—invertebrates, B2—Vertebrates, and so on. Within each subcategory the headings are arranged hierarchically from the most general to most specific. Seven levels of specificity are available.

## 4.5 Test (Thesaurus of Engineering and Scientific Terms)

Test was produced as a co-operative effort between the Engineers Joint Council and the Department of Defence of the United States in 1964. Its enlarged edition was issued in December 1967 and the second printing in March 1969. TEST covers a very wide range of subjects as its name implies. As well as the main alphabetical sequence listing used and unused terms, there is a subject category list in which the terms are arranged under COSATI subject categories (expanded where necessary). Thus the major part is the alphabetical list which is actually called 'Thesaurus of Terms' as distinct from other sections—an hierarchical index listing 'descriptor families' that is, 'Broader Terms' (BT) together with various levels of 'Narrower Terms' (NT) covered by the broader heading, Subject Category Index, and a computer-generated permuted index presenting an alphabetical listing of all significant words in single and multiword terms.

Main Alphabetical list :—

**Chlorides 0702**

BT Halides

NT Calcium Chlorides

Cerium Chlorides

Copper Chlorides

— Hydrochloric acid

RT Chlorine inorganic compounds

— Chlorine organic compounds

**Chlorinated hydrocarbons**

USE Chlorohydrocarbons

The dash (-) symbol in front of a descriptor indicates that the descriptor has narrowed terms (not shown) and the main entry should be consulted to determine these.

The dagger (t) in front of a term signifies that two or more descriptors are to be used in coordination for that term. The term entry should be consulted to determine these descriptors.

The number (0702) refers to subject categories. The Subject Category Index is in effect a broad subject classification with 22 major subject fields, each of which is subdivided into groups; Each Subject Category Field has a two-figure number and each Group has a further two-figure number.

Although these categories are identified by subject, they are in fact arranged in alphabetical order with the exception of No 10, which is meant for 'Nonpropulsive energy conversion'.

The Hierarchical Index displays descriptor families based on BT—NT reference as shown in the thesaurus of Terms. Only descriptors which in the Thesaurus of Terms have no BT and two or more levels of NT are selected as main entries for this index. The most generic terms appear along the left margin of the column with more specific terms below indented to the right at their respective hierarchical levels thus:

**Halides**

**Aluminium halides**

**Ammonium halides**

.....

Hydrogen bromide

.... Hydrobromic acid

.... Lithium bromide

.....

Chlorides

Calcium Chlorides

## 4.6 INIS

INIS Thesaurus was published in 1970 under the aegis of IAEA. Except for minor changes, terminology of the EURATOM Thesaurus, 1969 edition has been used in the formulation of INIS Thesaurus. It covers unclear physics and reactor technology in depth and related topies such as isotope technology, fabrication and use of unclear materials and instruments to a lesser degree. The terminoiogy is listed alphabetically but with each entry, the full 'word block' is displayed, giving all the terms associated with that particular entry. In order to attract the indexer's attention to the narrower terms in the word block, the reference indicators 'NT' are printed in bold face, upper case type. A full listing of narrower terms is provided in the Appendix to the thesaurus.

*BENZO1C ACID* [65; 65]

BTI monocarboxylic acids

BT2 Carboxylic acids

BT3 Organic acids

BT4 Organic compounds

***NTI ioglycamic acid***

RT benzolydroxamic acid

RT benzoyl peroxide

RT hypaque

The numbers in square brackets listed to the right of each descriptor represent frequencies of usage of that descriptor.

**Thesaurofacet**

The work of Ranganathan, which the CRG did so much to foster in the west, made its profound impact on thesaurus construction methods with the production of the Thesaurofacet by Jean Aitchison. The Thesaurofacet was originally planned as the fourth edition of the English Electric Faceted Classification for Engineering, but during its compilation it grew into a species of retrieval language so unlike its predecesors that it was deemed necessary to find a distinctive name for it.

When in the early sixties thesauri development dominated the scene, the English contribution to the modelling of thesauri resulted in Thesarofacet in 1970.

The thesauri of late fifties and early sixties were structured purely alphabetically. The limitations of the alphabetical arrangement led to the employment of classification aids, ranging from the broad to the detailed, and from auxiliary to integrated device. Thus the concept of thesaurofacet has come to stay as refinement of techniques in thesaurus construction, which has been evolving since the mid-sixties.

The Thesaurofacet is the integration of classification schedules and thesaurus. The full title **Thesaurofacet: a thesaurus and faceted classification for engineering and related subjects,'** indicates that there are two tools here, a classification and a thesaurus, but it is worth to note that the two have to be used if the best results are to be derived.

The Thesaurofacet the while of science and technology and hasborrowed parts of the London Graduate School's faceted classification of Business Studies for the relevant areas of management. Each term appears both in the thesaurus and in the schedules. In the schedules the term is displayed in the most appropriate facet and hierachy; the thesaurus on the other hand supplements the information by indicating alternative hierarchies and other relationships which cut across the classified arrangement. This thesaurus not only serves as an index to schedules but also controls synonyms and word forms in the manner of conventional thesauri.

The basic hierarchical relations of each index are displayed in the schedules of a faceted classification. Some examples will demonstrate the use of the two halves of the Thesaurofacet. Suppose we are asked for information on 'Echo sounders' when we look up in the thesaurus, we find :

Echo Sounders XEG

| | |
|---|---|
| RT | Echo Sounding |
| | Range finding |
| BT(A) | Electroacoustic devices |

If we turn to XEG in classified schedules we find :

| | |
|---|---|
| XC | *MECHANICAL MEASUREMENT* |
| XC2 | Dimensional measurements |
| XEB | Dept measurements |
| XEG | Echo sounders |

The thesaurus entry shows semantic relations to terms in other facets. BT(A)— Broader term (Auxiliary or additional) does not lie in the hierarchy to which 'Echo sounders' belongs. The net effect is to make classificatory structure behind cross references even more explicit.

Thus the thesaurofacet has become a multi-purpose tool, as easily applicable to shelf arrangement and conventional classified card catalogues as to coordinate indexing and computerised retrieval systems. With this systematic arrangement in the main part a thesaurus in fact becomes a classification system, reminiscent of the thesaurus method within the alphabetical index, including explicit indication of conceptual relationships for the control of synonyms and indication of additional broader concepts.

## 4.7 INSPEC Thesaurus

The thesaurus and unified classification, covering the fields of physics, electrical and electronics engineering, control engineering and computers, were developed in 1973. The thesaurus is linked to the unified classification in that each thesaurus term has been allocated one or more classification

codes. The second edition INSPEC Thesaurus : a thesaurus of terms for physics, eletrotechnology, computers and control was published in 1975. Besides UF (used for), NT (Narrower Term (s)), BT (Broader term (s)), it uses TT (Top term (s).to indicate the most general term in the hierarchy, CC (Classification Code), DI (Date of Input), PT (Prior Terms) to indicate terms used for the concept before establishment of the current perferred term. Examples of INSPEC Thesaurus listings are :

dynamic breaking

USE breaking

dynamic nuclear polarisation

UF dynamic nuclear polarization solid effect

NT CIDNP

Verhauser effect

BT magnetic double resonance

TT Resonance

RT nuclear polarisation

CC A 0758 A 3335D A 7670E

*DI* January 1977

PT magnetic double resonance.

## 4.8 UNESCO Thesaurus

UNESCO published its own **UNESCO Thesaurus** in 1977. It is intended to be used as a tool for indexing and retrieval of information processed through the Computerized Documentation System (CDS) of UNESCO as well as by any information / library service whose documentation coverage is closely connected or related to UNESCO's fields of activities. This was compiled by Jean Aitchison. The thesaurus covers major fields of knowledge

in outline, at least to the depth of subfields of UNISIST's draft 'Broad System of Ordering' (BSO). Within this framework, subjects of particular concern to UNESCO are covered in detail. The thesaurus is subtitled "A Structured List of Descriptors for Indexing and Retrieving Literature in the Fields of Education, Science, Culture and Communication."

The thesaurus is issued in 2 Volums Vol 1 Contains the Introduction, The classified thesaurus, the Permuted Index and Computer-generated the Hierarchical Dispiay of Terms. Vol 2 contains the alphabetical part of the Thesaurus. The thesaurus is essentially a species of Thesaurofacet, supplemented by computer-generated hierarchies in the INSPEC style.

A four-part structure of the thesaurus gives a multi-aspect approach to term interrelationship. The first part provides a classified arrangement of descriptors thus:

| | | |
|---|---|---|
| Z | 20/84 | Library and Information Science |
| Z | 45/56 | Information Processing |
| Z | 50 | Index Languages (By basic type) |
| Z | 50.01/15 | Controlled languages. |

The second part arranges the same descriptors in alphabetical order, with the conventional thesaurus relationship of synonym, narrower, broader and related terms as given below :

**Index languages** Z 50*. Z 02* x 50

| | |
|---|---|
| SN | An 'artificial language' used by documentation systems for the purposes of indexing |
| UF | Documentary languages<br>Indexing languages<br>Retrieval languages |
| RT | Controlled languages<br>Free language systems<br>Natural language systems |
| BT | Language Varieties |

TT Languages
Philology
RT Document description

The third part actually supplements the second part with a permuted index.

The fourth part holds the hierarchies of the generic relationships (BT/NT) in the alphabetical thesaurus. Here all the hierachical chains contained in the thesaurus are arranged in alphabetical order of the top terms (TT) of their hierarchy thus :

Languages
.... Index languages
.... Controlled languages
.... Alphabetical subject heading lists
.... Authority lists
.... Classification Systems
.....
.... Switching Languages
.... Thesauri.

Thus the UNESCO thesaurus differs from The Thesaurofacet in that, it on the one hand repeats in the alphabetical thesaurus relationships displayed in the schedules, on the other hand adds some relationships not shown in the classification.

## 4.9 ROOT Thesaurus

The British Standards Institution published their Root Thesaurus in 1981. It incorporates both an alphabetical list and subject display schedules and is printed in two volumes. The main sources of terms included in ROOT Thesaurus were Thesaurofacet, ISO Thesaurus and the list of terms used in the BSI Libraries for indexing. It contains 11,800 descriptors and 5,500 lead-in entries. The thesaurus is available both in printed version and on magnetic tape. It being a computer-based multi-lingual system was accompanied by a new set of indicators which replaced BT, NT and RT. A French version is available in computer printout and versions in other languages are also available. The second edition was published in 1985. The main branches of

technology are covered in depth and other fields such as social sciences and administration are included in less detail only to support the technological areas.

The Alphabetical list displays terms in conventional thesaurus format and serves the subject display as an index. The subject display presents the same terms in ciassifield order, fully cross-referenced. Thus there is practically.no fundamental difference between the ROOT Thesaurus and the Thesaurofacet. The Subject display is designed to be used in indexing and searching as the principal part of the thesaurus. It therfore includes related terms, scope notes and synonyms to descriptors. The main subject areas covered are as follows :

| | | | |
|---|---|---|---|
| A | General Section | 0 | Military technology |
| B | Measurements, testing and instruments | p | Production engineering |
| C/E | Science | Q | Transport engineering |
| F | Medical sciences | R | Construction |
| G | Environmental and Safety engineering | S | Mineral extraction technology |
| H | Agriculture forestry and fisheries | T | Materials |
| I | Food technology and tobaco | U | Metallurgy |
| J | Energy technology | V | Chemical technology |
| K | Electrotechnology | W | Wood, paper and texiles |
| L | Communication | X | Consumer goods and services |
| M | Control and Computer technology | Y | Administrative science |
| N | Mechanical engineering | Z | Social sciences and humanities |

Thus the main sequence is systematically arranged using case letter notation in the fashion of Thesaurofacet.

As stated earlier, the **ROOT** has been revelled with a set of following symbols

| Symbol | Significance |
|---|---|
| < | Broader term |
| > | Narrower term |

| | |
|---|---|
| — | Related term |
| *< | Broader term in an alternative hierarchy (Polyhierarchical relationship) |
| *> | Narrower term in an alternative hierarchy (Polyhierarchical relationship) |
| *— | Related term in an alternative array |
| = | Non-perferred synonym or Quaisynonym |
| 1 | Use (The term or combination of terms following the arrow should be used instead of the term preceding it) |
| + | It appears between terms which are used to synthesise a given concept. |
| / | It is used in the same sense as in UDC. |
| ** | Synthesised term. The term following the symbol is non-descriptor. |
| ** | The term (non-descriptor) following the symbolay be represented by the combination of descriptors preceding it. |
| [...] | Scope note. |
| (By ...) | Facet indicator to indicate the structure of the main sequence. It is not used for indexing purpose. |

## 4.10 Observations

Thus the ROOT thesaurus has emerged as the basis from which a whole series of thesauri and related products can be achieved. ROOT reminds one of Roget's original principle of a systematic list accompanied by an alphabetical one. In the alphabetical list nearly all of the information in the systematic display is repeated but with a change of format. There is in fact no fundamental difference between the BSI ROOT thesaurus and the thesarofacet as far as the structure is concerned.

Thesaurofacel represented a return to the pattern of Roget's Thesaurus. The ROOT Thesaurus has, of course, evolved *us* a logical development combining the tools for intellectual analysis developed over the past two

decades with technology needed to keep it current. To be more precise it has emerged as a generalized and modernised Thesaurofacet. Perhaps Roget has returned in a fuller torrent to ROOT.

## 4.11 Summary

This unit traces the origin, and development of thesaurus. Some thesauri for example, MeSH, TEST, INIS, Thesaurofacet. UNESCO thesaurus. ROOT Thesaurus are discussed in details. So far as structure in concerned there is no fundamental difference between BSI ROOT thesaurus and the Thesaurofacet.

## 4.12 References and Further Reading


1. Ailchison, J : Thesaurofacet : a multipurpose retrieval language tool. J. **Doc** 1970, 26(3). 187—203.
2. Chakraborly, A. R. and Chakrabarti, B. : Indexing : principles, processes, and products. World Press, 1984.
3. Foskett. A. C. The Subject approach to indormation. 5th ed. Library Associatior Publishing, 1996.
4. Foskett, D. J. Thesaurus. In Encyclopedia of library and information science ed. A. kent & ORs, vol. 30, 1980, 416.461.
5. London, G : Principles of Roget's thesaurus. **Rev. Int. Doc** 1965, 32(4), 146
6. Ramsden, M. J. : An introduction to index language construction. Clivt Bingley, 1974 (Programmed Text).
7. Vickery, B. C. : Thesaurus—a new word in documentation. J. **Doc** 1960, 16(4) 181-188.

# Unit 5 ❐ Bibliographic Record and Control

Structure



## 5.0 Objectives

After going through this unit you will be able to—

1. understand that the bibliographic records—their organisation and control have become crucial in modern times.
2. understand the endless interelationships among individual documents.
3. understand the difficulty in identifying intellectual responsibility of a document.
4. know the operations involved in bibliographic control.

## 5.1 Introduction

The written records of mankind were deliberately preserved and organised for the posterity. The invention of printing by movable types solved the problems of non-availability of multiple copies of a particular

document. The book has been performing its role and has been witnessing the whole range of basic changes occurring from time to time. The cumulative effects of such changes have tremendous impact over the intellectual advancement of human society and these have led to the growth and advancement of all kinds of human knowledge in all its ramifications. The nascent ideas, the insight into the past and present knowledge, as well as innovations in science and technology are documented in bibliographic records.

Libraries through centuries, have been making attempts to prepare lists, catalogues and indexes of such bibliographic records so that they can be retrieved and consulted in future. During seventeenth century journals came into being. A separate type of indexing service was ensued to list the writings of this genre. During twentieth century the non-print media appeared and stored in libraries. Separate retrieval systems and administrative agencies were created to serve the users of information from these media. The latest media are the electronic databases, on-line searching and networking both for bibliographic records and their indexes providing their own internal potential for automated self-indexing leading to bibliographic control and access.

Libraries have undergone a major change in the modern period. They have to play the role of comprehensive resource centres and librarians have to act as information providers. Therefore, libraries have to supplement the traditional carriers of information and intellectual contribution. At the same time the concept of single library services as stand alone institutions has become obsolete. The needs and requirements of the users in the present day libraries are so specific as well as sometimes so comprehensive on a particular field, that a single library, however large it may be, cannot serve the users' needs properly.

Because of the chaging needs and requirements of the users the information resources have become vital instruments for user's services. The bibliographic and non-bibliographic records—their organisation and control have become crucial in library management. The growth of information, information services, sources and systems and the electronic technology applied in these areas are the concern of the librarians as information scientists. The database agencies, commercial and noncommercial organisations having

specialised information management infrastructural facilities can manage and control the information resources.

Librarians can only attempt to keep track of the information sources, systems and retrieval services as well as documents as bibliographic records.

## 5.2 Bibliographic Data

Bibliographic data, as a whole, for a particular document and data elements in particular are the tools for bibliographic control. The data elements are the identification units of a particular document for the purpose of cataloguing and for preparation of bibliographies. Bibliographic data elements are the most significant for bibliographic record and biblographic control. Ronald Haggler mentions the significance and utility of data elements in the followings words :

"1. to identify a particular document uniquely in order to distinguish it from others (for example, a data of publication or a count of the number of pages in a printed book).
2. To show how two or more documents are associated with one another (for example, in that they have a common author or that one is a continuation or reprint of the other) and
3. to provide the basis for access points which enable a searcher to locale the record of a document in a file; these may relate to—
   (a) objective facts about the document itself (for example, its title or the fact that it constitutes the procesings of a conference) or
   (b) the topic (s) and concept (s) treated in its intellectual content; simplistically. the subject (s) of the document." Bibliographic data regarding three items mentioned above must be incorporated in bibliographic records both for catalogues and bibliographic tools. The primary value of the title of a document is as a description of the content of the document to the, users. The titie of a document in the bibliographic record is to identify the document itself under a specific tag, 'the name of a work.' Sometimes the term 'work' is ambiguous from the cataloguer's or bibliographer's point of view as well as from users' point of view. The intellectual content of a document is usually known as a 'work'. Bibliographic record and

control involve both the physical document with physical characteristics as well as the a intellectual content it embodies or the work. These two aspects cannot be separated in a document. But the users' approach may be either of the two aspects. For these reasons the identification of a particular document, listing of documents by cataloguers or bibliographers and bibliographical searching by users become complicated. Usually, the reason is that document entity (physical) gets precedence in bibliographic record because the primary function is not to interpret the document (work or intellectual entity) to users, but to provide access points to search the document in the form of one or more physical entities. Whatever term may be used to identify a document such as item, publication, edition or a work (intellectual content) the users are interested to get access to it for particular purpose.

## 5.3 Documents : Inter-relationships

The jobs of bibliographic record and control confront many complicated problems because of the endless interreationships among individual documents. All the relationships of any two or more documents are not available on records, may not be identified properly or these may not be anticipated. The bibliographic data on such relationships are not always found within the related documents themselves. The author of a particular document cannot imagine how many documents will be published related to that document during his lifetime and after his death by how many persons in various bibliographic formats having different authorship identity. Multiple shared authorship, multiple mixed authorship, various related and dependent works, various kinds of editions, compilations, translations, adapted works and the like may cause innumerable documents published and to be published in future. The identification of bibliographic record and management of bibliographic control of the documents have become practically impossible. In such cases the bibliographers have to investigate outside a concerned document in hand to ascertain all possible interrelationships.

Significant types of relationships may be summarised for basic bibliographic relationships. There may be many more but to begin with the

identification of relationships the following important points may be noted.

1. The physical formats of the same document may differ in the following ways:
   (a) Paper print, microcopy, CD-ROM, on-line access
   (b) Bound with hard cover, leather binding, paperback without illustrations, students edition, etc.
2. The bibliography entity may differ while the intellectual content remains the same in the following ways :
   (a) the same content may be published with different titles,
   (b) the same content republished by different publishers as reprints,
   (c) previously published book with a machine-readable disk.
3. The same content published in different editions may differ, such as,
   (a) simple print edition and illustrated edition, various literary forms, from drama to fiction or fiction to drama, revised text, supplementary materials, same content with editorial note, etc.,
   (b) research works or essays edited keeping the original form but with prefatory notes, additional updated materials, appendices, and the like,
   (c) the text as independent publication, same text in anthologies, compilations, collected works, critical editions, facsmile edition of rare books with or without editorial note, etc.,
   (d) different publications issued under a series with a composite or new title,
   (e) collection of several works of the same author published under a new title simultaneously with the individual titles,"
   (f) work being referred to another previously published work by a different author, a supplement to another work or a continuation of author work with different title.
   (g) keeping the title almost identical a work may be published with distinctly different intellectual content not identical to original work by the author, as edited, revised or completely revised edition written by other persons. In such cases the texts may vary from edition to edition.

The establishment of bibliographic relationships of documents is a complex and intriguing area for cataloguers and bibliographer because they have to inform the users about the intellectual contributions of the documents in different physical formats, variations in formats and they have to devise mechanisms to collocate all the documents showing their interrelationships and genealogical relations as information sources for getting access to the needed documents.

Bibliographic relationships are primarily shown in citations. references,bibliographies given at the end of documents, indexes, etc. as well as in library cataogues subject and author bibliographies in addition to documents themselves. The common areas of relationships may be revealed in the related documents, such as

(a) Various editions of the same work
(b) sequels, continuations, etc.
(c) items in the same series
(d) the same and similar works in various collections
(e) interdisciplinary and multidisciplinary works.
(f) citation analysis
(g) bibliographic references.

The bibliographic descriptions in annotated bibliography, data elements in library catalogues, bibliographic references, third level cataloguing of AACR-2. cross-reference citations and the like, have their own way of showing the above-mentioned and other relationships.

## 5.4 Intellectual Responsibility

The AACR-2R 88 uses the term "statement of responsibility" by giving less emphasis on what is known as "intellectual responsibility" because in most of the documents the intellectual content and its contributor cannot be easily identified except documents written by single authors. In machine-readable catalogues and bibliographies the physical description part has been standardized and there is no scope for annotation or critical evaluation of the intellectual content of the author or authors. It is not permitted to add such words as "by" or 'and' in square brackets to a statements of responsibility

if they do not appear in the work to be catalogued. However, Rule i.l F8 does not allow to add an explanatory word or phrase if the relationship between the title and the person or group named in the statement of responsibility, "is not clear" which means other than that of author and is not explicitly stated as such.

Despite its extensive instructions on construction of headings for persons the AACR 2R 88 makes explicit reference neither to personal **name authority control** nor to **authority control** in general. Both terms are absent from the entire text including the glossary. This situation makes it difficult to prepare bibliographic record and to make effective bibliographic control, AACR 2R 88 rules 22.2B2 and 22.2B3 sanction the practice of dividing authors' works among two or more headings. This practice goes against the collection of all the works together under one heading. This situation denies the fact that one individual produces several different types of works under several names. Although both rules instruct cataloguers to connect the various headings with explanatory references, but users cannot always be relied upon to take proper notice of these.

Intellectual responsibility has become increasingly difficult to define as more and more works are the contributions of a number of persons interacting in varied ways, whether as individuals or as members of corporate bodies. Intellectual contribution is authoritatively established in the work of a person created as singular contribution, and unquestionably the person is the sole owner of intellectual responsibility but in case if mixed authorship, shared authorship or otherwise multiple authorship, the extent of intellectual responsibility is very difficult to ascertain and most difficult to identify.

Intellectual responsibility of the surrogates as authors other than the principal authors has been accepted in cataloguing and bibliographic entries. The terms as editor, compiler, translator, etc. are being used to identify persons as holder of intellecting responsibility. The persons, so designated may perform any one or more of the functions as stated by Ronald Hagler. The functions are :

"1. cause a work to come into existence by conceiving its scope, general focus, and arrangement, then convincing others to do the actual writing.
2. reconstruct a deceased author's intended text by examining the existing editions, manuscripts, and other evidences.

3. add a commentary, glosses, footnotes bibliography, etc. (together called critical apparatus) to an existing work by someone else.
4. abridge, revise, paraphrase, bowdlerise, or otherwise modify an existing work by someone, or otherwise modify an existing work by someone else.
5. select for inclusion in a new publication material which was created for other purposes by one or many persons or bodies (his activity is often called compiling, as in the case of an anthology, but it may also be called editing, as in the case of conference proceedings)."

The description on the title page of a book records the intellecteal origin known as a statement of responsibility. Such description may be clear and explicit or sometimes vague and ambiguous. In the latter case some interpretation is required to determine which of the named persons have acted mainly to the creation of the intellectual content of the document for mentioning in the bibliographic record.

There is a basic difference between library catalogues and bibliographies. In library catalogues various access points or catalogue entries or references are used for access to a particular document in the library created by a person or persons with a focus to intellectual responsibility of any proportion. In bibliographies, not limited to the resources of a particular library, the name of a person, corporate body or an access tag is used as access or search point without much consideration regarding the intellectual responsibility of various types of 'authors' in a particular document.

Nancy Williamson in her 'Cataloguing and Bibliography : A comparative study of the Relationships as seen through their Principles and Practices '(1977) writes, "In many instances it was clear that the bibliographers were not concerned with establishing intellectual responsibility for a work, but rather with identifying it and locating it. Furthermore, they assumed that a name is a better means of identification than a title.....There was title in the sample of bibliographies which suggested that bibliographers considered title entries an improvement over author entries. In truth, there seemed to be a strong preference for the use of some other features of the work if possible, leaving title as the last possible choice ...". She further observes "In the light of the bibliographer's approach to authorship, the definition of authorship should be reexamined. In this context some thought rules toward

greater use of title entry is valid. With specific reference to works prepared under editorial direction and collections prepared by a compiler, reconsideration should be given to the possibility of accepting the name of the editor or compiler named on the title page of a work as viable entry."

The difference between catalogue entry and bibliographic entry is big in the manual catalogue system but in machine-readable catalogues and databases this difference does affect much. But in biblographic record and control the focus must be on bibliographic database and their organisation as well as searchability of access points and bibliographic collection.

## 5.5 Bibliographic Control

The bibliographic control is the basic apparatus for the knowledge management. The operations involved in bibliographic control are distinguished from the activities of cataloguing for identifying the needs and requirements of the immediate library users and access to documents in a library. Preparation of catalogues comes under technical services and access to bibliographic data comes under users' services or information services.

Bibliographic control makes a bridge between the two. It enables people to identify the documents, existence of documents useful for their' purposes and to know the nature and characteristics of each of the documents identification of a particular document as well as all the documents available on a particular topic, by a particular author or categorisation of documents for any particular purpose.

Bibliographic control is not just the preparation of list of documents or bibliographies, but it performs many functions. These are :

1. Identification of all documents produced in any physical format,
2. Identification of the individual works contained within a document or parts of the documents as distinct bibliographic entity including articles in journals, papers published in collections, units of anthologies, conference and seminar papers as individual entity in the proceedings, etc.,
3. Providing access points for multiple access of the documents and their parts, if any, under author, title, subject or any other access mechanism,
4. Producing and collecting all bibliographic records prepared according

to standard format of citation and collocation of documents as bibliographic instruments to help the users,

5. Providing locations of the documents, wherever they are available, to enable users for having access to them,
6. Indicating methods of access with mention to physical format, such as book, journal, online search, CD-ROM, disk and the like.

The operations of bibliographic control involve both the physical document and the intellectual content with the identity of intellectual responsibility, it incorporates. These two types of operations complicate the job of bibliographic control both for the bibliographers and the searchers or users. Bibliographic control is procedure through which the searches can have access to the documents in the form of one or more physical formats. The individual records in a bibliographic file is an organised sequence of document descriptions. All kinds of document description in machine—readable format are known as bibliographic records.

Uniformity and standardization of bibliographic records are essintial norms both for individual entry as well as bibliographic database. For adequate identification and consistency the following principles are to be followed :

1. Each bibliographic record must be unique in its identity and no two records are identical.
2. The bibliographic description of a record must be accessed by the searchable access points, such as author, title, subject, series, or any other heading,
3. The bibliographic description of a document must be presented in uniform and standardized manner so that their identity and interpretation can be unambiguous.

Standarization of bibliographic description, uniqueness in individual identity and information transfer mechanism ensure the following facilities :

1. Participant institutions can contribute to union catalogues ensuring resource sharing and cooperative cataloguing with the facility of broader bibliographic access,
2. Such institutions can download and upload records, and interchange bibliographic data of any category and of any physical format.

Bibliographic control covers a wide range of library and information techniques and disciplines. These may include cataloguing, systematic bibliography, indexing, abstracting, citation analysis, subject access by controlled vocabulary, personal name authority control, standardization of bibliographic format, bibliographic description, machine-readable cataloguing, indexing systems, thesaurus construction and other related areas.

## 5.6 Summary

This unit explains how with the arrival of non-print media libraries have undergone a major change. The concept of single library services as stand alone institutions has become obsolete. Bibliographic data have attained the most significant position for bibliographic record and bibliographic control. Because of the endless interrelationships among individual documents the situation for bibliographic control has become complicated. This unit also explains how intellectual responsibility has become increasingly difficult to define. It discusses the operations involved in bibliographic control.

## 5.7 Keywords

1. Access point : A name, term, code, etc. under which a bibliographic record is searched and identified.
2. Statement of responsibility : A statement, transcribed from this item being described, relating to persons responsible for intellectual or artistic content of the item, to corporate bodies from which the content emanates, or to persons or corporate bodies responsible for the performance of the content of the items.

## 5.8 Exercise

1. What is bibliographic data? Discuss the significance and utility of bibliographic data elements.
2. Discuss the difficulties in defining the intellectual responsibility.

3. What is the basic difference between library catalogues and bibliographies? Explain.
4. Discuss the operations involved in bibliographic Control.

## 5.9 References and Further Reading


1. Anglo-American Cataloguing rules. 2nd edition, 1988 Revision.
2. Downing, M- H. and Downing, D. H. : Introduction to cataloguing and classification. 6th ed. McFarland & Co, 1992.
3. Mahapatra, P. K. and Chakrabarti, B : Knowledge management in libraries. Ess Ess publications, 2002.
4. Tait, James A and Anderson, Douglas : Descriptive cataloguing, clive Bingley, 1071.

# Unit 6 ❐ Bibliographic Description and Access

Structure



## 6.0 Objectives

After going through this unit you will be able to—

(a) learn the vital issues in bibliographic description and access.

(b) understand an international standard format for bibliographic record.

(c) know the major components of a bibliographic record

(d) know the ISBD(M), ISDS and machine-readable storage and exchange format.

(e) have an idea about ISO 2709, UNIMARC, SGML, DTD and CCF.
(f) understand how the acrivities of bibliographic description and control pave the way to direct access to the content of a particular document.

## 6.1 Introduction

Bibliographic description typically includes areas such as title, and statement of responsibility, edition, physical description, series, standard number, etc. each of these areas is likely to be further divided into a number of elements, which vary according to the type of material. In a specialist sense, the term 'bibliographic description' is applied to the very detailed description of the physical and bibliographical characteristics, printing and publishing history and visual manifestation of early printed books.

Computer and telecommunication technologies greatly facilitate storage and access to information and documents. Computers can themselves promote such activities through compatibility of bibliographic records and transfer of such records to be performed automatically. Bibliographic and information services are very largely involved in manipulating records representing physical items, such as books, periodicals, periodical articles and other carriers of information in machine— readable storage and transfer systems. Standardization of bibliographic descriptions, identity of data elements, users, online searches, storage in users' computers are the vital issue in bibliographic description and access.

During past three decades much has been achieved in standardization and compatibility among bibliographic records as well as information transfer protocols. An international standard format for bibliographic record exchange has been devised and widely adopted. It is ISO 2709. Based on this standard, UNISIST Reference Manual provided a complete working manual for the international standardization of the form and content of computer-readable bibliographic descriptions prepared by abstracting and indexing services and others. The MARC exchange format and its various applications have also been widely adopted for establishing significant level of compatibility and standardization of exchange formats.

The contributions of IFLA (International Federation of Library

Associations and Institutions) for international standard bibliographic descriptions for a wide vaiiety of materials have become vital for achieving standardization of bibliographic formats of various documents. The international Standard Book Number (ISBN) and the Internationa) standard serial Number (ISBN) facilitate unique identification of these bibliographic items and interlining of different databases. Much work has been done for the areas of author auihority-control and subject authority control. Standards developed for programming languages, command languages, input / output devices, storage devices, character sets, telecommunication and networking links have made a break through in electronic storage and transfer of information.

Information storage and transfer systems are concerned with the processing of records representing some bibliographic item or document in various physical formats. The system to process bibliographic records with the help of electronic technologies stimulated the global activities of machine-readable storage and access to databases with standardized bibliographical description by means of achieving international compatibility of such records so that machine-readable records can be integrated and exchanged at international level. The ultimate objective was described by D. Anderson in "Universal Bibliographic Control : a long term policy— A plan for action (1974)," the purpose of which is :

"...to make universally and promptly available, in a form which is internationally acceptable, basic bibliographic data on all publications issued in all countries.

The concept of UBC (Universal Bibliographic Control) presupposes the creation of a network made up of component national parts, each of which covers a wide range of publishing and library activities, all integrated at the international level to form the total system."

A bibliographic record for a document can be subdivided into following four major components :

1. Description of the document or physical item itself with such elements as authors, title, publisher, and other essential elements to identify a particular item;
2. Highlighting the elements from complete description of the bibliographic record, such as personal or corporate author for intellectual responsibility

or title, etc. as search tag, to act as access points through which the record can be searched and retrieved;

3. A unique identifier for the document; and
4. Subject heading terms representing the subject matter of a document, which might be controlled vocabulary terms or authority file based on thesaurus, or words and phrases drawn from the document itself. In the last two decades and a half, IFLA has produced a series of "Standard bibliographic description" of documents in various physical formats to cover the first component. Such bibliographic descriptions of a particular physical format specify the data elements needed to completely describe a bibliographic item, the sequence in which they should be presented, and the punctuation conventions to be used with them. The bibliographic record format for description is flexible enough to allow an individual user to select those parts of the description needed to serve a particular puspose.

The ISBDs (International Standard Bibliographic Descriptions) for various physical formats have been accepted internationally as standard bibliographic description. All these have been recognised as the basis for rules for bibliographic description in AACR 2 for English speking countries as well as catalogue codes for some other countries. But these bibliographic descriptions deal with the content of the 'record', rather than with the representation of that content or with the content of the 'document' described. Two important steps towards Universal Bibliographic Control for the purpose of unique identification of a document were the adoption of a system for uniquely indentifying publications by International standard Book Number (ISBN) as well as International Standard Serial Number (ISSN) Since ISBNs or ISSNs are included in IFLA's standard bibliographic descriptions, these descriptions serve to identify documents uniquely as well as describing them.

The ISBDs give the bibliographic description of a document as a whole but these are not designed to contribute directly to the choice of access points. The access points for a document are to be decided by the library for its users. The standard description of a document may highlight some words to be used as access points. D. J. Hickey has given his observation in "Bibliographic Control in Theory" (1980) as stated below :

"Bibliographic information not included in the descriptive process can

hardly be selected as the basis for an access point; thus it is important, wherever possible, for the descriptive information to be identified and recorded in a standardized way so that all significant access points will be embedded in an appropriate context and can be uniformaly identified for retrieval purposes ... The ISBD ... lays the groundwork for possible future agreement about the way in which access points are chosen and their form of representation in a bibliographic file."

The information storage and transfer systems are concerned with the access to a particular document as well as a group of documents or access to information for such document or documents. Therefore, various access points for different types of needs and requirements of the users should be ensured in the mechanism of information storage, organisation and search strategy.

## 6.2 Bibliographic Records Format

The International Meeting of Cataloguing Experts at Copenhagen in 1969 adopted the following resolution :

"Efforts should be directed toward creating a system for the International exchange of information by which the standard bibliographic description of each publication would be established and distributed by a national agency in the country of origin ... The effectiveness of the system would depend upon the maximum standardization of the form and content of the bibliographic description."

Computer application in library and information services and networking necessitated international standardization of bibliographic description and record. Thus international standards were formulated. IFLA produced a series of "standard bibliographic descriptions" to specify the data elements needed to completely describe a bibliographic item and the sequence in which they should be presented. The International standard Bibliographic Description for Monographic Publications ISBD(M) was published in 1974 and served as the basis for rules of description of monographic materials in AACR 2. In 1975, the General International Standard Bibliographic Description, ISBD (G), was developed by agreement between JSC, AACR (Joint steering Committee for the Revision of AACR), members of the IFLA Committee on

Cataloguing and Chirman of IFLA's specialised ISBD working groups (See Table). The ISBD(G) serves as a single framework for the description of the types publication in all types of media, thereby ensuring a uniform approach in bibliographic description.

Other ISBDs have been formulated for serials, music, for non-book materials, for maps and for antiquarian materials and an 1SBD (CP) for "analytical" (Component part of works). The ISBD texts are subject to a five-yearly review. The Society of American Archivists (SAA), the Library of Congress, the Research Library Group all together formulated standard data elements used in Archives, Manuscripts and Records Repositions Information Systems as the Format for Archives and Manuscripts Materials in machine-readable form.

The unique identification of a document or publication becomes possible towards Universal Bibliographic Control by the adoption of the unique number as international Standard Book Number (ISBN) for monographs and International Standard Serial Number (ISSN) for monographs and International Standard Serial Number (ISSN) for serial publications. The ISBN number consists of four segments—the country segment, publisher segment, work segment, and Check digit-each separated by a hyhen. The ISBN was introduced in 1969, so earlier works will not have such numbers. Because inclusion of the ISBN is voluntary, not every work will include one. ISSN was introduced in the following year.

Publishers use ISBN for their computerised inventory control and financial accounting systems. An ISBN is ten digits in length. The ten digits used are the Arabic numberless 0—9, in the case of check digits only an X may sometimes be used. An ISSN uniquely identifies all issues of a serial publication while they have the same key-title. The International Data System of UNISIST coordinates the assignments of ISSNs and Key-titles by national agencies.

## 6.3 ISBD for Monographic Publications

International Standard Bibliographic Description was formulated to make available all bibliographic data of a particular document in any physical format. The bibliographic formats were designed for the purpose of :

1. standardisation of the format of descriptive portion of the bibliographic record for all types of documents,
2. making it easier to recognise data elements regardless of the language of their content,
3. making the data stored and retrieved in machine-readable format.

Obviously, the title page of a book remains the most important source for the data required for bibliographic description. ISBD(M) recognises the title page of a book as the main source of information. Different main sources have been recognised for each physical carrier of content in which there is no title page formally or its equivalent. But the arrangement of information on the main source no longer governs either the choice of which data elements are to be included in the record or their arrangement in the record. These are governed by the standardized format. Therefore, ISBDs specify other sources of data required for standardized bibliographic description. These sources include verso of the title page, cover and external sources depending on the type of data element involved.

The scope of ISBD(M) specifies requirements for the description of printed monographic publications with the order to the elements of the description, and specifies a system of punctuation for that description. The primary purpose of the ISBD(M) is to aid the international communication of bibliographic information by making records from different sources interchangeable, assisting in the interpretation of records across language barriers and assisting in the conversion of bibliographic records to machine-readable format.

ISBD(M) specifies eight areas for bibliographic data to be presented in the sequence given below :

| | |
|---|---|
| Area 1 | Title and Statement of Responsibility Area |
| Area 2 | Edition Area |
| Area 3 | Material-specific details Area |
| Area 4 | Publication, Distribution, etc. Area |
| Area 5 | Physical Description Area |
| Area 6 | Series Area |
| Area 7 | Note (s) Area |
| Area 8 | Standard Number, and Terms of Availability Area. |

There is enumeration of the elements is each of the above-mentioned areas.

## 6.4 International Serials Data System (ISDS)

The ISDS is an intergovernmental organisation established within the framework of the UNESCO-PGI programme. The aim of ISDS is to provide a reliable and up-to-date machine-readable database of world serials publications covering the full range of recorded knowledge and containing essential information for the identification and bibliographic control of serials. The ISDS functions as a network of national and regional centres established in various countries and coordinated by the International Centre located in Paris. The ISDS provides almost complete bibliographic control of the current serial publication of the world output.

This international serials directory provides all accurate ISSN. The ISDS has the sole responsibility for the assignment and control of ISSN. Records on all serials titles numbered by the ISDS network are published in the Register. Updating of the Register is ensured by the ISDS centres throughout the world. The ISDS record contains the ISSN, the key title, other forms of title, the abbreviated key title, the imprint, the issuing body, the classification number (UDC or DDC), all types of related titles, and other information. All ISDS data elements can be used as access points to the record.

The national and regional centres register systematically all new serials and title changes which occur. The records are forwarded to the International Centre of ISDS in machine-readable format. The records are varified, processed, added to the ISDS database. As one of the objectives of the ISDS is to have complete bibliographic control, the database also contains short records of serials of local interest or ephemeral nature which are excluded from general dissemination. Complete records of such serials are maintained at the respective national centres.

## 6.5 Machine-Readable Storage and Exchange Format

Machine-readable bibliographic record format for storage and exchange of bibliographic and information databases refers to the method of organising data so that each item of data can be unambiguously identified. In manual catalogues and printed bibliographies, the records and data elements are readily identified visually, by such characteristics as position, sequence and

spacing. In machine-readable records, there must be explicit way of determining the end of one data element and the beginning of another. In machine-readable format, the need for precision in determining the begining and end of every data element is essential and, therefore, must in accordance with internationally recognised codes that can indicate what kind of element or part of the element it is and how many characters it occupies.

The machine-readable record format of a particular bibiographic item, a book, serial, periodical article, etc. has following there major components :

1. The.structure of the record, which is physical representation of data on the machine-readable medium.
2. The content designators lables, which are the means of identifying the data elements or providing additional information about each.
3. The content of the record, which are the data elements themselves, such as names of authors, titles, and other elements to identify a particular document.

The most important record format is MARC (Machine-Readable Cataloguing). It refers to a specific set of conventions for representing an identified ensemble of bibliographic data. Initiated by the Library of Congress MARC I evolved in 1966 as a format for a machine-readable catalogue record, an instrument in the MARC Pilot project and the MARC Interim System that followed. In 1968 MARC II format was put into use and the full MARC System began functioning. It was proved useful in the United States. It became a model to be followed in other countries and changed a bit to accommodate local needs. USMARC was identified simply as MARC until other versions were developed, and then it was called LC-MARC. Finally, the name USMARC has come to distinguish it from the more than twenty other national versions (for example, CAN / MARC, UK MARC, Den MARC, etc.)

US MARC is based on ANSI standard Z 39.2, American Standard for Bibliographic Information Interchange (1971, revised 1985). The international version is ISO 2709 : 1996, **Information and Documentation—Format for Information Exchange.** The United States (USMARC) and Canada (CAN / MARC) have agreed on a "harmonised" format for two countries. The harmonised USMARC and CAN / MARC formats were published in a single edition in 1999 under a new name : MARC 21.

US MARC actually has five types of data. The following formats are currently defined :

bibliographic format—for encoding bibliographic data in records that are surrogates for information packages.

authority format—for encoding authority data collected in authority records created to help control the content of those surrogate record fields that are subject to authority control.

holdings format—for encoding data elements in holdings records that show the holdings and location data for information packages described in surrogate records.

community information format—for encoding data in records that contain information about events, programs, services, etc., so that these records can be integrated with bibliographic records.

classification data format—for encoding data elements related to classification numbers, their captions associated with them, their hierarchies, and the subject headings with which they correlate.

### 6.5.1 UNIMARC (UNIversal MARC)

It was developed in 1977 by the IFLA as a vehicle for interchange of MARC records between national bibliographic agencies. It conforms to ISO 2709, as does US MARC. The proliferation of national MARC formats necessitated the development of UNIMARC. At first it was thought that UNIMARC would act as a conversion format. In this capacity it requires that each national agency create a translator to change records from UNIMARC to the particular national format and vice versa. When a translator is in place, records can be translated to UNIMARC to be sent to other countries, and records received from other countries can be translated from UNIMARC to the national format. In addition to this use, a few national agencies that did not already have a MARC format have adopted UNIMARC as the standard in their countries. UNIMARC has been adopted by the countries of the European Community in order to produce "Unified Catalogues".

According to some UNIMARC differes from USMARC are that UNIMARC uses the ISO character set instead of USMARC's ALA Character set; UNIMARC allows for embedded fields (for example, an authority record

number may appear at the beginning of an access point field); and it deemphasises "main entry".

## 6.6 Standard Generalized Markup Language (SGML)

SGML is an international standard for document markup and conforms to ISO 8879 : 1996, **Information Processing—Text and Office Systems—Standard Generalized Markup Language.**It is a set of rules for designing markup languages that describe the structure of a document so that documents may be interchanged across computer platforms. It allows documents to be represented in such a way that text may be separated from structure without using a word processor. Structure means that the coding says : this is the title; this the first chapter; this is a section heading; this is a paragraph; this is a quoted statement, etc. SGML is flexible enough to define an infinite number of markup languages.

SGML defines data in terms of elements and attributes. An element is a particular unit from the text such as a title, a chapter title, a section heading, or a paragraph. An attribute gives particular information about an element (for example, giving the name of the thesaurus from which the subject term has been taken).

### 6.6.1 Document Type Definition (DTD)

A DTD is an SGML application. A DTD defines the structure of a particular type of documents; that is, it defines :

1. all elements that might be part of that particular document type.
2. element names and whether they are repeatable
3. in what order the elements should be placed
4. the contents of elements (in a general way, not specifically)
5. what kind of markup can be omitted.
6. Tag attributes and their default values.

### 6.6.2 Hyper Text Markup Language (HTML)

The HTML **DTD** was developed to enable the creation of web pages. It is a

basic markup language that allows almost anyone to be a web author. It provides for creation of simple structure, enables display of images, and provides for establishing links between documents.

## 6.7 Common Communication Format (CCF)

The International Symposium sponsored by UNESCO was held in Sicily in 1978 for the development of a common communication format that would meet the needs of libraries and documentation centres of all types and to bridge differences between the library community and information service community. Technically, the CCF represents a major advance in formatting because it provides for different levels of description within the same record. CCF differs from UNMARC by specifying no rules for. description, permitting minimal records, and introducing the concept of groups of fields called record 'segments'. Through the use of segments CCF records permit specific kinds of relationships to ecist between field, groups of fields, and records.

CCF has designated only eight data elements as mandatory. Five are required chiefly for computer processing of the record and most of these can be computer-generated. They include such elements as the unique record identifier, the name of the agency preparing the record, the date when the record was entered into a file, and so on. Another two or three elements are mandatory because they are considered necessary to identify the item uniquely. These include the language and script of the item and the title, and in case of a serial, the ISSN and key-title. All other data elements are optional.

The CCF has been designed to provide a standard format for the following purposes :

1. To permit the exchange of bibliographic records between groups of libraries and abstracting and indexing services.
2. To permit a bibliographic agency to use a single set of computer programmes to manipulate bibliographic records received from both libraries and abstracting and indexing services.
3. To serve as the basis of a format for an agency's own bibliographic data base by providing a list of useful data elements. To assist the development of individual systems, other UNESCO documentation

centres will provide implementation notes for CCF. and guide for AACR 2 Cataloguers who use CCF.

## 6.8 Bibliographic Access

The purpose of all the activities of bibliographic description and control is to provide the most direct access to the content of a particular document as well as a number of doucuments essential to the stated needs and requirement of the users. Every user wants the largest possible assemblege of potentially relevant documents and information to be available for any search. The compilers of catalogues, bibliographies and information sources are expected to build the files in a consistent and predictable manner in a system for accessing the descriptions of single documents and groups of documents. The bibliographic fields must be created as the best organised kind of information assess tools.

The objective of creating the files for users' search implies a mechanism form isolating a single record or a related group of them from among mant others. The mechanism involves two kinds of systems as follows :

(a) storage and arrangement of documents and information sources in a systematic way so that these can be retrieved at any point of time for any purpose and

(b) indexing of the documents and information sources available to the users for manipulation of multiple access.

These process consist of the following methods :

1. determination of useful access points for each document or information source from among its elements;
2. structuring and arranging these access points is a predictable way so that the users can make a purposeful and useful search;
3. linking the access points of document or information source for multiple search, and
4. interlinking of access points for related and surrogate materials.

The staff of the library or information centre are responsible for the organisation and maintenance of the above mentioned activities at input stage. Steps involved at the output side of the service are almost similar to those involved at input. The users submit request to the library or the

information centre and the staff members prepare search strategies for the requests of the users. After preparation of search strategies, these are matched in some way against the database of document representations. Document representations that match the search strategies and satisfy the logical requirements of the search are received from the database and delivered to the users.

The purpose of bibliographic access is primarly the organisation of information about the doucments the library or information centre collects. It is not just restricted to a particular library or information centre. The electronic storage media and communication technologies have made it possible to have global access in various formats, physical and electronic.

## 6.9 Summary

This unit describes the areas of bibliographic description. IFLA has produced a series of standard bibliographic description of documents. ISBDs have been formulated for book and non-book materials. This unit discusses the ISBD for monographic publications. It explains the machine-readable storage and exchange format. It discusses MARC formats, UNIMARC, SGML, DTD, and CCF. Finally it highlights the direct access to the content of a particular document.

## 6.10 Exercise

1. What is the objective of universal bibliographic control?
2. What purpose do ISBN and ISSN Serve?
3. Discuss the ISBD for monographic publications.
4. What are the major components of the machine-readable record format of a particular bibliographic ten?
5. What led to the development of UNIMARC?
6. What is SGML?
7. What is the purpose for designing CCF?

## 6.11 References and Further Reading


1. Chan, Lois Mai : Cataloging and classification. McGraw-Hill, 1981.
2. Downing, M. H. and Downing, D. H. : Introduction to cataloging and classification. 6th ed. McFarland & Co. 1992.
3. Hunter, Eric J : Computerised Cataloguing. Clive Binley, 1971.
4. Mahapatra, P. K. and Chakrabarti, B : Knowledge management in libraries. World Press, 2002.


1 The ISBD(G)

| *Area* | *Prescribed Preceding (or Enclosing) Punctuation for Elements* | *Element* |
|---|---|---|
| *Note : Each area, other than the first, is preceded bo a point, space, dash, space (.—).* | | |
| 1. Title and statement of responsibility area | [ ] | 1.1 Title proper |
| | | 1.2 General material designation |
| | = | 1.3 Parallel title |
| | : | 1.4 Other title information |
| | | 1.5 Statements of responsibility |
| | / | First statement |
| | ; | Subsequent statement |
| 2. Edition area | | 2.1 Edition statement |
| | = | 2.2 Parallel edition statement |
| | | 2.3. Statements of responsibility relating to the edition |
| | / | First statement |

| Area | *Prescribed Preceding (or Enclosing) Punctuation for Elements* | | *Element* |
|---|---|---|---|
| | | | Subsequent statement |
| | | 2.4 | Additional edition statement |
| | | 2.5 | Statements of responsibility following an additional edition statement |
| | / | | First statement |
| | ; | | Subsequent statement |
| 3. Material (or type of publication) specific details area | | | |
| 4. Publication, distribution, etc.. area | | 4.1 | Place of publication. distribution, etc. |
| | | | First place |
| | ; | | Subsequent place |
| | | 4.2 | Name of publisher. distributor, etc. |
| | [ ] | 4.3 | Statement of function of publisher, distributor. etc. |
| 4. Publication. distribution, etc., area | | 4.4 | Date of publication. distribution, etc. |
| | ( | 4.5 | Place of manufacture |
| | : | 4.6 | Name of manufacturer |
| | ,) | 4.7 | Date of manufacture |
| 5. Physical description area | | 5.1 | Specific material designation and extent of item |

| Area | Prescribed Preceding (or Enclosing) Punctuation for Elements | | Element |
|---|---|---|---|
| | : | 5.2 | Other physical details |
| | ; | 5.3 | Dimensions of item |
| | + | 5.4 | Accompanying material statement |
| 6. Series area | | 6.1 | Title proper of series |
| | = | 6.2 | Parallel title of series |
| *Note : A series statement is enclosed by parentheses. When there are two or more series statements, each is enclosed by parentheses.* | : | 6.3 | Other title information of series |
| | | 6.4 | Statements of responsibility relating to the series. |
| | / | | First statement |
| | ; | | Subsequent statement |
| | , | 6.5 | International Standard Serial Number of series |
| | ; | 6.6 | Numbering within series |
| | . | 6.7 | Enumeration and / or title of subseries |
| | = | 6.8 | Parallel title of subseries |
| | : | 6.9 | Other title information of subseries |
| | | 6.10 | Statements of responsibility relating to the subseries |
| | / | | First statement |
| | ; | | Subsequent statement |
| 6. Series area | | 6.11 | International Standard Serial Number of subseries |

| Area | Prescribed Preceding (or Enclosing) Punctuation for Elements | Element |
|---|---|---|
| | ; | 6.12 Numbering within subseries |
| 7. Note area | | |
| 8. Standard number (or alternative) and terms of availability area | | 8.1 Standard number (or alternative) |
| | = | 8.2 Key title |
| | : | 8.3 Terms of availability and/or price |
| | ( ) | 8.4 Qualification (in varying positions] |

SOURCE International Federation of Library Associations Working Group on the General International Bibliographic Description. *ISBD(G) : General International Standard Bibliographic Description : Annoted :* IFLA International Office for UBC 1977.

do rules 2.5 (tor books), 3.5 (for cartographic materials), 4.5 (for manuscripts), and so on..

As stated in the introduction to AACR 2, in choosing the appropriate chapter or chapters to be used in cataloging a particular document, the cataloger should start with the physical form of the item being cataloged, not the original or any previous form in which the work has been published. For example, a monographic publication in microform could be described according to the rules in Chapter 11 (for microforps), augmented by those in Chapter 2 (tor books, etc.). and Chapter 1 (general rules) when required.

## SOURCES OF INFORMATION

AACR 2 specifies sources of information to be used in describing a publication; in the case of a printed monograph, for example, such sources include the title page, the verso of the title page, etc. Of these, the source of bibliographic data to be given first preference as the source from which a bibliographic description is prepared is called the *chief source of information.* The rules identify a chief source of information, for each type of material.

**Chief Sources of Information**

| *Type of Material* | *Source* |
|---|---|
| Books, Pamphlets and printed sheets | Title page |
| Cartographic materials | a. Cartographic item itself |
| | b. Container or case, the cradle and stand of a globe, etc. |
| Manuscripts | Titie page |
| | Colophon |
| Music | Title page |
| Sound recordings | |
| Disc | Label |
| Tape (open reel-to-reel) | Reel and label |
| Tape cassette | Cassette and label |
| Tape cartridge | Cartridge and label |
| Roll | Label |
| Sound recording on film | Container and label |
| Motion pictures and videorecordings | Film itself and its container (if integral part of item) |
| Graphic materials | Item itself including any labels and the container |
| Machine-readable data files | Internal user label |
| Three-dimensional artefacts and realia | Object itself with any accompanying textual material and container |
| Microforms | Title frame |
| Serials (printed) | Title page |

In each chapter of AACR 2, prescribed sources of information for individual bibliographic areas are enumerated. Information taken from sources other than the prescribed ones is enclosed in brackets.

# Unit 7 ❐ Bibliographic Description of Non Print Materials

Structure



## 7.0 Objectives

This unit helps one

- to assess the role of Non Print Materials as an 'information source
- to identify variety of users of the non print materials.
- to indentify the common features in bibliographic description of books and other print materials with non-print materials.
- to provide guidelines to describe the non-print materials.

## 7.1 Introduction

The huge growth of information in this information society has led to many challenges, one of which is the importance of the non print materials as a major storage system. The development of mass-storage systems using

CD-ROMs, videodisks and other non-print formats has made it possible to input, store and display vast amount of information, including digitized information as well. A wide variety of methods is applied to the organization of and retrieval from these media. Retrieval of information, i.e., fast search and access systems is an important criteria in today's complex society. For this rapid access of information, the users need a proper search strategy, and for this reason an appropriate and detailed bibliographical description must be provided. This unit provides an overview of this aspect of non print materials.

## 7.2 Non Print Materials—Users

In the information society, we come across a wide range of specialized users who are adopt to the use of print as well as well as non print materials. The scope of the emergence of exciting and new creative forms for the presentation of scientific knowledge which contribute to the progress of education has led to the advent of a wide range ofusers. The self education and learning process in many disciplines has moreover added to this tendency. The users of the non print materials fall into two groups—institutions and individuals. In many institutions, there are a variety of non print materials. Moreover, many institutions are members of many libraries, which store a large number of non print materials. The needs of non print materials users are focused on three questions which have direct relevance to libraries—(a) what documents and equipment exist? (b) how does the user gain access to documents and equipment? (c) how does he or she use the documents and equipment provided? These questions are readily answered by the individual and institutional producers who produces the document and which ultimately reaches the user, individual or institutional, who takes the finished product in order to extract information from it.

## 7.3 Bibliographic Description of Non Print Materials

Many non print resources are common as their print counterparts in the collections of libraries, but the standards for bibliographic description of new print materials continue to evolve. Without a bibliographic record, access

is severly limited or non existent. Since nonprint resources may non be browsed in the manner of books and other print resources, reliance on bibliographic record for access to their contents is particularly important. The rationale provided by the Online Audiovisual Cataloguers (OLAC) is important—"Full and standardized bibliographic description of nonprint resources facilitates a heighetened awareness of the full range of information resources a library offers its user population; a consistent means *for* both the local and remote user of the catalogue to search the entire collection through a single interface; identification of material that represents a significant expenditure of library funds; international efforts in cooperative cataloguing by sharing bibliographic records in the utility databases." The manner in which a library publicizes and makes available the non print resources through its online catalogue and web page is very important.

AACR 2R(1998) outlines three possible levels of bibliographic description that may be used to describe resources. The level of cataloguing depends on a particular library's needs, available staff, and the emphasis given to a specific collection and / or material type. The first level of description, (Rule I. ODI), is comparable to minimal level cataloguing that is used for items of a temporary nature that libraries may wish to make available in their online catalogues, including reserve resources, recent acquisitions that have not yet been catalogued, etc., or it may be used by libraries with limited staff and monetary resources that have a minimum of time to provide descriptive treatment for nonprint resources. The second level of description, (Rule 1.0D2) is fuller and is probably the closest of the three in the degree of detail provided by most libraries in their bibliographic records. The third level of description (Rule 1.0D3) is the fullest and most complete, and can be compared to enhanced cataloguing. This type of description can also be purchased through a vender service. An item given third level descriptive treatment would include all elements outlined in the rules that one applicable to the item being described.

The framework of chapter 1 (AACR 2R) discusses general rules for description, as a basis to discuss the physical description of non-print resources in the foflowing points :

The physical description of any item should be based on the first instance on the chapter dealing with the class of materials to which that item belongs.

1.1 Title and statement of responsibility area—

1.1.1 Title proper

1.1.2 [Genera Material Designation]

1.1.3 Other title information

1.1.4 Statement of responsibility

1.2 Edition area—

1.2.1 Edition statement

1.3 Material specific details area.

1.4 Publication, distribution etc. area—

1.4.1 Place of publication, distribution, etc.

1.4.2 Name of publisher, distributors

1.4.3 Statement of function of publisher, distributor, etc.

1.4.4 Date of publication, distribution, etc.

1.4.5 Place of manufacture, name of manufacturer, date of manufacture, etc.

1.5 Physical description area—

1.5.1 Extent of an item (including specific material designation)

1.5.2 Other physical details

1.5.3 Dimensions

1.5.4 Accompanying material.

1.6 Series area—

1.6.1 Title proper of series

1.6.2 Numbering within series

1.7 Note area—

1.8 Standard number and terms of availability area—

1.8.1 Standard number

1.8.2 Terms of availability.

The above mentioned framework gives all the elements that are required to describe non print materials and also assigns an order to these elements.

In case of computerised bibliographic description, the following information may be used to create bibliographic records for non print resources :

- Chief source of Information
- Prescribed sources of Information
- Choice of Main Entry
- International Standard Book Number
- Cataloguing Source
- Language Code
- Geographic Information
- Library of Congress Call Number
- Title
- Title Variations
- General Material Designation (GMD)
- Statement of Responsibility
- Edition
- Place of Publication and / or Distribution, or Manufacture, etc.
- Name of Publishers and / or Distributors
- Date of Publication, Distribution, Copyright, Manufacture, etc.
- Physical Description
- Series
- Notes
- Subject Access

- Added Entries (Personal and / or Corporate Names, Title Added Entries)
- Classification.

For nonprint materials' description, we find sources of information as follows:

1) The material itself, including the NPM which forms the integral part of the item, for example, videocassettes.
2) The accompanying print-on-paper text describing the contents of audiocassettes, videocassettes etc.
3) Other sources, which act as referral guides for information.

## 7.4 Standards for Bibliographic Description

The ISBD for non print materials prescribes the following structure for standards for bibliographic description—

1) Title and statement of responsibility area—

Title is given by the auther or publisher or producer of nonprint materials, which may be of three types :

**uniform title**—when variants of titles exist;

**supplied title**—when title for a non print material does not exist, but has to be supplied;

**collective title**—when the individual item does not form part of different groups of materials.

Responsibility for the creation of NPMs—There can be several credits involved in the creation of a document.

Physical description of the materials—This varies from material to material in case of now print resources. This includes the range of the item, physical details, such as size, demension and other characteristics.

## 7.5 Physical Description of Non Print Materials— Examples

The most important part in describing a non print material is its physical description, because it varies from material to material.

The features of these materials with examples are given below :

(i) **Cartographic Materials**: These are an important part of the collection of many libraries. Cartographic materials are those meterials that represent the whole or part of the earth or any celestial body at any scale. These include two and three dimensional maps and plans, vcelestial, aeronautical charts, atlases, globes, aerial photographs, etc. The physical description includes colour, physical medium (paper, textile, etc.), type of reproduction (facsimile), type of mounting for maps (mounted on cloth), sheet size for maps when a map is printed with an outer cover within which it is intended to be folded or if the sheet itself contains a panel or section designed to appear on the outside when the sheet is folded. An example with the USMARC format for this area is given below :

300 1 map on 7 sheets : Col. : vellum; 140 x 196 cm., on sheets 82 x 108 cm., folded to 27 x 22 cm.

(ii) **Sound Recordings** :

Sound recording is a recording on which sound vibrations have been registered by mechanical or electrical means so that the sound may be reproduced. The various physical peculiarities of sound recordings are speed (16 rpm), configuration of playback channels (mono, quadraphonic, stereo), groove width / groove pitch (microgroove), dimensions, tape configuration (full track, half track), kind of disc, cylinder or tape (master tape), kind of material (plastic), kind of cutting (lateral or combined), special playback characteristics (CX encoded, digital recording), capture and storage technique (digital / direct storage) etc, duration (90 min.) etc.

(iii) **Motion Pictures & Video Recordings**: These are recognized as a powerful medium for communication and education. AACR 2R defines motion pictures as "a length of film, with or without recorded sounded, bearing a sequence of images that create the illusion of movement when projected in rapid succession". and a videorecording as "a recording on which visual images, usually in motion and accompanied by sound, have been registered; designed for playback by means of a television set". In the physical characteristics area, the extent of the material, the duration (playing time,

e.g. 30 min.), sound characteristics (sound or silent), color characteristics (b & w, col..) and dimensions (width of motion picture films in millimeters or width of videocassettes in inches) are recorded.

e.g. 300 1 film reel (52 min.) : si, b & w

300 2 videocassettes (ca. 180 min.) : sd., col.

(iv) **Electronic Resources** : Electronic resources (formerly "computer files") are an essential collection of modern libraries. These typically receive full level cataloguing. These are defined as a file (data and / or programs) encoded for manipulation by computer. The physical description includes extent (number and Specific Material Designation of physical parts of an item), sound characteristics (provided only when an electronic resource has sound capability), color characteristics (only when a program is in color or has portions in color), dimensions (size, width etc.), additional physical characteristics (number of sides, density etc.). Example (in MARC format) is shown below :

300 2 computer discs : sd., col., double sided, double density; ½ in.

(v) **Kits** : Kits are a set or collection of various material types intended for use as a unit. It usually contains two or more categories of materials, no one of which is predominant constituent of the item. Kits can be described in a variety of ways. The physical description may be general or specific, or a multilevel description may be provided to describe each component part in full detail. Component parts may be described as "various pieces" for resources with a large number of different parts. A description of characteristics other than extent or dimensions are also provided.
Examples :—

**Specific description**

300 2 sound cassettes (ca. 49 min. ea.) : analog, $1^7/_8$ ips + flash cards + 2 puppets in container 40 x 45 x 40 cm.

**Multilevel description**

300 2 scrend cassettes (ca. 49 min. ea.) : analog, $1^7/o$ ips.

300 1 videocassette (30 min.) : sd., cal.; $^1/_2$ in.

300 xi, 200p. : ill.; 28 cm. + 1 teacher's guide + 1 student workbook (100 p. : ill. ; 28 cm.)

Information is to be provided on size, width, etc, as appropriate,

(vi) **Microforms :**

Microforms enables libraries to provide access to large collections that may no longer be available in print, or when print counterparts are fragile and handling is not advised. They allow libraries to provide access to older runs of serials or other large collections that could not be housed as easily if available in print. Microforms is a generic term for any medium, transparent or opaque, bearing microimages. Microforms consists of microfilm, microfilm, microfiche, microopaques and aperture cards. For physical characteristics, information is provided on extent, negative characteristics, color, illustrative characteristics, dimensions etc. For microfiche, the number of frames is important. Detailed information on type of illustrations, colored illustiations and size is important. In MARC format, this is also represented in the 300 field.

e.g., 300 2 fiche : col. ill., maps ; 35 min. The standard size for fiche is 10.5 (height) x 14.8 (width) cm. Dimensions are not provided for fiche unless larger than the standard size. The width of film for reels is provided in millimeters (min.).

For all the above nonprint resources, the physical description is provided in MARC format in 300 field, and also accompanying materials are given as and when needed with the physical descriptions.

## 7.6 Exercise

1. What are the general rules for description of non print materials as given in AACR 2R?
2. Discuss the physical description of microforms and sound recordings as given in AACR 2R.

## 7.7 References


1. Andrew, J.—Developments in the organisation of non-book materials. London : The Library Association, 1977.
2. Fothergill, R. and Butchart, I.—Non-book materials in libraries. London : Clive Bingley, 1990.
3. Weber, Mary Beth—Cataloging now print and internet resources. London : Neal-Schuman, 2002.

# Unit 8 ❐ Metadata

Structure



## 8.0 Objectives

After going through this unit you will be able to :—

1. know that metadata emergesto cope with the problem of representing digital information.
2. know a number of metadata standards that have been evolved.
3. know the different types of metadata.
4. understand the advantages of using metadata.

## 8.1 Introduction

Metadata was initially applied only to electronic information and network information. Since the very beginning of computerised data processing there has been a tendency for ever growing amounts of data to be processed and stored by computers. Probably not by accident, the modern computerised data processing was also referred to as mass-data processing. Especially in the environment of the so-called large-scale information systems, as for example, statistical ones, there was an ever-growing necessity to find the ways and means how to handle those repidly expanding amounts of statistical data. The technological advancement and users needs finally led not only to introduction of very-large databases and their distribution to the database networks but also to the necessity to invent and introduce the particular tools for handling their content that is, data and information in the form of data and information on another source and / or object data and information, which started to be regarded as metadata. Whether we call it cataloguing, indexing, or metadata, the concept is a familiar one for information professionals. While the metadata has become a buzzword in the information business, the concept important for both authors and seekers of electronic information. The connotation of the term has been expanded to include all practices for information representation because metadata has become a very fashionable term and the Internet is increasingly becoming a platform for accessing a wide variety of information.

## 8.2 Definitions

Metadata nowadays can be defined in two different ways : one is narrower in scope, implying descriptions provided for networked information and digital information by following a standard or framework (e.g. Dublin Core) that is specifically created for this purpose. The other is broader in coverage, including cataloguing and indexing data created for any kind of documents by using traditional methods for describing and organising information. In this sense, for example, Cataloging data produced with Dewey Decimal Classification and Anglo-American Cataloguing Rules / Machine Readable Cataloguing (AACR / MARC) is also regarded as metadata.

Following are the few definitions of metadata :

In the broadest sense metadata is data about data. The same thing can be both data and metadata, depending on one's perspective. The Macquaric Dictionary defines the prefix "meta"—as meaning "among", "together with", "after" or "behind". That suggests the idea of a "fellow traveller" : that metadata is not fully-fledged data, but it is a kind of fellow-traveller with data, supporting it from the sidelines. Metadata, literally "data about data", is an increasingly ubiquitous term that is understood in different ways by the diverse professional communities that design, create, describe, preserve and use of information system and resources. Metadata can also be defined as "data associated with objects which relieves their potential users of having to have full advanced knowledge of their existence and characteristics". In other words, standard bibliographic information, indexing terms, and abstracts are all surrogates for the original material, hence metadata. Gregory Wool believes that metadata provision is essentially an extension of the traditional cataloguing process.

## 8.3 Purpose of Metadata

Metadata serves many important purposes, including Data browsing, Data transfer and Data documentation. Metadata can be organised into several levels ranging from a simple listing of basic information about available data to detailed documentation about an individual dataset. At a fundamental level, metadata may support the creation of an inventory of the data holdings of a state or local government agency. Metadata is also important in the creation of a spatial data clearinghouse, where potential users can search to find data they need for their intended application. At a more detailed level, metadata may be considered as insurance. Metadata insures that potential data users can make an informed decision about whether data are appropriate for the intended use. Metadata also insures that the data holdings of an agency are well documented and the that the agencies are not vulnerable to losing all the knowledge about their data when key employees retire or accept other jobs.

## 8.4 Types of Metadata

There are at least four types of metadata described below :—

A. **Descriptive Metadata**:Descriptive metadata describe the intellectual content and associations of a document or resource in a way that facilitates search, identification and collection of information contained within or exemplified by the resource, for example, Dublin Core.

B. **Subject Metadata** :Subject metadata can facilitate effective searching. They can also be used to organise the browsing structure of the gateway.

Keywords, Classification Codes, Classification System, Terms from Thesauri, Subject Heading.

C. **Structural Metadata**: It defines the physical structure of a complex digital entity to facilitate navigation, information, retrieval and display.

Provides information about the internal structure of resources including page, section, chapter numbering, indexes, and table of contents.

Describes relationship among materials (for example, Photograph B was included in manuscript A).

Binds the related files and scripts (for example, File A is the JPEG format of the archival image File B).

A simple example might be metadata that facilitate page-turning functionally.

A more complex example world use metadata to link searches run on encoded text to the images of the pages where hits occur.

D. **Administrative Metadata**: Administrative metadata encompasses a variety of data related to viewing, interpretation, use and management of digital objects over time.

Includes technical data on creation and quality control.

Includes right management, access control and use reuqirements.

Preservation action information.

## 8.5 Characteristics of Metadata

Some of the characteristics of metadata are the following :—

1. It is readable by both humans and machines.
2. Metadata takes a variety of forms, both specialised and general and many be part of a larger framework.
3. New metadata sets will develop as the networked information infrastructure matures.
4. Different communities will propose, design and maintain different types of metadata.

## 8.6 Examples of Metadata Standards

Although the concept of metadata is relatively new in the realm of information representation, a number of metadata standards have been developed and more are being conceived. Dublin Core and Resource Description Framework are some examples of metadata standards.

### 8.6.1 Dublin Core (DC)

Dublin Core is a simple content description model for electronic resources. DC is a joint effort among experts from the library world, the networking and digital library research communities, and a cariety of content specialists in a series of focussed, invitational workshops called the Dublin Core Workshop series. Dublin Core consists of 15 elements :

1. Title
2. Creator
3. Subject and Keywords
4. Description
5. Publisher
6. Contributor
7. Date

8. Type
9. Format
10. Indentifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights Management

As Weibel Stuart explained : "... the Dublin Core is not intended to replace richer description models such as AACR 2/MARC Cataloging, but rather to provide a core set of description elements that can be used by eatalogers or non-catalogers for simple resource description."

In practice, Dublin Core is mainly used to provide simple descriptions for networked like websites. It seems to be the best known standard among all the metadata models, moreover, Dublin Core is the prototype application that drove the development of the Resource Description Framework (RDF) in the World WideWeb consrtium (W 3C). Developers of the Dublin Core are closely involved in the W 3C's RDF effort.

### 8.6.2 Resource Description Framework (RDF)

Resource Description Framework was developed under the auspices of the W3C, becoming a W3C recommendation in February 1999. Shafar, Keith explained;

"RDF is an infrastructure for encoding, modeling and exchanging metadata. At the heart of RDF is a simple three-part model : metadata is about a resource, the resource has one or more properties, and each property has a value."

RDF uses Extensible Markup Manguage (XML) as the transfer syntax. RDF is evolving to support the many different metadata needs of vendors and information providers. Unlike Dublin Core, RDF does not specify the particular elements the framework should have. Rather, it lets users choose

and define the specifications within its framework based on their needs. RDF is therefore a foundation for processing metadata.

Other metadata standards such as Digital Object Identifier (DOI) have been or an being developed for describing digital information on Net.

## 8.7 Advantages of Metadata

The Hiawatha Island Software Company, Inc. : suggests the following advantages of using metadata :

1. Time—tested and reliable : Metadata follows universal bibliographic standards established for the card catalogue and developed by technology pioneers.
2. No. database pointer records : Using metadata is similar to querying a database using document descriptors. But unlike a database, which uses pointer records to direct the document location, metadata lies inside a document and remains there no matter where the document moves.
3. Support for all media formats and document types : Metadata allows the users to catalogue all types of digital resources.
4. Cost-effectiveness and performance : Metadata systems require minimal user programming, and schedule updates for growing document collections. Highspeed document retrieval is measured in milliseconds.
5. Integration with existing methods : Metadata works with existing archival procedures, search engines, and search portals. In eddition, an automated metadata system can index any files which can be accessed from the administrating server.
6. Mainpulate metadata easily : Metadata can be added and removed from documents. In addition, unlimited data can be added to meta tags. Meta tags can be customised for organisation, and carry unlimited potential for search and retrieval accuracy.
7. Manage maturing document collections : Metadata maintains growing document collections and maturing documents.
8. Ease the transition from paper to e-files : Many organisations are

using specialised scanning techniques to convert paper to electronic files. Metadata eases this transition by search-enabling scanned documents.

9. Language independence : Other knowledge management systems struggle to establish multiple language support for global organisations. Metadata, using universal bibliographic standards, features taken. Tokens overcome archive and retrieval inconsistencies caused by semantics and multiple languages. A token represents the idea-not group of words that describe it.
10. No implementation time or training required : Since Metadata integrates readily with existing methods, no trading is required for organisation members. Similarly, when an organisation beings using metadata, no implementation time or "transition" time is necessary.

## 8.8 Who should create Metadata?

The answer to this varies by discipline, but it is almost always a cooporative effort. Many projects have found that it is more efficient to have library cataloguers or other information professionals create the descriptive metadata, because the authors or creators of the data donot have the time or skills. The researcher may create a skeleton, completing the elements that can be supplied most readily. Then the results may be supplemented or reviewed by the cataloguer for consistency.

## 8.9 How to create Metadata?

Normally, a single disk file for each metadata record can be created. Thus one disk file describes one data set. Then some tool is used to enter information into this disk file so that the metadata conforms to the standard. The following steps may be followed :

Assemble information about the data set.

Create a digital file containing the metadata, properly arranged.

Check the syntactical structure of the file. Modify the arrangement of

information and repeat untill the syntactical structure is correct.

Revised the content of the metadata, verifying that the information describes the subject data completely and correctly.

## 8.10 Conclusion

Metadata is like interest it accrues over time. Carefully designed metadata results in the best information management in the short and long-term. If thorough, consistent metadata has been created, it is possible to conceive it being used in an infinite number of new ways to meet the' needs of non-traditional users, for multi-versioning and for data mining. Librarians have used the concepts behind metadata for generations. Metadata, information about a document's content, brings the power of a database to electronic documents. Just as we could search'for an item using the title, subject or author in the library card catalogue, we can search for electronic documents using an unlimited number of fields, or met tags. Metadata is stored within a document. Although we cannot see it when we open the file using an information browser, metadata remains inside a file no matterwhere the file moves. Metadata, no matter how the current and future practices may change, is ultimately intended to facilitate the representation of digital information so that it can be more effectively retrieved later.

## 8.11 Summary

This unit describes some of the basic facts about metadata like definitions, purpose of metadata, types of metadata, characteristics and examples of metadata standards. It explains Dublin Core as content designator model for electronic resources. It mentions other metadata standards like resource Description Framework and Digital object Identifier. This unit also explains who is competent to create metadata and how is metadata created.

## 8.12 Exercise

1. What is metadata? Explain its purpose.

2. Describe different types of metadata.
3. What are the advantages of using metadata?
4. Illustrate two examples of metadata standards.

## 8.13 References and Further Study


1. Chu, Heting : Information representation and retrieval in the digital age. Thomas H. Hogna, Sr. 2003.
2. Dempsey, L and Heery, R : Metadata : A Current view of practice and issues. **Journal of Documentation** 1998, 54(2), 145-172.
3. Shafer, Keith : Mantis project provides a toolkit for cataloging. OCLC **Newsletter** 1998, 21-23.
4. Weilbel. Stuart : The Dublin Core : A simple contenet description model for electronic resources. **Bulletin of the American Society for Indormation Science,** 1997, 24(1), 9-11.
5. Wool, Gregory : a meditation on metadata. **The Serials Librarian,** 1998, 33(12), 167-178.
6. The Dublin Core Metadata Element Set, ANSI/NISO 239-85–2012. ISSN : 1041-5635; ISBN-13 : 978-1-937522-14-8, (Lttp : dublincore.org/).

# Unit 9 ❐ Indexing in Theory and Practice

Structure

## 9.0 Objectives

Indexing has been a widely adopted method for information representation. Information, recorded in different types of documents needs to be represented before it can be retrieved. After reading this unit you will be able to :—

1. understand what indexing is;
2. understand how indexes provide a vital link between stared bibliographic data and its later retrieval;
3. know two basic categories of indexing systems : pre-and post-coordinate.
4. understand the different indexing systems and their mechanism; merits and demerits.
5. understand the general theory of indexing
7. evaluate indexes in order to determine how good they are,

## 9.1 Introduction

The term 'Index' is derived from the Latin root 'Indicare' which means to indicate. An alphabetical list of topics, names of persons, places, etc., mentioned in a book or series of books, indicating at what place or places they appear in the source document's usually by page number (sometimes with an additional symbol denoting a position on a page) but sometimes by section or entry number). In contemporary usage, a finding-aid to the position of material in a library collection. Used more or less as a synonym for catalogue. However, although the principles of analysis used are identical, an index entry merely locates a subject, whilst a catalogue entry also includes descriptive specification of a document concerned with the subject.

Analysing the contents of a document (book, pamphlet, audio-visual or machine-readable item, or collection of documents) and translating the results of the analysis into terms for use in an index—an organised grouping of such terms to allow location and retrieval of information. An index is a gate and not a surrogate of the original document.

## 9.2 Indexing

Indexing is the process of creating entries in an index. It is a service operation. The essential operations in this process are : (a) scanning the collection; (b) analysing its content; (c) tagging discrete items in the collection with appropriate identifiers, and (d) adding to each identifier the precise location within the collection where the item occures, so that it may be retained. Indexing is neither strictly an art nor a science, but mixes characteristics of each. It is an art in the sense that it requires sensitivity, intuition and taste. It is a science because it appreciates the use of rules and patterns. But these rules are empirical and do not lead us to universally accepted results as should be in science. Thus strictly speaking indexing is not a science. Similarly it forfeits its claim as true art as it prohibits the projection of individuality and the contribution of creative ideas during the process.

## 9.3 Need for a theory

The role of any science is to explain, control and predict behaviour. In information science, we are concerned with the behaviour of information, and we behaviour of information, and we seek to investigate the properties of information, of forces that govern the flow of information and the techniques for processing information for optimal storage, retrieval and dissemination. In the course of information science investigations, many specific facts concerning information transfer have been accumulated. These have not been propertly integrated into a fully comprehensive theory that will enable us to explain, predict and control the transfer of information.

Indexing is an important part of the information storage and retrieval process. It uses terms (that is, words or phrases), either derived or assigned, to represent important facets of the original document. We know how to do indexing, but we lack knowledge that will enable us to predict accurately the retrieval effectiveness of different indexes. In order to make indexes and indexing process more effective,we must base our work on scientific knowledge and not rely solely on existing practice. In a word, we need a theory to help us explain, predict and control the dissemination of information,

and this theory must encompass indexing and indexes.

A few researchers have started researching on different aspects of a theory of indexing for information retrieval. We shall take an initial look at the theory building efforts of four selected pioneers.

## 9.4 Theory of index terminology

An early contribution toward a theory of indexing was made by Fredrick Jonker, a pioneer in the field of information science and information storage and retrieval. Jonker is the inventor of the Termatrex system of optical concidence cards and other devices for use in information retrieval systems. He formulated a general theory of indexing as expressed by a "Terminological continum" and a "Connective Continum." The theory explores the nature of formalised index languages, the functions they perform and the relationship to living languages.

The terminological continum consists of a general law expressing the relationship between the size of the vocabulary and the specificity by which a concept can be described. Whereas terminological continum deals with words and not their relationships, the connective continum deals with the connections between terms. The variable involved is the average number of words per index term. These connections between words may be based either on the principle of post-coordination as embodied in Jonker's optical coincidence cards, on hierarchical classification, or on a combination of both whereas the terminological continum delas with words and not their relationship, the connective continum deals with the connections between terms. The variable involved is the average number of words for index term. These connections between words may be based wither on the principle of post-coordination as embodied in Jonker's optical coincidence cards; on hierarchical classification; or on coincidence of both.

### 9.4.1 Lawrence Heilprin

Heilprin is another leader in the field of information science who has devoted his considerable energies toward improving the scientific basis of information systems. Soon after Jonker published his paper on "The

Descriptive Continum, A Generalised theory of Indexing," Heiprin attempted to provide this model with a more formal structure and mathematical basis. He introduced the concept of a search path corresponding to the number of available paths that led him from questions to documents. He also introduced the concept of noise in the system which results from any deviation from theoretical limits of indexing permutability or hierarchy. Most importantly, he suggested a three dimensional indexing system model rather than Jonker's two dimensional model.

### 9.4.2 B. C. Laundry

Approximately ten years have lapsed between the initial work of Jonker and Heilprin and the attempt by Laundry to develop a general theory of indexing and thus provide of firmer foundation for information storage and retrieval systems. Laundry begins the discussion of his model by showing the similarities between the indexing process and the general communication process. He provides a set of fundamental definitions and postulates. His work is predicated upon the existence of three conceptual classes in any and all information storage and retrieval systems : sets of documents, sets of arrtibutes and sets of relationships expressing a connection between documents and attributes. The theory relies heavily on the concept of data elements, and he identifies three important features of this concept : 1. A data element allows for independent manipulation, 2. A data element does not decompose into two or more units, 3. A data element has definite meaning or interpretation.

Although the data element can be any desired entity such as words, strings or words, titles, etc. the indexing system serves to define those data elements that can be recognised and processed by the system.

A document is any well-ordered set of data elements. The documents that are input to the indexing system form a document space. An index is the output from the indexing process. This leads to Laundry's theorem which states :

"An index, by definition, must preserve the well ordering of its parent documents and the ordering of document space. Thus, an index is a well-ordered set of data elements."

### 9.4.3 Gerard Salton

Salton, an information seientist was one of the pre-eminent figures in the field of automation techniques of information storage and retrieval. In 1975 he completed a monograph titled 'A theory of Indexing.' In this monograph he brought together much of the work by himself and others, on evaluating and improving computer-based document storage and retrieval systems, with special emphasis on the need for an effective indexing vocabulary. In the course of this work he defines functions of index terms, or key words in structuring a document collection, proposed procedures for measuring the characteristics of good and bad index form, suggested techniques for modifying the indexing vocabulary so as to improve retrieval effectiveness. His work is both theoretical and experimental. The monograph describes the theory in formal mathematical terms so as to demonstrate its generality and internal consistency, and it reports experimental tests of the theory on three different data bases.

## 9.5 Characteristics of an indexing Vocabulary

The properties of an indexing vocabulary are specificity and exhaustivity. Specificity denotes the level of detail at which concepts are represented in the vectors and exhaustively designates the completeness with which the relevant topics in the document are represented in the indexing vocabulary and in the document-index vector. When retrieving documents from the collection, specific indexing leads to high precision searches in which a reliatively few, highly relevant documents are retrieved. Exhaustive indexing leads to high recall wherein a large percentage of all relevant documents are retrieved along with many non-relevant documents. The indexing vocabulary properties of specificity and exhaustivity account, impart at least, for often mentioned inverse relationship between recall and precision. For example, high precision entails low recall, and vice-versa.

Documents of a store may be divided into relevant and non-relevant groups on the basis of a query. In a search, usually a part of this relevant documents is retrieved Simultaneously, a part of non-relevant documents is also retrieved in spite of indexing filter. This retrieval scene is usually

represented by the following matrix :

| | Retrieved | Not-retrieved | Total |
|---|---|---|---|
| Relevant | a | b | a + b |
| Non-relevant | $\frac{c}{a+c}$ | $n\frac{d}{b+d}$ | $\frac{c+d}{a+b+c+d}$ |

From the above matrix we can observe that out of the total relevant documents a + b only 'a' has been retrieved. This has been used as a measure of retrieval efficiency. Thus $\frac{a}{a+b}$ is considered a measure of recall and $\frac{a}{a+b} \times 100$ is used to denote recall ratio of a search. Precision is also an important aspect of retrieval efficiency.

Mere recall of large number of documents may not serve the purpose of the user unless they are relevant. This aspect of retrieval efficiency is measured by considering number of relevant documents out of the total documents retrieved. The ration $\frac{a}{a+c}$ quantities precision or relevance, and the precision ratio is measured by $\frac{a}{a+c} \times 100$.

It is evident from the above that because of their inverse relationship no indexing system can provide hundred percent recall and precision ratio simultaneously.

## 9.6 Progress and Status

The development of a comprehensive theory of indexing is a complex task that will involve the separate contributions of many information scientists and the integration of these concepts into a unified structure. Theory building will strong then the foundations of information science and will foster understanding and experimentation. A theory of indexing will help explain the nature, the effectiveness, and the quality of indexing. It will provide a basis for automating some or all of the indexing procedures and should expedite efforts to achieve a universally acceptable agreement on specifications and standards of indexes. At the very least, a well formulated theory would make indexing easier to teach and learn.

## 9.7 Definitions of Index

For formal definition of indexing we may turn to either Laundry or Salton, for both their definitions are similar. An index is defined as a well ordered set of data elements. A data element is a word or term such that (a) it has a definite meaning. (b) it cannot be decomposed into two or more meaningful units, and (c) it is capable of being manipulated independently of other data elements.

An accurate index is one that preserves the ordering of the data elements in the present document.

## 9.8 Indexing System : General Features

Indexing systems can be divided into two basic categories : pre-coordinate and post-coordinate, since the days of Cutter the urge for providing 'specific subject heading' has been engaging indexers' minds. But providing specific subject index became a complicated process as subjects developed complexities due to interdisciplinary nature of research. As a result, most of the documents of present days cover composite subjects. This trend has dramatised the values of good indexing. In fact, representation of subject contents of documents is now impossible with a single index term and it calls for a combination or coordination of a number of terms. This coordination can be carried out in two different places— either at the input stage while building up the index file or at the output stage in constructing the required query formulation.

Indexing systems like Chain, POPSI and PRECIS share the mechanism of coordination at input level. Thus these are precoordinated indexing. Pre-coordination does not reuqire sophisticated search logic. These are simply looks under the terms that he expects to find the subject described with, and with a good indes, finds and follows instructions from his first entry point until appropriate document references have been retrieved. The problems of precoordinate indexing, specially the rigidity of citation order, have led to the development of indexing techniques in a new direction, where the problem of taking a decision on the citation order is avoided by isolating the concepts

of a composite subject and keeping them separate for manipulation at search stage. This device has shifted the coordination of index terms from input to output stage making the input word simpler and easier. As the coordination of terms takes place at the output stage, that is, after the indexing operation itself, the system is called post-coordinate indexing or simply coordinate indexing. In general, a post-coordinate index is simpler to produce than a precoordinate index because it shifts the responsibility for coordination of index terms to the searcher.

## 9.9 Specific Indexing Systems

The determination of the subject heading is an important piece of work in cataloguing. But the complexity of composite subject heading baffiled cataloguers and indexers for quite long years. There emerged consistent approaches to subject cataloguing and indexing at different stages of refinement of procedure. This process of refinement has opened up an apparently never-ending line of research for cataloguers and indexers. Here we shall discuss different indexing systems evolved to solve many-fold problems.

### 9.9.1 Chain Indexing

The Chain indexing has been widely used in generating subject indexes to classified catalogues. In principle, it is a manufa means of generating index entries, although some aspects of the system may be seen in computer-generated printed indexes such as Curent Technology Index. It was developed by Dr. S. R. Ramganathan who did most of the early work on analytic-synthetic approaches to classification, its application in the dictionary catalogue was also conceived by its creator and E. J. Coates made its application in a decont way. 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.

A chain is a hierarchy of terms in a classifiaction scheme, each term includes all those which follw it. From this analytical study of classification Ranganathan could distinguish between division of a basic class which are

coordinate and those divisions which are in a hierarchical order and hang like a chain from the basic class. Thus in any basic class we can have as many chains as there are subdivisions. 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. 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. 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 consistent 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.

The steps in the chain indexing are :

1. Assign the appropriate class number from the classification scheme in use to the document being indexed.
2. Examine the hierarchical or other subject structure of the area of the schedules from which a class number has been drawn.
3. Prepare the first index entry by accepting a natural language description of the most specific concept in the 'chain[7], as reflected in step 2 above, and qualify with terms which represent other concepts in the chain. These qualifying terms are to be included in the order that they appear in the chain, moving from specific to more general subjects.
4. Prepare the second entry by choosing a term to represent the second most specific concept in the 'Chain-' and then qualify the term with more general concept terms, in the same way as in the first index entry.
5. Prepare the remainder of the index entries in a manner consistent with the previous index entries.
6. Determine the different kinds of links as noted earlier.
7. Derive a subject heading from each of the sought links in the chain.

The procedure for deriving subject heading is to start from the terms of last sought link and proceed towards the terms of upper link, in a reverse rendering process.

8. Construct the specific heading for specific subject entry or subject reference entry with the minimum number of terms of such upper links as are necessary and sufficient to make the subject heading meaningful and individualised.
9. The specific Subject Entries, Subject Reference Entries, and Entries for Cross References should be merged and arranged in a single alphabetical sequence. The actual operation of chain indexing can be studied with the help of examples. A document titled **German poetry in classical period.**

   The class number according to DDC 19
   —831.6
   The hierarchical chain is as follows :
   8 Literature
   83 German, Literature
   831 Poetry, German, Literature
   831.6 Classical period, Poetry, German, Literature
   Index entries are :
   Classical Period : Poetry : German : Literature 831.6
   Poetry : German : Literature 831
   German : Literature 830
   Literature 800

The document titled **Harvesting of wheat** will generate the following chain :

class number J 382 : 7 (CC6)
J Agriculture (Sought Link)
J3 Food, Agriculture (Unsought Link)
J38 Seed, Food, Agriculture (unsought link)
J382 Wheat; Agriculture (Sought Link)
J3'82 : False Link
J382 : 7 Harvesting, wheat, Agriculture (Sought Link)
Index entries will be as follows :
Wheat, Agriculture J 382

Agriculture J

**Merits :**

Chain indexing is the first systematic procedure laid down for subject indexing. Some of the principal merits are as follows :

It is a systematic, consistent and almost mechanical method of deriving subject entries. It scores heavily over earlier systems on grounds of economy and speed.

With its postulational approach and principle, it is based on a strong theoretical foundation of classification which given it a logical syntax.

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 classification schemes. 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.

**Deficiencies :**

An index linked with a classification scheme has to share the defects and rigidity of the scheme.

Out of the subject headings generated for a document through chain indexing, only the last one is specific and others represent broader subjects. The specific subject heading will be available to only that user who has aparticular search formulation. The problem of disappearing chain lands us in this condition. This problem has been subjected to severe criticism.

This method may generate a number of entries for empty links in the chain in case of a document of a highly specialised field. Here the subject of a document will come at the end of a long chain. It is argued that these empty links and to a little extent all generic entries create noise in the file.

Sometimes a step of division may go unrepresented by a further digit of the class numbers. If the original allocation of subjects was faulty or if a few semicomprehensive subject has developed, a subject may be represented by some of the coordinate classes in an array instead of all of them. The schemes in DDC indicated this by 'Centered Headings'. These should be treated as separate steps in indexing. For example, Crops in Agriculture with class number 633-635. In false links steps of digits in the notation,

donot represent subject, but are meant for structural devices such as facet indicators or synthetic devices. False links are those that connect the facet isolates or phase isolates.

The chain procedures have problems as well as benefits. It operates well backed by a faceted classification, system which carries a regulated and modeulated schedule. With some seeming deficiencies it is a powerful method of subject indexing.

### 9.9.2 Preserved Context Indexing System (PRECIS)

The System intends to allow the user of an alphabetical subject index to enter the index at any signigficant terms which together make up a compound subject statement. The user finds the full context in which his chosen term has been considered by the author. Viewed from another standpoint a full statement similar to the word PRECIS presenting a brief statement of ideas, is offered to the user under every term in the subject which the indexer regards as significant enough to be used at any entry word. In other words index phrases of PRECIS represent brief statements of subjects of documents through the component terms of phrases, designated as strings in the vocabulary of the system. Thus all the important components in a string serve as approach points in turn. Other than these approach points the system also sets up the context in which the terms have been used. The acronym is therefore, significant in both the senses. The system designed and developed by Derek Austin by about 1970, was intended to provide a new system of subject, indexing for the British National Bibliography which was launching the UK/MARC Project.

Some of the features of index entries generated according to PRECIS are :

1. All entries contain all the index terms used in the description of the topic.
2. Access is possible via each and any of the index terms.
3. All terms in the line following the lead term appear in an order that ensures that specific terms are listed first, followed by more general terms.
4. The subject descriptions clearly stated in each entry.
5. Index entries are placed in two lines, the role of the string in establishing

citation order for the components parts of each index entry is the most important feature of PRECIS. PRECIS relies on citation order (with some support from prepositions) to record syntactical relationship, role operators, which are a central feature of the system, are assigned to concepts in such a way that this is achieved.

**Terms, Strigns and Role Operators.**

A concept is usually defined as a single idea or unit of thought. But in PRECIS a concept is a unit of thought, can be logically matched by one of the role operators. A term is a verbal representation of a concept. It may consist of one or more words.

A string is an ordered sequence of terms preceded by role operators. The role operators are code symbols which indicate the function of the indexed term and determine its position in the string of terms representing the subject of the document. The order of terms achieved by the role operators is based on the principle of context dependency. The role operators are meant for the guidance of the indexers only and do not appear in the index entries.

**Formation of Subject Headings**

The labour involved in the formation *of* subject headings is divided between the human indexer and the computer. The indexer undertakes all the intellectual tasks if subject analysis-is to set the indexing terms in a logical sequence according to the syntactical rules. In the second stage, the computer takes over the routine task of producing the desired entries from the inputs provided by the human indexer. The process requires the following steps :

1. Context analysis of documents to determine their specific subject;
2. Selecting appropriate indexing terms;
3. Putting these indexing terms into a string to fall into a sequence;
4. Formation of entries;
5. Alphabetical arrangement of entries.

**Entry Format :**

One of the aspects of PRECIS is context dependency. A string of terms organised into a context-dependent order can be shown in the following manner :

A > B > C > D

The display indicates that the author has considered D in the context of C, C in the context of B and B in the context of A. In another way we can say that each term is context-dependent on the term (A) to its left, and sets the term (A) to its right in their wider context.[1] An example will make the picture clear :

India > Jute industries > Personnel > Selection

Here the first term 'India' sets all the succeeding concepts in their geographical context. The second term 'Jute industries' provides the context in which the third term 'Personnel' occurs and the last term 'Selection' is set in the context by the term 'Personnel', In order to set down the selected indexing terms from a document in the sequence of context dependency and to ensure that different indexers or the same indexer on different occasions consistently, arrives at the same conclusions concerning this input order, a Schema of 'Role Indicators' are used. This Schema determines the order in which terms should be cited. Schema of Operators.

**Primary Operators :**

| | | |
|---|---|---|
| Environment of Core Concepts | 0. | Location |
| | 1. | Key System<br>Thing towards which action is directed e.g., object of transitive action, performer of intransitive action, |
| | 2. | Action : Effect of action. |
| | 3. | Performer of transitive action (agent, instrument). Intake; Factor. Extra-core concepts |
| | 4. | Viewpoint-as-form; Aspect |
| | 5. | Selected instance, e.g. study region, sample population. |
| | 6. | From of document; Target user. |
| Secondary Operators | f | 'Bound' coordinate concept |
| | g | standard coordinate concept |
| Dependent elements | p | part, or property |

| | | |
|---|---|---|
| | q | member of quasi-generic group |
| | r | assembly |
| Special classes of action | s | Role definer, directional property |
| | t | Author-attributed association |
| | u | Two-way interaction |

**Standard Format** : The three parts of a typical PRECIS entry—Lead, Qualifier and Display—are derived in the following way :

**Lead** **Qualifier**

**Display**

The 'Lead' is the term which serves as the user's approach point to the index that is, it is occupied by the filing term. Terms in the 'Qualifier[1] position provide wider context of the approact or 'Lead' term. The 'Dsiplay' position is occupied by those terms which are context dependent on the lead term.

A context-dependent string of terms formulated the following syntax of PRECIS is manipulated by a mechanism known as shunting, to place the component terms in turn in the lead position. Before the shunting is started formulated string is marshalled into the display position leaving the lead and qualifier vacant.

For a document on "Training of personnel in iron industries in India", we can apply relevant role operators to these terms to obtain an input string.

**Concept—Analysis**

| Term | Role | Role operator |
|---|---|---|
| Training | Transitive action | 2 |
| Personnel | Object of action and part of key system | p |
| Iron Industries | Key system | 1 |
| India | Location | 0 |
| Input string | (0) | India |
| | (1) | Iron Industries |
| | (p) | Personnel |
| | (2) | Training |

The standard entries for input string are as follows :

INDIA

Iron Industries. Personnel. Training

IRON INDUSTRIES India
Personnel. Training

PERSONNEL Iron Industries. India
Traning

TRAINING. Personnel. Iron Industries. India.

### 9.9.3 Computer-Aided Subject System (COMPASS)

The ORECIS was developed for the purpose of generating multiple subject strings in a printed index. But its complex coding and system of role operators are unnecessary in an online system. In order to reduce the unit costs of cataloguing the British Library has been using COMPASS since January 1991. COMPASS is a simplified restructuring of PRECIS. The index string is organised by the PRECIS principles of context dependency and role operators.

### 9.9.4 Postulate-Based Permuted Subject Index (POPSI)

POPSI can be applied to micro and macro levels documents available in the form of non-print / non-book forms. It is not based-on any particular system of classification. It is built around a set of fundamental theoretical ideas on classification both in the analysis of subjects as well as in the structuring of the names of subjects. The deep structure of POPSI arises from a Subject Indexing Language (SIL) which should build the basic frame work for any system of subject indexing. POPSI was designed by Dr. Ganesh Bhattacharyya. POPSI is based on certain postulates and principles and is conceptually a generic name and all other systems like chain indexing are different versions of POPSI.

**Principles :**

One of the basic principles of POPSI is the recognition of Elementary Categories. They are : Discipline (D); Entity (E); Action (A); Property (P); Modifier (M). They are explained as follows :

**Discipline :**

An elementary category covers conventional fields of study or any aggregate of such fields. For example, Physies, Chemistry, etc.

**Entity :**

The elementary category E is manifested in perceptual correlates as contrasted with their properties and actions preformed by them or on them. Energy, Light, Place, Time, etc.

**Action :**

An elementary category includes the concept of 'doing.' Thus self-action like function and external action like treatment belong to A.

**Property :**

An elementary category that includes ideas denoting the concept of 'attributes[1]— qualitative or quntitative. For example, Property, Effect, Power, Efficiency. Colour, etc.

**Modifier :**

Other than these elementary categories the concept of modifier for qualifying a manifestation without distrubing the conceptual wholeness of the latter. With the help of a modifier extension of a qualified manifestation is decreased and the intension is increased. Thus 'Infection' in 'Infectious diseases' is a modifier.

**Syntax :**

The Syntax of POPSI is based on Ranganathan's general theory of classification. Precise citation order of the components of subject formulation is achieved with the help of numerical devices as available in the POPSI table. Punctuation marks like comma (,), full stop (.) and hyphen (-) have a distinct role.

**Semantics :**

The system requires a mechanism of vocabulary control characterised by special features. According to the implication of the general theory of the Subject Indexing Language (SIL) the standard vocabulary that controls the use of different concepts, has to be a faceted thesaurus. A hybrid of faceted classification and thesaurus has been suggested for this purpose and christened as **'classaurus'**. Another feature of POPSI semantics is its allowing the use of preposition, **conduction,** participles, etc.. so that the exact meaning is conveyed without any ambiguity.

**Sequencing of Components :**

The POPSI table helps in the sequencing of terms. The components are arranged following decreasing sequence of the ordinal values of the **category** numbers.

**Stages of Operation :**

The index entries according to this system are generated in a systematic manner with the help of the following steps of operation :

1. Analysis of the Subject Indicative Expression
2. Formalisation of the Subject — Proposition
3. Standardization of the Subject — Proposition
4. Modulation of the Subject Proposition
5. Preparation of Entry for Organising Classification (ECO)
6. Decision about Terms of approach (TA)
7. Preparation of Entries for Associative Classification (EAC)
8. Alphabetization.

Let us examine these stages with the help of a sample **title— Chemical treatment of tuberculosis of lungs**

1. Analysis

   D = Medicine (Impticit in the above title)

   E = Lunge (Explicitly stated in the title)

   P of E = Tuberculosis.

   A on P = Chemical treatment (Explicit) = Chemotherapy

2. After analysis the formulisution of the sequence of components is done. The accepted sequence is D. E, A (Modified or unmodified) properly interpolated or extrapolated by P (modified or unmodified). Applying this principle, the components are sequenced in the following manner. Medicine (D), Lungs (E), Tuberculosis (P), Chemical treatment (A).
3. The third step is standardization and is concerned with the semantics, this step helps to decide the standard terms for synonyms and the terms for reference generation. It is done by vocabulary control. The above chain after this step will be Medicine (D), Lungs (E), Tuberculosis (P of E). Chemotherapy (= Chemical treatment).
4. This step augments the standardised subject formulation by interpolating or extrapolating the successive subordinates by using standard terms along with their synonyms, Medicine. Man, Respiratory system > Lungs.

Disease > Tuberculosis > treatment > Chemotherapy (= Chemical treatment).

5. The fifth step is the preparation of the entry for Organising Classification. Numbers of POPSI Table are used to show the categories and the positions of the components in the modulated subject formulation. The above chain will take the following shape :

   Medicine 6 Man, Respiratory system, Lungs 6.2 Disease, Tuberculosis, 6.21 Treatment, Chemotherapy.

6. The next step is to decide about the Terms of Approach. It is done carefully so as to achieve economy. Synonyms are also controlled and references are generated from synonyms to stand rand terms, for example,

   Chemical treatment (Medicine)
   See Chemotherapy

   Use each term other than medicine, and man as be term of approach.

7. This step arranges entries under each approach term. Thus for the present example one of the entries will be Chemotherpy

   Medicine 6 Man, Respiratory System, Lungs 6.2 Disease, Tuberculosis

   6.21, Treatment, Chemotherapy.

   Similarly under each of the other terms.

8. Here all the index entries including references are arranged in a word-by-word sequence ignoring sings and punctuation marks. It is worth to note that in steps 1 to 5 the special interest at each step is to arrive at an organizing classification. This part of the procedure is postulate based. In steps 6 to 8 the special interest is to arrive at various associative classifications. The foundation of these steps rest on the technique of 'Permutation' which refers to 'transformation' or the process of changing the line and order in a definite sequence. Thus the system is rightly known as Postulate-based Permuted Subject Index (POPSI).

POPSI has successfully solved the problem of disappearing chain. It has made the indexing system free from classification scheme because the system is based on the general theory of classification and is not tagged with any classification scheme, POPSI is amenable to computer use.

**POPSI Table**

| | |
|---|---|
| 0 | From modifier |
| 1 | General treatment |
| 2 | Phase relation |
| 2. | General |
| 2.2 | Bias |
| 2.3 | Comparison |
| 2.4 | Similarity |
| 2.5 | Difference |
| 2.6 | Application |
| 2.7 | Influence |
| | Common modifier |
| 3 | Time modifier |
| 4 | Environment modifier |
| 5 | Place modifier |
| 6 | Entity (E) |
| 7 | Discipline (D) |
| | .1 Action (A) |
| | .2 Property (P) |
| | Part |
| | Species/Type |
| | Special modifier |
| | Preceded by the number of the manifestation in relation to which it is action/property |
| 8 | Core |
| 9 | Base |
| | Features analogous to 6 entity/7 Discipline. |

### 9.9.5 Keyword Indexing

Andrea Crestadoro introduced keyword indexing system as early as 1856 under the name 'Keyword in title' for a catalogue in Manchester Pubiic Library. Crestadoro included a 'Concordance of titles' as a quasi-subject

approach with his anther catalogue of 1864. Hans Peter Luhn of IBM revived this systme under the name of Kcyword-in-Context (KWIC) in 1958. In fact, KWIC seems to be the most significant addition to Washington conference in 1958. Hans Peter Luhn of IBM and H. Ohlman of System Development Corporation independently developed and distributed copies of machine produced indexes. Both of them produced permuted title indexes but used different methods. Luhn used IBM computer and named his index KWIC, where as Ohlman used an ordinary punched card machine. But the professional triumph for Luhn was the adoption by the American Chemical Society in 1960 of the KWIC method of indexing and the publication of Chemical Titles by this method.

### 9.95.1 Principle of KWIC

KWIC indexing is based on three principles : (a) titles are generally informative; (b) words extracted from the title can be used effectively to guide the searcher to an article or a paper likely to contain desired information, and (c) althrough the meaning of an individual word viewed in isolation may be ambiguous or too general, the context surrounding the word helps to define and ex-plain its meaning.

A simple KWIC index, is as the name suggests, based on the 'Keywords' that appear in the title of the documents. A title is usually considered to be one line abstract of the document. It is, therefore, expected that words in the title should be significant or keywords should project the subject covered in it. Thus the success or otherwise of the system depends on the efficiency of titles.

### 9.9.5.2 Structure of KWIC

The Keywords-in-Context indexing has three parts : (i) Keywords; (ii) Context: and (iii) Identification Code. The keyword which is displayed at the window provides access point to the search. It should be noted **that** all keywords are displayed in turn at the window and they appear in the index **at** the appropriate place according to alphabetical order. Excluding the keywords, other terms in the index line.express the context in which the

keywords have been used. Identification code at the extreme right indicates the location of the document.

In KWIC indexing, titles are sorted by computer so that each significant word appears in the centre of a column or page in its correct alphabetical sequence surrounded by the words preceding or following it in the title. Computers cannot distinguish the insignificant words unless properly instructed. As a measure of economy and to aviod noise in search, insignificant words should be isolated and steps should be taken so that they do not appear as keywords at access point. This is done by providing the computer with a 'stoplist'. In such a list all initial articles, prepositions and certain insignificant words are excluded.

### 9.9.5.3 Format of KWIC

A sign, either '/' or '=' is used to indicate the **end of the title. The format of KWIC** can be explained with the help of an example. Let us take a document with the identification code 79 **entitled as 'Classification of books in a College Library.' This will have the following entries according to KWIC indexing system.**

| | |
|---|---|
| BOOKS in a college library / classification | 79 |
| **CLASSIFICATION of books in a college library** | 79 |
| **LIBRARY / Classification of books in a college** | 79 |
| **COLLEGE Library / Classification of books in a** | 79 |

**The first word in each entry—Books, Classification, Library, College—is a keyword.**

### 9.9.5.4 Variations of Keyword Index

Some variations in the keyword indexing system have been introduced to overcome its limitations and to improve its working. Important among the variants are :.

(i) Keyword out of context (KWOC); (ii) Keyword Augmented in Context

(KWAC); (iii) Keyword with context (KWWC); and (iv) Key-term Alphabetical (KEYTALPHA).

KWOC (Keyword Out-of Context): Unlike KWIC the keyword at the winddow is followed by the complete title.

| | | |
|---|---|---|
| LIBRARY | classification of books in a library | 45 |
| LIBRARY | Introduction to library cataloguing | 48 |
| LIBRARY | Library user | 49 |

#### 9.9.5.5 Evaluation

KWIC indexes can be produced speedily, no intellectual effect being involved. Title indexes are generally computer produced and thus can have certain advantages. The cumulation of indexes is simple and straight forward. Title indexes can provide a useful current awareness service. As subject retrieval devices they have certain limitations. For exhaustive approach, one has to depend on another index. KWIC indexes are normally produced by a computer and the format is of field length. Owing to this fixed field length, part of the title is truncated when the title is lengthy. As a result, complete context may not be available in all index lines. Some authors relish catchy titles to impart flavour to his style. In such cases, extraction of keywords from titles become difficult. The effectiveness of KWIC depends manly on the titles provided by authors.

### 9.9.6 Post-Coordinate Indexing Systems

Post-Coordinate indexing systems can be grouped into two main categories : **'Term records'** and **'Item** records'. According to the former category a concept is represented as a heading in a term card and accession numbers of all documents carrying the concept are posted on it. Thus we need to make to make as many entries (term cards) for a document as we select terms to describe its contents. In each of these entries accession numbers of the document will be available in appropriate columns. Uniterm indexing of W.E. Batten is an example of term entry system. According to the other approach, only one entry is made for a particular document and all aspects

of the document are coded on the card. The mechanism followed here provides multiple access to the document through the coded concepts. This method is known as item entry system. Zatocoding system of Calvin Mooers comes under this category.

### 9.9.6.1 Uniterm System

The simplest form of term entry system was introduced by Mortimer Taubc in 1953 to organise a collection acquired by the US Armed Services Technical Information Agency. The very name of the index suggests that single term is used as heading in place of composite subject formulation.

The operation of the system involves a number of steps. Initially, document profiles are prepared for each document in the store. Profiles are added as documents are added to the collection. This record is generally maintained in card form. A serial accession number is given to each card, which determines the order of arrangenet of document profile. Other than this accession number the document profiles contain bibliographic description and also some descriptors. Subject content of the document is analyzed to identify the concepts covered. These concepts are then translated into index terms and posted as headings on the term cards. The space at the top of Uniterm card is reserved for this purpose. Rest of the card is divided into tern columns, numbered 0 to 9 as shown in the figure below. For example, if we have document dealing with the compound subject "Library Science Training in West Bengal"; the subject analysis of this document may consists of three simple concepts, namely,

Library science

Training

West Bengal

The document is indexed under each of these simple concepts and sindex entries are prepared. Thus the compound subject of the document receives an index description consisting of three concepts entered individually in the index.

| TRAINING | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 20<br>150 | | 2<br>42 | | 64 | 75 | 106 | | 28<br>88 | 19 |

| WEST BENGAL | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 50<br>70<br>150 | 51 | 22<br>42 | 63 | 64 | 15 | | 117 | 68 | 39 |

| LIBRARY SCIENCE | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | | 42 | 63 | 64 | | 86 | | 68<br>88 | |

| TRAINING | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | | 42 | 63 | 64 | | 86 | | 68<br>88 | |

Search by coordinator of terms

## 9.10 Citation Indexing

The revealed world of seekers with its profundity transcends the bounds of time and space and elevates the posterity. In the sphere of science this elevation is reflected in the sentiment of Newton when he candidly expresses—"If I have seen anything farther that is standing on the shoulders of giants." This indebtednes of researchers, whether in the field of sciences or social sciences, has found expression in citations or bibliographical references accompanying research communications. Ethics of communication claims that previous concepts, methods, apparatus and the like used in the current document, should be duly recorded with the help of citations. Citations are used not only to pay homage to pioneers and give credit for related work but also to criticise, correct and dispute previous contributions.

### 9.10.1 Citation Indexing

A citation index is an ordered list of cited articles each of which is accompanied by a list of citing articles. The cited article is identified as a reference and citing article as a source. The association of ideas existing between the cited and the citing articles is utilised in the preparation of this index. This method has provided a new approach to the problem of file organisation, which overcomes many shortcomings of traditional indexing systems. It is free from the intellectual involvement of indexers and the problems of assigning appropriate subject labels to documents, which bother traditional methods of indexing. It also identifies relationships between documents, which are overlooked in traditional methods. It is amenable to computer manipulation.

Citation Index in general and science citation Index (SCI) in particular owe much *to* Eugene Garfield and his Institute for Scientific Information (ISI) of their popularity and reputation throughout the world. An experimental version of SCI was brought out in 1961. Regular march started 1963 onwards. Frequency of SCI is quarterly. December issue, the last one is cumulated for the whole year. Initially SCI consisted of two parts—Citation Index and Source Index. Later since 1967 a third part, Permuterm Subject Index, was incorporated to provide an additional approach.

### 9.10.2 Citation Index

This part of SCI is arranged alphabetically by cited author. An entry for a cited item carries names and initials of first author (in case of many), year of publication of the cited item and details like volume and page number of the source. The title of the source periodical is represented in abbreviated form. When different papers of a particular author are cited, the papers are arranged in chronological order. When a particular document is cited by different authors in different places, all these citing items are displayed immediately under the cited item. In this source authors are arranged in alphabetical order. Citing item carries information about author and details of citing documents like title, volume, page and year of publication. Coded symbols like A (abstracts), C(Correction), D(Discu'ssion), E(Editorials), I(Tributes, Obituaries), L(Letters), M(Abstfacts from meetings), N(Technical notes) are used to indicate the nature of citing item. As current issues of periodicals are scanned for producting SCI, all the citing items are of the current year.

### 9.10.3 Source Index

The source is an alphabetical list of citing authors accompanied by co-authors and other bibliographical details of the source items like periodical title, volume and issue numbers, page, year of publication and titles of papers. Within the source index a separate section is maintained to meet the corporate approach of users. In this corporate index source items are listed alphabetically by author under the name of the organisation where the work was done. When more than one organisation are involved in a particular project all of them are taken care of making entry under each of them.

### 9.10.4 Permuterm Subject Index

Permuterm is a contraction of the phrase 'Permuted term'. Here the term 'Permuted' is used in its correct mathematical sense. Thus for a title containing significant words, there will be n(n - I) pairs. According to this

arrangement all the significant words appear in turn in primary and co-term positions.

The most important citation indexes are the products of ISI : SCI, SSCI, Arts & Humanities Citation **Index,** together with their machine readable and searchable databases : SCI SEARCH, **SOCIA SCI** SEARCH and ARTS and HUMANITIES SEARCH. **ISI's products are** also available on CD-ROM.

## 9.11 Summary

This unit discusses the importance of indexing in the process of information storage and retrieval. We need a theory of indexing to help us explain, predict and control the dissemination of information. This unit highlights the contributions of four selected authors viz Lawrence Heilprin, B. C. Laundry, Gerard Salton in building the theory of index terminology. The unit discussed many indexing systems viz chain indexing, Preserved Context Indexing System, Computer-Aided Subject System, Postulate-Based Permuted Subject Index, Keyword" Indexing, Post-Coordinate Indexing Systems, Citation Indexing along with their characteristic features.

## 9.12 Exercise

1. Why do you need a theory for indexing? Discuss the contributions of Jonker, Gerard Salton in theory building process.
2. Describe the characterisic features of an indexing vocaboulary.
3. In the context of Chain indexing elucidate the different types of links.
4. What are the merits and deficiencies of chain indexing?
5. How are index entries generated in PRECIS? What is the purpose of role operators in PRECIS?
6. Discuss the elementary categories of POPSI. Enumerate the stages of operation in this system.
7. Elucidate the entries in KWIC indexing system. What are its limitations?
8. What are the main categories of post-coordinate indexing system? Illustrate the sumplest form of term entry system?

9. Write a short note on Eugene Garfield.
10. What are the advantages and limitations of citation index?

## 9.14 References and Further Reading


1. Borko, Harold : Toward a theory of indexing. *Information Processing & Management* 1977, 13, 355-365.
2. Chakraborty, A. R. and Chakrabarti, B : Indexing : principles, processes and products. World Press, 1984.
3. Cleveland, Donald B and Cleveland, Ana D : Introduction to indexing and abstracting 2nd ed, Libraries Unlimited, 1990.
4. Lancaster, F. W. : Indexing in theory and practice. Graduate School of Library and Information Science, 1991.
5. Rowley, Jennifer E : Abstracting and indexing. London Clive Bingley, 1988

# Unit 10 ❐ Indexing Languages : Types and Characteristics

Structure

## 10.0 Objectives

Language in Information Representation and Retrieval Assumes and the form of either natural language or controlled vocabulary. Whenever there is a choice, a question arises as to which type of language one should use for representing and retrieving information. After reading this unit you will be able to :

1. understand the advantage and disadvantage of natural language
2. understand that both are used for specifying subject and organising the searchable file.
3. know the various types of indexing language
4. know the paradigmatic and syntagmatic relations amongst concepts.

## 10.1 Introduction

One of the important fuctions of information processing is to specify the subject of a document available in the store. This information is available in different parts of a document depending upon its nature. Title of a book usually carries this information. For a periodical, title is expected to be one line abstract of the paper. But in many cases titles are illusive. Instead of representing the subject of a document, in many cases, catchy titles represent the literary style and mood of the author. For a periodical article, or a technical report, abstract of the document is another source of information. In case of a book, this information is also available in the preface, introduction, contents page and book jacket. In extreme cases browsing through different sections and chapters may be necessary to ascertain the subject of a document. But whatever may be the source of information, once it is ascertained, this is to be recorded in some languages. This may be natural or artificial.

## 10.2 Natural Language

### 10.2.1 Advantages of Natural Language

Now we are to decide in which language this is to be recorded. So far as the question of specifying the subject is concerned, this can be done in

any language, natural or artifical. Natural language is the main conveyance of communication of ideas, owing to a number of facilities offered by it. With the advancement of time and research news concepts and naturally new terminologies to denote them have been growing continuously in the field of knowledge. Acceptance or inclusion of these new terms in the vocabulary of natural language does not create any problem. But mere acceptance in the vocabulary is not sufficient. A document may deal with a simple or compound subject. For the later, may be specified not by a single term but the association of a number of terms, In such cases the syntax of the natural language for phrase or sentence construction helps us to project the correct meaning. Importance of this aspect will be evident from the following example. Teachers. Students and Assessment are three significant terms in the subject of a document dealing with "Assessment of students by teachers." But these significant terms without prepositions used, project a side-view of the subject. Full view of the subject is obtained by theuse of the preposition 'of and 'by' in proper places. Any wrong associations of terms and the prepositions may convey a different meaning. As in this case, wrong association of the terms and the prepositions will lead us to the following sentence, which is not only different but contrary to meaning—'Assessment of teachers by students.'

### 10.2.2 Disadvantages

The natural language with its flexibility of incorporating new terms and syntax to represent correct association of terms, is competent to specify any subject. But specifying a subject is not enough for organising an index file. This has to be recorded in a system for organising the file so that it can be operated by users with minimum efforts to retrieve information. A concept in natural language may be identified by different terms by different authors. For a natural language, this flexibility is considered its richness. But this richness of natural language is an obstacle for using it in organising index files. A 'mansion' may be expressed by the terms 'Sky-scrapper' a 'High-rise-building' by different authors. Thus in an alphabetical file this concept will be scattered in three different places, if natural language is used in index without any control mechanism and the user will be in a dilemma to

select the entry point or will be compelled to search all possible places. This problem of synonymy in natural language crops up due to a number of reasosn.

(a) The same concept may be expressed by different terms by the specialists and others, for example, Philately and Postal stamp, Ornithology and Bird-watching;

(b) Separate terms are used in popular and descriptive language, for example, Hovercraft and Air-cushing vehicle;

(c) Origin of the term in different places, for example, wolfram and Tungsten.

In organising an index file, one of the many terms for a particular concept is selected for indexing and references are made from the rejected terms guiding the user to the approved term in the file.

Homograph is another problem in natural language making it unsuitable for indexing. The same word is sometimes used in natural language to represent different concepts.

For example, order (command)

order(Sequence)

other (Indent)

The problem may be solved by providing in context in parenthesis.

Would-form is another peculiarity of natural language. Different word-forms may be constructed by combining free morphenes with different suffixes. Thus compination of free morphenes like 'Sail' and 'Heat' with bound auxiliaries like 'ing', 'or' and 'ed' iead us to the following results.

Sail—in—Sailing

Sail—or—Sailor

Heat—ing—Heating

Heat—ed—Heated

A loose relation among the words of a pair may be formed in the idea plane but their spelling difference will scatter them in an alphabetical file.

Abbreviations in the forms of acronym, contraction, initials, and the like used in natural language pose another hurdle in indexing some of them

like Vitamin, Rader—are so widely used that their origin is less known and may not be suitable for entry point.

Again the user is not only interested in a specified topic, he may be interested in broader, related or more specific topic for comprehensive study. Any language used for information processing should have this quality. But this quality of showing different types of relations of concepts like super-ordinate, sub-ordinate, coordinate and collateral, is not inherent in natural language. This unless this relation is shown articially by "See," "See also" and other mechanisms, the purpose of organising the index file will not be achieved.

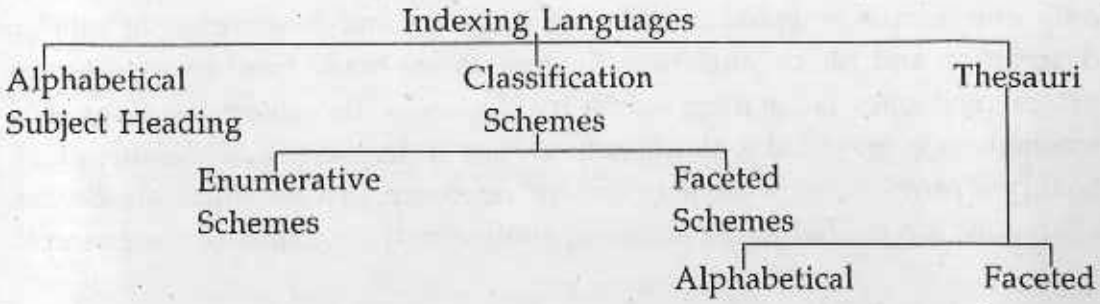
The purpose of indexing is not just to specify the subject-contents of documents it has also got a retrieval function Terms or phrases, describing the subject, must be arranged in a pre-determined order so that the user can find an access point during his search in the file to locate the documents of his interest. Title of the document or some keywords derived from the test may determine the subject of a document. An abstract of the document is a step forward in this direction. But titles or abstracts, however, good efficient they may be in specifying subject, are not helpful in organising a searchable file to locate documents in a store. Thus came the necessity of an artificaial language with its power to control vocabulary, flexibility to show relations of different concepts and facility to build up a searchable file providing access to possible approach points of users. The different procedures adopted in libraries and information centres to perform these functions are now called indexing languages.

## 10.3 Indexing Language

An indexing language has come artificiality in it in the sense that it is different from a natural language. Subject indexing with terms or words and classification with notations, used for subject analysis of documents were considered different types of activities. Only in recent years it has been realised that there is a symbiotic relationship between them. Both are used for the same purpose of specifying subject and organising a searchable file. May be they are performing these functions in different ways. But even in natural languages vocabulary and syntax of one differ from another. In spite

of these differences they all belong to the family of natural language. Besides this realisation of similarity of subject indexing and classification, it was also accepted later that classification is only one form of indexing. All those realisations have come in successive stages in different periods. Other than these two, another device used for document representation and file organisation has come up in the form of thesauri. All these devices combined together are how called indexing language.

### 10.3.1 Types




Once these devices are accepted as indexing languages, naturally some of the linguistic concepts have been applied in their study. Thus in the assessment, comparison and refinement of indexing languages, different linguistic concepts like semantic, syntax, size of the vocabulary, vocabulary control, facilities for representing paradigmatic and syntagmatic relations, etc. of indexing languages are taken into consideration. These are the characteristics where an indexing language differs from a natural language.

### 10.3.2 Characteristics of Indexing Language

Indexing language is designed for a special purpose, Hence, apart from the common objective of being a vehicle of communication of ideas it has to perform some special functions. Other than subject description indexing language is also used in organising a searchable file to be used by the information seekers. A match between the description of document done by the indexer and description of 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. In successful organisation of this file and

subsequent matching of doculent and request, semantic and syntax of indexing language play very important roles.

### 10.3.2.1 Vocabulary Control

In respect of semantic aspect of indexing we need to consider the size of the vocabulary and devices for vocabulary control. Vocabulary of indexing should be precise and exact. A complete one to one relationship among 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 file organisation. References are made from rejected terms to accepted ones facilitating search by the users. In subject headings 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.

### 10.3.2.2 Coordination of concepts

Specifying a concept by a particular term, out of available synonyms, dose not solve all the problems in indexing. Most of the documents of present days cover compound subjects comprising a number of concepts. Naturally, in these cases the complete subject of the document can be represented not by a single term but by a coordination or combination of a number of terms. A subject class generated by coordination of two or more terms, representing different concepts, will differ from the classes represented by the individual terms or by the terms in some other combination. For example, the terms Library, College and Building may be coordinated in different ways like **library college building** and **College library building.** The meaning of one coordinated from differs from differs from the other. In indexing language this coordination of constituent terms frequently made to generate index phrase at input stage. But there are exceptions also. This coordination at input stage is carried out only for precoordinate indexing. But is post-coordinate indexing the complete subject is analyzed into terms, representing simple concepts and coordination of these constituent terms are carried out

only during search or output stage. Thus the two systems differ in coordination during input stage but for both of them coordination of terms is necessary during search stage.

### 10.3.2.3 Sequencing of Terms

The coordination or combination of terms gives rise to the problem of sequencing of terms. This sequencing or ordering of terms is important in indexing. The four terms a, b, c, d may be arranged $\underline{|4}$ or $4 \times 3 \times 2 \times 1$ or 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 should also be helpful in search.

In a natural language other than the basic terms preposition, conjunction, etc. are also used to convey the correct relationship of the terms. Absence of this facility in indexing language is compensated by the correct sequencing of term. A simple example may be cited to emphasise the importance of this aspect. A sentence in natural language—'Assessment of students by teachers' conveys certain meaning. Correct relationship of the substantive terms, 'Assessment,' Teachers' and 'Students[1] is indicated by the prepositions 'of and 'by' in this example. Any change of place of the terms and prepositions used in this example may convey a different meaning. Thus sequencing of terms in approapriate manner is important in indexing language.

In spite of the rules of sequencing of terms, prescribed in the syntax of indexing language, it becomes difficult at times to show the correct relationships of terms. To solve this problem indexing language has made the provision of relational symbols known as role operators, categories, etc. in different systems. Ranganathan introduced different punctuation marks among the Fundamental Categores PMEST.

### 10.3.2.4 Syntax of Indexing Language

Importance of sequencing of terms is accepted in all indexing languages. But the rules prescribed by them for ordering of terms differ. J. Kaiser in his Systematic Indexing suggested that compound subjects may be analyzed into a combination of a 'concrete' and a 'process.' Again within this

combination 'concrete' is more important and should be placed in the first position. Thus subject formulation for 'Extraction of iron ore' will be entered as :

Iron ore—Extraction.

Kaiser also realised that 'concrete' and 'process' are not sufficient enough to represent all the concepts in a composite subject. So in his analysis he also made provision of 'place' is involved in a composite subject, he prescribed double-entry. Once under 'concrete' and again under 'place.' A specific locality is represented first by the country followed by the place. Thus subject formulation for 'Extraction of iron ore from Bihar,' will be rendered as :

Iron ore—India, Bihar—Extraction
India, Bihar—Iron Ore—Extraction

The three categories concrete, process and place, as suggested by Kaiser are not sufficient enough to represent the composite subjects of modern times. Ranganathan went deep into the problem and came out with his idea of Five Fundamental Categories—Personality, Matter, Energy, Spece and Time or in short PMEST. The citation order of PMEST represents the order of decreasing concreteness of the concepts in a document. He introduced these categories for subject analysis of documents in his Colon Classification, which according to modem idea is nothing but an idexing language.

A new approach to subject formulation of a document has been provided by J.E.L. Farradane. This approach created a new type of syntax in indexing language. Instead of considering attributes of component terms, in isolation or within a class, Farradane highlighted the relationships that existed between each pair of terms. Nine types of relationships like, association, comparison, concurrence, dimensional, distinctness, equivalence, Appurtenance, Reaction, causation, have been recognised in this device. The relationships are indicated by different symbols or operators by placing them in between the pair of terms.

| Concurrent | Concurrence | Comparison and Self activity | Association |
|---|---|---|---|
| | /q | /* | /; |
| Non-Distinct | Equivalence | Dimensional | Appurtenance |
| | /= | /+ | /( |
| Distinct | Distinctness | Reaction | Causation |
| | /) | /– | /: |

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. All these operators stand for specific types of concepts. Subject of a document is formulated initially in the form of a tile like phrase. Syntactical role of different components of this is ascertained, and the operators, expressing these roles, are assigned to the components to form the subject formulation, which is a string according to the vocabulary of PRECIS. An example will be helpful in understanding the method of string formulation.

Assessment of students of medicalcol leges in West Bengal

The following component concepts can be identified in the phrase :

Assessment
Medical colleges
Students
West Bengal

Out of these concepts, 'Assessment' represents an action and thus the role operator (2), representing action, is perfuxed to the term. 'Medical college' is the key system and "students' are part of it. These are presented by prefixing (1) to the former and (p) to the letter. The whole picture is displayed within the environment 'West Bengal'. Thus location operator (o) is prefixed to it. The total analysis presents a string in.the following form :

West Bengal Medical College Students. Assessment.

Sequencing of terms is based on the ordinal value of the operators. A non-numerical operator is attached to the concept with which it is related.

### 10.3.2.5 Rotation of Component Terms

Rules of syntax of indexing language help us to formulate subject representation of a document. But in a linear representation like this, 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 out 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 used not 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.

It may be noted that this idea of rotation of component terms was in rudimentary form in chain procedure of Ranganathan. The chain procedure lacked the facility of projecting context as the terms were deleted in successive stages. But the idea of different lead terms was very much present. Application of computer has accelerated this rotation of terms in indexing language as this mechanical part of indexing work can be done by a computer efficiently and quickly. Other than KWIC (Key-Word-in-Context) index and thie varieties some indexing languages like PRECIS and POPSI have also maintained this trend.

#### 10.3.2.6 Syndetic 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 syndetic devices. These include guides, cross-references, glosses, inversion of headings, introduction to indexes, etc.

Guides and introduction to specific indexes provide useful information regarding scope and stricture of the index, terminology and symbols used, rendering of subject headings, depth of indexing, format and typography, etc. While in some cases these guides are brought out as separates but for others this is published as an integral part of secondary periodicals, either at the beginning or at the end of it.

The cross references of different types are the most important of all the syndetic devices. They correlate similar concepts scattered throughout the index due to its alphabetical overtone. They also guide the user from his present position in the index to where he should locate his desired information. These cross-references are mainly of two types—'see' and 'see also.'

These references guide the users from synonyms to the preferred terms of the index. In thesaurus, a type of indexing language 'see' reference is performed by 'use' reference. For example,

Index Terms **See** Headings.

Brinjal **See** Eggplant.

In a thesaurus this guidance is normally provided with the help of UF

(Used For) Symbol. For example,

Grain Crops

UF Barley.

Inverted headings are also used as a syndetic entries together. For this purpose words, terms and pharases are placed in an inverted order contrary to the normal order of construction. This inversion is carred out usually with help of a comma (,). For example,

Current Awareness Service, definition of

Zinc, Metallurgy of

### 10.3.2.7 Relational Symbols

In a natural language correct relationship between two or more terms can easily be established with the help of preposition, conjunction, etc. Thus 'photographs of albums' and 'albums of photographs' convey different meanings. This differentiation is possible with the help of preposition, conjunction, etc. This facility is absent in an indexing language, which is an artificial one. To overcome this difficulty some indexing languages use relational symbols or indictor digits to bring out correct relation between the words. For example,

| Indexing language | Role operators |
|---|---|
| Colon classification | , ; : . ' |
| Universal Decimal classification | + / : = ( ) (= ...) " ... " |

### 10.3.2.8 Paradigmatic and Syntagmatic relations

Linguistic concepts like semantic and syntax have been introduced in the study of indexing languale. Other linguistic concepts introduced in the study of indexing language are paradigmatic and syntagmatic relations. Paradigmatic relations are those which are known in advance before scanning any particular document, while syntagmatic relations are inderstood only after scanning a particular document. Paradigm means a set of words having the same stem. In the context of indexing languale this term has been used in a slightly different manner to show different types of genus—species relations that exist among different concepts. The relations include broader-narrower and associative relations, and are displayed in a classification

schedule. For example, magnetism include eletromagnetism but itself a part of physics. Magnetism and electricuty are of equal rank as both of them are part of physics. This relation is known to an user without reference to any particular document and is displayed in a classification schedule. This is paradigmatic relation.

A syntagmatic relation is restricted to a particular document. On analysing the subject contents o a document we may establish the syntagmatic relation of different concepts, covered in the document. But it is difficult to tell in advance the different syntagmatic relations that will occur in collections to be procured in future. In a feceted classification scheme isolates under each facet may be combined appropriately to represent different subjects. These combinations are not permanent relations and thus are not displayed in the schedule. These combinations are made for specific purpose of representing the subject of a document. Thus syntagmatic relation is restricted to a particular document. Let us take an example to clarify, the point. A document on 'Impact of artificial light on food habits of noctural birds,' taken from broad subjects Physics and Zoology. This particular combination of concepts is a speciality of this document which is not normally necessary and naturally is not shown in the notations of a classification scheme. In a faceted scheme this mechanism is provided to show these relations of concepts covered in specific publications. Indexing language should be able to denote the subjects and show their relations and directions. While paradigm takes care of showing general relations, specific relations and direction of subjects are shown by syntagma. Thus an effective indexing language, used to represent the subject of document should have the mechanisms to show both the relations.

### 10.3.2.9 Structuring of indexing language

Indexing languages should be designed with the ultimate objective of meeting the subject approach of users. To meet this objective indexing languages have to take into consideration different characteristics of users' approach. Normally the user may be interested on a specific topic but during search he may also look items dealing with narrower aspect of the topic. He may also broaden his search and look for a document on borader subject

covering his topic of interest. Users may also be interested in collateral subjects during their search for information. Indexing languages have to take care of all these characteristics by showing all these relationships of concepts in different fashions. Thus indexing languages are structured. Different indexing languages viz., classification schemes, subject heading listsa and thesauri follow different methods and thus differ from one another.

In an enmmerative classification scheme this relationship is shown with the help of notations and their display. In a list of subject headings this structure is shown with the help of 'see' and 'see also' references at approapriate places. A thesaurus takes the help of different symbols like BT, NT and RT for the purpose of showing this structure. This structuring helps in describing subject contents of documents at input stage and also during search stage by broadening or narrowing range of search. This peculiarity of indexing language is a hurdle for introduction of new terms in the vocabulary.

## 10.4 Summary

This unit on 'Indexing languages : types and characteristics' presents a vivid picture of natural language and its advantages and disadvantages as well as indexing language, its characteristic and types. It describes how vocabulary is controlled. It discusses coordination of concepts, sequencing of terms, syntax of index language, rotation of component terms, syndetic devices used, relational symbols, paradigmatic and syntagmatics relations and finally structuring of indexing language. It explains the peculiarity of indexing language.

## 10.5 Exercise

1. Discuss the advantages and disadvantages of natural language.
2. What is indexing language? Describe the various types of indexing language.
3. Discuss the characteristics of indexing language.
4. Discuss why is sequencing of terms important in inde[illegible]guage? Illustrate.

5. How does Ranganathan solve the problem of syntax in indexing language?
6. What do you understand by 'syntagmatic relations'?
7. How is indexing language structured?

## 10.6 Keywords

1. Homograph : A word of the same spelling and pronunciation as another, but of different meaning and origin.
2. Morphism : Noun Combining form.
3. Natural language : It is the language people speak and write. What people use for representing information or forming a query without consulting a controlled vocabulary is called natural language.
4. Paradigmatic relations : To exhibit side by side. Paradigmatically related terms are normally to be found within a facet.
5. Role operators : In PRECIS some of the aspects are like pervious categories and other are of different nature. They have been named as Role operators.
6. Semantic : 'relating to meaning.'
7. Syndetic device : Connective devices.
8. Syntagmatic relations : Syntagmatically related terms are to be found in the facet analysis of a particular document.
9. Syntax : Grammatical structure in sentences.

## 10.7 References and Further Reading


1. Bernier, C. L. : Syndetic Systems. In Kent, A & ORs ed. Encyclopedia of Library and Indoemation Science. 1980, Vol. 29, 334-356.
2. Chu. Hating : Information representation and retrieval in the digital age. Thomas H. Hogan. Sr, 2003.
3. Chakraborty, A. R. and Chakrabarti, Bhubaneswar : Indexing : principles, processes and peoducts. The World Press, 1984.
4. Guha, B. : Documentation and information : Services, techniques & systems. 2nd rev. ed. the World Press, 1983.

# Unit 11 ❐ Controlled Vocabulary

Structure



## 11.0 Objectives

What people use for representing information or forming a query without consulting a controlled vocabulary is called natural language. After reading this unit you will have :

1. an idea about controlled vocabulary
2. an idea about major types of controlled vocabulary
3. an idea of comparison of three types of controlled vocabulary
4. an idea about the coexistence of natural language and controlled vocabulary in IRR (Information Representation and Retrieval).

## 11.1 Introduction

An index language is an artifical language which is adopted for the purpose of indexing. Such a language will not only reflect a controlled vocabulary but also seek to indicate the relationships between the terms in its vocabulary. In other words, an index language which has a controlled vocabulary and which seeks to show relationships between terms in the index vocabulary, is said to be structured. Controlled vocabulary must be constructed and maintained with specific subject area (s) in mind. Terms to be included in controlled vocabulary are selected by following the principle of either literary warrant or user warrnat. Literary warrant means that terms to be included in the controlled vocabulary must be chosen from existing literature. Similarly, user warrant implies that terms to be selected for inclusion in a controlled vocabulary must have been used in the past. There are various types structured index language, classification being the most familiar example. List of Subject headings is an example of second type and the thesaurus an example of third type.

## 11.2 Classification Schemes

The history of classification would almost necessarily be a history of all attempts to organize human thoughts. It was attempted in many ways, shapes and forms as knowledge advanced, a classification scheme is a controlled vocabulary of terms in alpha-numerical that are designed for pre-coordination. Units in classification schemes are called classes, which are further labelled numerically or alphabetically. Classification schemes have witnessed and weathered the changes over the 20th century. On the other hand, they have been updated and revised again and again as time passes.

Unlike the other two types, classification schemes are built on an artificial framework of knowledge. For example, Dewey Decimal Classification chooese to present the world of knowledge on the basis of ten classes. Exactly 10 classes are formed at each level within the scheme, Which indicates its rigidity. The hierarchical relationship among classes is displayed using different lengths of classification notations. That is the longer a classification notation, the deeper the class locates in the hierarchical structure. Associative relationships among the terms are displayed via the "See" and "See also" mechanism which is only occasionally used.

Traditionally, classification schemes are most often selected for representing and retrieving monograph information.

## 11.3 Subject Heading Lists

Subject heading has been defined as "a word or a group of words indicating a subject under which all material dealing with the same theme is entered in a catalogue or a bibliography, or is arranged in a file " Credit should go to Crestadoro who for, the first time in his book 'the Art of Making Catalogues' published in 1856 could realise that the cataloguer should provide a standardized guide to subject content of a book by giving it a heading.

A subject heading list is a controlled vocabulary of terms in natural language that are designed both for precoordination and post-coordination. Pre-coordination, predominant before the 1940s, combines terms before they are used for representation and retrieval. If one intends to, for examople, represent and retrieve documents about **Internet retrieval** systems, the three-word phrase would appear exactly in a pre-coordinated language. Procoordinated controlled vocabularies are thus nonmanipulative.

Since subject heading lists allow both pre-and coordination, their flexibility seems better than classification schemes.

Terms in subject heading lists are called subject headings and are arranged alphabetically. Notations for subject headings include "See" for pointing to preferred terms from nonpreferred ones while "X" meaning "See from", presents the reciprocal expression of the "See" statement. For instance, **handicapped** See **physically Challenged** and **physically challenged**

**X handicapped.** Another set of notations for subject headings lists is "See also" and "XX", which stands for "See also from". "See also" indicates both the hierarchical and associative relationships among the choosen subject headings. That practice, however, undermines the specificity of subject heading lists. Similar to X notation, XX shows the reciprocal expression of the "See also" statement.

Generally, subject heading lists are used for both information representation and retrieval although they appear to be applied less frequently as the thesauri for representing and retrieving nonmonograph information.Sears' List of Subject Headings is one example of subject heading lists. Another well-known example is the Library of Congress Subject Headings, whose notations from 11th edition, however look more like that for thesauri.

The National Library of Medicine calls MeSH. a thesaurus and in the sense that it provides a strict hierarchical structure and it is aubject-oriented, it is a thesaurus. But in the sense that it precoordinated phrases (for example, "Life change Events") it is a subject heading list. It also has a subdivision list from which terms are to be taken to subdivide the terms and phrases found in the list proper. In print from MeSH has three volumes : a hierachical listing; an alphabetical arrangement that includes scope notes; and a permuted alphabetical listing, in which every word of a phrase heading is brought into the lead position and arranged alphabetically.

## 11.4 Thesauri

A thesaurus is a controlled vocabulary of terms in natural languages that are designed for post-coordination. Post-coordination is manipulative. Post-coordinated languages allow users to coordinate terms at the time of retrieval. The main drawback of post-coordination is false coordination. The absence of recognised relationships could result in ambiguity, for example, a search for fertilizers for sugar beet by the simple coordination of Sugar and Beet and Fertilizers would also produce documents on the use of sugar-beet tops as fertilizers.

Standard notaions are used in thesauri to specify the. hierarchical, associative, homographic, and other relationships among the terms included.

USE and UF (Used for) identify which terms in the thesaurus are preferred descriptors while SN (Scope Note) defines the meaning of a descriptor. NT(Narrower Term) and BT(Broader Term) display the hierarchical relationships among descriptors. The associative relationships among descriptors are expressed by the notation RT (Related Terms).

Thesauri are the most commonly used controlled vocabulary in nonmonographic information retrieval for their specificity, flexibility and ability to tackle complex concepts. There are many thesauri for different subject areas.

## 11.5 A Comparison of Classification Schemes, Subject Heading Lists and Thesauri

The three types of controlled vocabularies are designed to perform some common functions and as such they have much in common. Basically all of them are lists of subjects, mostly composite ones, represented either by words or by notations. These are the mechanism by which vocabulary is controlled in indexing and thus anarchy is avoided. Other than the projection of preferred term and notation they also summarises relationship between terms in an indexing language.

In spite of these similarities the three processes differ in respect of their structure and use. Considering the arrangement of entries we may bring subject heading lists and thesauri in one group and classification schemes in another. The mode of arrangement in the first group is alphabetical. Both the preferred and rejected terms in this group are listed in the same sequence of alphabetical block, within which relationships between terms are displayed. A thesaurus may often boast an additional explicit statement of the lists or graphic displays. In a classification scheme notations, representing concepts or isolate ideas, are arranged in a linear order. The classification scheme, whether enumerative or faceted, is structured and supported by an index, which guides the user from term to the notation used.

Rules for using these tools also differ. Specific instructions are provided in case of subject heading list to select the appropriate index term. The user is guided to move from his term to the accepted one by the help of "see" reference. In a theasaurus the mechanism is built within the system. Two

types of symbols, 'USE' and 'UF' are used to identify the preferred term of the vocabulary. In a classification scheme, whether enumerated or faceted one, a built-in mechanism helps the use of the system. Further index serves as additional approach point.

The precoordinate vocabulary is enumerative whereas the postcoordinate language is synthetic. Analysis and coordination methods also determine the specificity and flexibetity of controlled vocabulary. A synthetic and postcoordinate vocabulary appears more specific and accommodating than an enumerative and precoordinate language.

Thesauri seem to have more specificity and flexibility than classification schemes while subject heading lists fall in between on this spectrum of measurements which explains why thesauri have been the most used controlled vocabulary among the trio in information retrieval.

## 11.6 Natural Language Vs Controlled Vocabulary

We have discussed the features and characteristics of both natural language and controlled vocabulary. Now it is time to contrast how they differ from each other and what they have common as **IR** languages.

### 11.6.1 Different Eras of IR Languages

Controller vocabulary came later than the natural language in the history of IR. Four different eras can be identified in this context.

The first era refers to the time before any controlled vocabulary was introduced. Natural language was the only language applied in information representation and retrieval. Soon people began to realise the limitations (e.g. homographs and synonyms) of natural language when it was applied for representing and retrieving information.

The second era is signified by the introduction of controlled vocabulary to information representation and retrieval. In the beginning, precoordination systems such as classification schemes were heavily used. Other types of controlled vocabularies viz. subject heading lists and thesauri, were implemented later in the field. But the debate on natural language vs controlled vocabulary started and has been continuing.

The third era is symbolized by the resurgence of natural language due to keyword retrieval technique and the development of full-text systems. Controlled vocabulary is still applied **for** IR but most often to bibliographic IRR systems. Will controlled vocabulary remain a viable IR language? More debates have been provoked in this era the issue of natural language vs controlled vocabulary. The fourth era started when natural language interface was implemented in information representation and retrieval while vocabulary control mechanism was employed behind the scenes. Since the vocabulary control mechanism is invisible to the end-user, Milstead, Jessica L called it invisible controlled vocabulary in a natural language retrieval environment.

### 11.6.2 Co-existence of natural language and controlled vocabulary

The reasons behind the coexistence are the complementary present cons of the two kinds of Information Retrieval languages.

#### 11.6.2.1 The Synonym Issue

The issue of synonyms is one of the most cited reasons against using natural language in IR. In natural language, there are synonyms that use different terms referring to the same entity. For example, microcomputers, personal computers, PCs, desktops, Pentiums and the like are all synonyms. But which term should be chosen and used for representation and retrieval purposes? This is an issue in the debate about **IRR** languages. When controlled vocabulary is the selected **IRR** language, synonyms are not a source of concern because one and only one term is chosen from the controlled vocabulary as the perferred term. The rest of the synonyms are treated as non-preferred terms and cross-referenced to the perferred term.

#### 11.6.2.2 The Homograph Issue

Homographs are terms spelled the same but carrying different meanings in different contexts. Terms such as order, record, drug, bank and duty are a few examples of homographs. In natural language representation and retrieval, homographs can produce problems such as ambiguity, due to lack

of context. One common method is to use parentheses to specify context for homographs, for example, Order (classification) order (Indent), duty (responsibility) and duty (tax).

#### 11.6.2.3 The Syntax Issue

Language has syntax. But how can syntax be expressed when natural language is used for representing and retrieving information? Suppose is represented by three terms Idea, automobiles, and china in natural language. This document can be about, for example, China exporting automobiles to India, or the opposite, India exporting automobiles to China. It is not clear which country is the exporter when given only these three terms without any syntax information. This problem can be solved easily in controlled vocabulary with devices such as roles, which are symbols or numbers that indicate the syntax relationship between or among terms. Here, we may define the number 1 as the role of exporter and put it next to the term China i.e. China (1). Similarly, we can designate the number 2 as the role of importer and then assign it to the term India (2). The mechanism provided in controlled vocabulary allows people to tackle the syntax issue while natural language cannot.

#### 11.6.2.4 The Accuracy Issue

It is always preferable to have an **IRR** language that could accurately represent information and search queries. That objective can be attained if natural language is chosen as the IRR linguage. The rationale for the statement seems two fold. First, no additional manipulation (e.g. explanation) is required in using natural language for IRR. Second, interpretation is unnecessary in natural language interpretation and retrieval what the author / user uses will be the language for IRR. In comparison, controlled vocabulary is artificial, not reflecting the richness of natural language used for composing documents and expressing search queries. Controlled vocabulary also lacks specificity as a result of the language manipulation process. **Interpretation of controlled vocabulary** seems indispensable as the connotation and denotation **of each term are defined with particular target audience in mind. Such interpretation could** invariably introduce inaccuracy in **IRR** using controlled vocabulary.

#### 11.6.2.5 The Updating Issue

Natural language needs no updating while controlled vocabulary does. New terms can be used for IRR purposes as soon as they occur in natural language while any change in controlled vocabulled vocabulary cannot be implemented until it goes through the rigid updating process. Consequently, terms in controlled vocabulary are not always upto date.

#### 11.6.2.6 The Cost Issue

It takes time to create, maintain and learn how to use a controlled vocabulary in information representation and retrieval. The time needed for these activities will eventually get translated as cost for information representation and retrieval. In comparison, natural language is just the language people commonly use. Neither training nor maintenance is reuqired for doing IRR in natural language.

#### 11.6.2.7 The Compatibility Issue

There are times when a IRR system needs to change its language during the course of its development or when a user would prefer to perform multiple database searching. The compatibnility issue thus emerges for controlled vocabulary IRR as each controlled vocabulary has its own unique features and characteristics. It seems impossible, for instance, to use Universal Decimal Classification to perform IRR tasks in an OPAC that actually uses Library of Congress Classification as its controlled vocabulary. But natural language should always be compatible with itself. So if the systems are built on natural language, the switching or migration from one system to another does not arise.

In summary, the strengths for controlled vocabulary in synonyms, homographs and syntax are weak points for natural language. Likewise, the weaknesses for controlled vocabulary in accuracy, updating, cost and compatibility are strong points for natural language. Both have found their own places in information representation and retrieval.

## 11.7 Summary

This unit introduces three types of controlled vocabulary. Each type has its own characteristic features. A comparative study has been attempted. Natural language and controlled vocabulary also with their strong and weak points exist. There is some agreement as to the relative merits of each of these systems.

## 11.8 Exercise

1. Identify different eras of IRR languages.
2. Discuss the complementary pross and cons of natural language and controlled vocabulary.
3. Why have thesauri been the most widely used controlled vocabulary in information representation and retrieval?
4. Which type of controlled vocabulary would you use for retrieving monograph information?

## 11.9 References and Further Reading


1. Chakraborty, A. R. and Chakrabarti, B. : Indexing : principles, processes and products, World Press, 1984.
2. Chu, Heting : Information representation and retrieval in the digital age. Thomas H. Hogan, Sr. 2003.
3. Lancaster, F. W. : Vocabulary control for information retrieval. Information Resources Press, 1986.
4. Milstead, Jessica L : Invisible thesauri : The year 2000. **Online & CR-ROM Review,** 1995, 19(2), 93—94.
5. Rowley, Jennifer E : Organizing knowledge : An introduction to information retrieval. Gower, 1992.

# Unit 12 ❐ Construction of IR Thesaurus

Structure

## 12.0 Objectives

Thesauri that are components of reference retrieval systems must conform to the goals of the systems, must be related to the system content, and must be designed to serve as a retrieval device for a designated category of users. This unit will make you aware of purposes and functions of thesaurus; of the different approaches for the construction; of controlling vocabulary; and finally, of displaying the thesaurus.

## 12.1 Introduction

Thesaurus design is complex and may be approached from serveral directions. Harold Wooster describes two approaches and in his whimsical style, characterises these as the stalagmitic and stalactitic. More precisely, these two approaches for the selection of terms in a thesaures are : (a) terms derived from a relevant set of documents, and (b) terms derived from a consolidation of existing thesauri. A third, closely related approach is through the development of micro thesauri and a forth approach is a compromise,

which uses indexers to select terms and lexicographers to organize them.

The stalagmitic approach is one that begins "down on the floor of the cave among the documents, slowly building toward the ceiling." Less whimsically, this can be described as the empirical or original—indexing approch. A representative sample of documents is collected and indexed, by means of a free vocabulary consisting of the most approapriate terms with no controls or restrictions. After completion of this task, the designers of the thesaurus review, organize and structure of the proposed terms. If the documents have been truly representative and the indexing satisfactory, then all of the index terms of concern to the information system would have been included. This is the idea and ideal of the stalagmitic approach to thesaurus. In practice, the ideal is never fully achieved.

The stalactitic approach is what Wooster describes as being "much more fun— one convenes groups of experts who hang up on the roof of the cave, twittering and chirping among thesmesleves but as far away from the actual documents as they can get." Let us call this consolidation approach. In this approach, committees of experts are convened and are assigned reponsibility for preparing a consolidated list of terms that can be used for indexing. The committee members donot work from documents; instead, they use previously compiled lists of terms. The committee strives for completeness. It may be recalled that the original-indexing approach often leads to incomplete list of terms. In the consolidation approach, terms can be added even **though** no document has, as yet, been indexed by that **term.**

## 12.2 Purposes of thesaurus

The purpose of thesaurus is to store reference information data in the most logical form of organisation. The family library can hold hundreds of books, photo albums, and scrap books, all of which can be located in the dark, by a blind folded family members, from age 5 to 95, but businesses cannot work that way. They need systems and planned space for their records and files. Secretaries come and go, file clerks can forget things, and vice Presidents of research are not infallible. 1Information science experts cannot bring peace to the world or change the rate of inflation but they can organize

what had been overflowing from file cabinets, They can keep the most accurate and current information available, something that is almost as hard to do as predicting the weather.

It you want a little more order in your business world, add a thesaurus to straighten things out. It will remove some of the chaos. It will remove some of the chaos. It will determine the exact term or concept from among other similar, ambiguous or overlapping terms. It will save time, money and labour wherever possible.

A fascinating aspect of thesauri is the preceise and narrow definitions of terms found in the best examples. The systems of coding and organisation can be so refined that cataloguers and indexers devote lifetime to their complexities,

The best thesauri control the arrangement of material so you will not obtain too much, not enough, too simple, or too' difficult material. It you want information on plastics, for example, you must be specific as possible or you receive much inaccurate information. Precise labeling and retrieval of information is goal of thesaurus creation. When we say 'Plastics', we could mean malar, polystyrene, polypropylene or superficial and shallow people.

## 12.3 Functions of thesaurus

A thesaurus seeks to make the meaning of wors clear to everyone. An accurate, precise, and thorough definition of an idea is vital to good communication. It is the foundation for accurate information searching. Accurate definitions eliminater ambiguity from the core of a concept and delineate the parameters of its connotation. A thesaurus helps to reduce contradictions by clarifying variations in preference, function, and intent for the specific needs of a particular user question.

As quantities of research and technological developments progress with overwhelming speed and volume expert organisations and analysis are required to obtain the precise information wanted in the shortest possible time. A thesaurus assists in this goal by providing orderly, logical arrangements of similar terms and concepts. A thesaurus allows a near infinite

growth and addition of need terms and concepts to proceed in a continuously logical manner. One of the more important functions of a thesaurus is to display relations among terms of the vocabulary and thus help the user to select the most appropriate terms when idcxing a document or formulating a search request.

In terms of function "A thesaurus is a terminological control device used in translating from the natural language of documents, indexers or users into a more constraied 'system language'." (Documentation language, Information Language). UNISIST (World Science Information System of UNESCO).

## 12.4 What is in a thesaurus?

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. A thesaurus gives several types of information to indexers and searchers.

### 12.4.1 Preferred Terms

The thesaurus has to indicate which terms indexers and searchers are allowed to use. These terms are called preferred terms. This is a major part of vocabulary control—restricting the vocabualary so that it is easier to predict what words might have been used to index a concept.

### 12.4.2 Non-preferred Terms

In addition to preferred terms, a thesaurus also needs to indicate some terms that indexers and searchers are not to use. These terms are called non-preferred terms. It should be possible to look up a non-preferred term and see what preferred term should be used unstead. This will save time and make it less likely that the best preferred term will be missed. A thesaurus also usually allows you to look up a what the term is supposed to mean.

### 12.4.3 Semantic Relations

As well as linking preferred terms with non-preferred terms, a thesaurus also shows links between different preferred terms. These links are usually semantic relations. Like a link between a preferred term and a non-preferred term, one of these semantic links can help to direct you to the right term and make the meaning of a term clearer.

### 12.4.4 Guides to Application

A good thesaurus should make it clear what a term is meant to cover. It can accomplish this to some extent by showing non-preferred terms and semantic relations. Other ways of guiding people in using a thesaurus include introductory matter and scope notes. A scope often takes the form of a definition of the term. Ensuring that terms are used consistently with the same meaning is another major aspect of vocabulary control.

### 12.4.5 Rules for Synthesis

Usually, a thesaurus lists all its preferred terms explicitly. Such Thasauri are enumerative. Some thesauri indicate some preferred terms indirectly : instead of listing all the preferred terms they give rules for creating then out of components.Such thesauri are at best synthetic.

## 12.5 Collecting Terms

Thesaurus construction requires collecting a set of terms. Some of these will end of becoming preferred terms and others may not appear in the thesaurus at all in their original form, but they may suggest concepts that need to be covered in some way.

### 12.5.1 Sources of Terms

Sources from which terms can be collected include existing lists of terms;

other thesauri, indexes, dictionaries, glossaries, etc.

tests from which terms can be extracted titles, abstracts, or full texts of indexed terms queries by patterns.

**People** — Subject specialists, etc.

### 12.5.2 Kinds of terms to be collected

Where possible, terms in a thesaurus should be nouns or noun phrases.

A term should be general enough that it might be used to index a number of items. For example, a theasaurus susally does not include proper names.

But a term should not be so general that it might be used to index too many of the items in the thesaurus' subject area. For example, the term 'NEWS' would not be use in a thesaurus for indexing news items.

## 12.6 Modifying and Inverting Terms

### 12.6.1 Standardizing the form of words

| Guidelines | Examples |
|---|---|
| Plural for things than can be counted | "TUBES" |
| Singular for mass nouns | "WOOD" |
| Singular for processes, properties, and conditions | "REFRIGERATION" "WEIGHT" "POVERTY" |
| Not inverted | "RADAR ANTENNAS" (rather than "ANTENNAS, RADAR") |
| Excluding prepositions | "CARBOHYDRATE METABOLISM" rather than "METABOLISM OF CARBOHY DRES" |
| Excluding punctuation marks, diacritics, special characters and abbreviations | "COOPERATIVE PROGRAMS" (rather than "Co-operative programs" or "Co-operative programs" "MUSICAL NOTES" rather than "(Musical) Notes or "Mus Notes" |

### 12.6.2 Homograph

A homograph is an expression that has the same spelling as another expression, but a different meaning. A thesaurus needs to distinguish between homographs.

A unique term may be created out of a homograph by adding a parenthetical qualifier, for example, "PORT (WINE)"

It may be noted that including parentheses is contrary to the guideline given above: namely to avoid punctuation, a unique term may be created out of a homograph by adding another word without punctuation, for example "PORT WINE"

### 12.6.3 Introducing New Terms

In addition to terms extracted from various sources, we may sometimes choose to introduce new terms of our own.

For example,

(a) Broad concept terms

(b) Structural terms

(c) Terms for nontextual material

#### 12.6.3.1 Broad Concept Terms

Terms that represent broad concepts may be introduced because they are useful in broad searches.

For example, "Traffic Stations", because it can be used to replace a search **for** "Airports or Bus terminals or Train stations or Heliports or ..."

#### 12.6.3.2 Structural Terms

Terms may also be introduced because they help to clarify the structure *of* semantic relations.

**For example,** "Employment of specific groups" to clarify the relationship

between "Employment" and "youth Employment."

#### 12.6.3.3 Terms for Nontextual Material

If you are constructing a thesaurus for indexing material which is not in the form of text, you have fewer sources for terms. You may therefore find yourself inventing your own terms more.

## 12.7 Preferred Terms and Non-preferred Terms

### 12.7.1 Equivalent Terms

After collecting terms for your thesaurus, you need to decide which are equivalent terms. For purposes of indexing and searching, a set of equivalent terms will all be treated as though they meant the same thing and will be represented by a single preferred term.

#### 12.7.1.1 Spelling and Synonyms

Sometimes, equivalent terms really do mean the same thing. So, it obviously makes sense to use a single preferred term to represent that one meaning.

(a) A word may have more than one spelling; For example. "AESTHETICS" and "ESTHETICS."

(b) Two different words may have essentially the same meaning; for example, "AUTOMATION" and "MECHANIZATION."

#### 12.7.1.2 Quasi-synonyms

Sometimes, equivalent terms mean different things in ordinary language. For indexing and retrieval, it is better to group the different meanings together. Such euqivalent terms are called **quasi-synonyms.**

#### 12.7.1.2.1 Types of quasi-synonyms

Terms with overlapping meanings are sometimes treated as equivalent. For example, "GENIUSES" and "PRODIGIES" might be treated as equivalent, even, though the two terms mean different things.

A term whose scope is included in that of another term is cometimes treated as equivalent. For example; "STEEL" might be treated as equivalent to "METAL" if it is not important to distinguish items on steel from items on other metals.

Sometimes **opposites** are treated as equivalent, because items on one are likely to be relevant to a query for the other. For example, "TRANSPARENCY" might be treated as equivalent to "OPACITY."

### 12.7.1.3 USE / UF

**A non-preferred term is normally linked to a corresponding preferred term by a USE reference. The corresponding reference in the opposite direction if US ("Used For") For example,**

| | |
|---|---|
| **Periodicals** | **Serials** |
| **USE Serials** | **UF Periodicals** |

Here the preferred term is "Periodicals".

### 12.7.1.4. Choosing Preferred Terms

The following are some principles for choosing preferred terms, together with examples of applying them.

| **Guidelines** | **Examples** |
|---|---|
| Usage | Cooking |
| | UF Cookery |
| | ("Cooking" is the more commonly used word) |
| Breadth | Plastics |
| | UF Polyethylene |

| | |
|---|---|
| | ("Plastics clearly means all plastics, of when polyethylene is only one) |
| Disambiguation | American Library Association |
| | UF ALA |
| | ("ALA" could stand for something else) |
| Collocation | Railway stations |
| | UF Train stations |
| | **(In an alphabetical sequence, "Railway Stations" would appear near to "Railways" and other terms related to railways)** |
| Conciseness | **Muckrakers** |
| | **UF Muckraking movement** |
| | **(one word rather than two)** |
| **Plural for** | Geese |
| **Countable** | **UF Goose** |
| Countable objects | **(Geese are Countable)** |
| **Internal** Consistency | **If you have decided to prefer the Latin** names for plants, consistencydo so consistently |
| External consistency | You might prefer 'Piers & Wharves[1] to 'Landings', 'Boat Landings', 'Docks', 'Quays', or 'Wharves' partly because thai is what LC Thesaurus for Graphic Material does. |

### 12.7.1.5 Compound USE References

Instead of a single non-preferred term, one may sometimes instruct indexers and searchers to use more one preferred term in combination. In such cases, the USE reference points to all the preferred terms, and the UF reference is often marked in some special way.

For example, —

Snowmobiles

USE Vehicles + Snow

Snow

UF + Snowmobiles

Vehicles

UF + Snowmobiles

you are especially likely to do this if the non-preferred term consists of more than one word.

School Cafeterias

USE Cafeterias + Schools

Cafeterials

UF + School Cafeterias

Schools

UF + School Cafeterias

On the other hand, you may choose not to make such a term a non-preferred term, even if it consists of more than one word.

**12.7.1.6 Making Multi-Word Terms Preferred**

A term consisting of more than one word should typically be made a preferred term if.

(a) combining terms is not possible either at the indexing stage or at the searching stage,

(b) too many terms would otherwise be required to index an item,

(c) the resulting number of preferred terms in not too large,

(d) indexing and searching are generally easier using the compound term,

(e) the term is likely to be used frequently in indexing and searching,

(f) the term's components occur frequently in different syntactic relations; for example, 'Library schools', 'School libraries',

(g) the term is needed in the structure of semantic relations; especially, if

any narrower concepts are represented by preferred terms,

(h) you are in doubt.

## 12.8 Semantic Relations

Indicating semantic relations helps in several aspects in information management :

1. Checking whether a term should be used in indexing a given item or in formulating a given search specification,
2. Choosing the correct level of generality in indexing and searching,
3. searching in response to broad inclusive queries,
4. sharing indexing by facilitating translation from one scheme to another.

### 12.8.1 Semantic Relations between Terms

The main semantic relations indicated between preferred in a thesaurus are hierarchical relations and non-hierarchical relations.

BT and NT links are used to indicate hierarchical relation, one term is viewed as being "above" another term because it is broader in scope. There are various definitions of what constitutes a hierarchical relation. You are advised, however, to restrict yourself to the following cases :

1. Genus / Species : "Animals" is a broader term to "Cats" ("Cats" is a narrower term to "Animals" because all cats are animals. On the other hand, 'Pets' is not a broader term to 'cats' because not all cats are pets.
2. Class / Member : The narrower term can sometimes name a class with only one member.

   For example, "Universities" is a broader term to "University of Calcutta", because to University of Calcutta is a university.
3. (a) Hieraichical whole—Part: In a medical thesaurus, "Head" might be a broader term to "Nose" because noses are normally parts of heads.

   On the other hand, "Forests" would not be a broader term to

"Trees" because not every tree is part of a forest.

(b) Geographical whole / Part : In a hieravchical whole / part relation, both the broader term and the narrower term may name a class with only one member. This is often true of geographical names.

For example, "India" is a broader term to "Kolkata" because "Kolkata" is a part of India.

### 12.8.2 BT, NT and RT References

Normally BT and NT are "inverse" links. For example, if a thesaurus contains the entry.

Pens

BT Writing Materials

you should expect it also to have the entry

writing Materials

NT Pens

A thesaurus is usually "polyhierarchical"; this means that a term can have more than one immediately border term and more than one BT reference. For example,

Social Psychology

BT Psychology

BT Sociology

Polyhierarchy avoids futile arguments about the "best" broader term to choose.

Some terms in a thesaurus have no broader terms and so no BT references. Such terms are usually fairly broad in meaning, at least within the subject area covered by the thesaurus. For example, in a sports thesaurus, "Sport/' might have no broader terms.

An RT reference is used for non-hierarchical semantic relations in a thesaurus. Normally, RT is its own "inverse" link type. For example,

Pens

RT Calligraphy

You would expect it also to have the entry

Calligraphy

RT Pens.

### 12.8.2.1 Semantic Categories of RT References

In constructing your thesaurus, you may find it useful to list categories of Semantic relations that you think should be covered by RT references. Here are some categories sometimes used.

| Categories | Examples |
|---|---|
| Time | Leisure Reading |
| | RT Leisure Time |
| Place | Foreign Languages |
| | RT Language Laboratories |
| Product | Still cameras |
| | RT Photographs |
| | Shipbuilding |
| | RT Ships |
| Cause | Vandalism |
| | RT Hostility |
| Agent | Coaching |
| | RT Coaches |
| Device | Painting |
| | RT Paint Brushes |
| Application | Computers |
| | RT word Processing |
| Part | Vehicles |
| | RT wheels |
| Complement | Parent |
| | RT Children |

## 12.9 Scope Notes

The most common type of guide to applying terms in a thesaurus is the **Scope note.** A scope note is normally preceded by the notation SN. Scope notes take a variety of forms.

A scope note may be a definition; for example,

Space Error

SN Tendency to be biased by the spatial position of stimuli in relation to the observer.

A definition in a scope note should apply to the noun form, not to a related verb or adjective. For example,

This scope note for "Indexing"

SN. To assign Natural Language terms to documents

should be changed to following :

SN Assigning to Natural Language terms to documents.

A thesaurus term should have a single meaning. Any definitions in the term's scope note should reflect that meaning.

For example, this scope note for "Accent"

SN stress placed on a syllable; variation in pronunciation due to linguistic back ground—is incorrect because it confuscs two different meanings of the term.

A scope may indicate a concept that is included in the scope of the cerm; for example,

Mechanised Information Retrieval

SN Includes Pre-computer methods, such punched card systems.

A scope note may indicate a concept that is excluded from the scope of the term. This may be done to show that the term has a narrower meaning than some users of the thesaurus might have in mind. For example,

Bears

SN Does not include Pandas

It may also be done to draw attention to an excluded meaning of

ambiguous term; for example,

Parties

SN Political parties only, do not use for social gatherings.

Some scope notes refer to other terms, especially to indicate how to deal with a concept that is excluded; for example,

Licensing

SN Excludes aspects covered by the terms 'School Accreditation* and Teacher Accredifation".

A scope note may give additional instructions to indexers. For example, it may remind indexers of other types of terms that they should assign :

Hospitalization

SN Assign also terms for the conditions for which patients were hospitalized, if applicable.

A scope note may suggest that the term not be used if a more specific term is appropriate; for example.

Equipment

SN Broad term; Prefer terms specifying types of equipment if possible; for example,' office Equipment'

In a synthetic thesaurus, instructions for synthesis may appear in scope notes; for example,

History

SN Append also as a subdivision after terms designating disciplines, activities, living, things, etc. For example, 'Intercropping—History',

"Goods—History'.

Information included in a scope note should be helpful to users of the thesaurus as indexers or searchers. A thesaurus is not a dictionary, an encyclopedia, or even an index."

## 12.10 Thesaurus Displays

For any thesaurus display, you may needd to make several decisions. The decisions are likely to include

1. which types of terms will have entries?
2. how to indicate special types of terms?
3. what types of links will be shown to other terms?
4. how many levels linking will be shown?
5. how to indicate link types?
6. where the linked terms are placed relative to the entry term relative to each other?

1. Which types of Terms will have Entries?

A thesaurus display might have entries only preferred terms; for example.....

Ex-convicts

Eye Examinations

Eye Patches

Eyeglasses

Eyes

Fables

At least one of the displays, however, should provide entries for non-preferred terms as well, to allow users to browse through these for the correct preferred terms:....

Extremism

Ex-Convicts

Ex-Military Personnel

Eye-Catchers

Eye-Examinations

Eye-Patches

Eyeglasses

Eyes

Fables

Fabric Design Drawings

2. How to Indicate Special Types of Terms.

You may want to mark certain kinds of terms in special ways. For example, you might put all the non-preferred terms in italics :...

*Extremism*

Ex-Convicts

*Ex-Military Personnel*

*Eye-Catchers*

Eye-Examinations

Eye Patches

Eyeglasses

Eyes

Fables

of course, users of the thesaurus should be able to tell that a term is a non-preferred term if it has a USE reference after it, bat displaying the term differently will serve as an added reminder. You may prefer a mixture of upper and lower case for your thesaurus displays to make them easier to read. Mixing upper and lower case may be especially helpful for longer elements such as sope notes :

Goggles

SN Protective eye coverings

3. What Types of Links should be shown to other terms.

Taken as a whole, your thesaurus displays should cover all the term links that are important to the people who will use the thesaurus. In individual displays, you may choose to include only certain links. For example, a brief display might include only 'USE' references :

*Extremism*

USE Radicalism

Ex-Convicts

*Ex-Military Personnel*

USE Veterans

*Eye-Catchers*

USE Architectural Follies

Eye-Examinations

Eye Patches

Eyeglasses

Eyes

Fables

*Fabric Design Drawings*

USE Textile Design Drawings.

At bast one of the displays would include the scope notes : Alidades

SN Telescoping siting devices used as part of a shop's navigational equipment

for taking bearmgs.

BT Scientific Equipment

BT Telescopes

RT Navigation

An entry in one of displays will usually give all links :

Eyeglasses

UF Spectacles

BT Medical Equipment and Supplies

BT Optical Devices

NT Monocles

NT Sunglasses

RT Contact Lenses

RT Goggles.

4. How many Levels of Linking will he shown. In one of your displays you may wish to show indirectly linked terms as will as those linked directly to the entry term, litis is mostly used with links representing hierarchical relations. The display could indicate more than one level of broader term :

Monocles

BT Eyeglasses

BT Optical Devices

BT Equipment.

BT Medical Equipment and Supplies.

Likewise, the display could indicate more than one level of narrower term :

Optical Devices

NT Binoculars

NT Contact Lenses

NT Eyeglasses

NT Monocles

NT Sunglasses

NT Goggles.

5. How to indicate Link Types.

You can often omit symbols for different kinds of links if it is obvious what they are. So; you need to use a symbol such as RT only once.

RT Employee Eating Facilities

Employee Fringe Benefits

Employee Rights

Employment

Unemployed.

6. Whare the Linked Terms are placed

In a printed display, you will usually want entry term to appear at the top left, because this makes it easy to search for. Variations are possible, though.

- Medical Equipment and Supplies
- Equipment
- Optical Devices
- Eyeglasses
- Monocles
- Sunglasses

If all the links are indicated in the same position relative to the entry term, the best order to follow is generally

Scope notes

non-preferred-equivalent terms

broader terms

narrower terms

related terms.

For example.

Employees

SN Persons identified as working for another, but where the nature of the occupation, business, or industing is not known.

UF Personnel

Staff

Workers

BT People

NT Hotel Employees

Railroad Employees

RT Employee Eating Facilities

Employee Fringe Benefits

Employee Rights

Employment

Labourets

Unemployed

Within a group of terms linked in the some way to the entry term, the order is most commonly alphabetical. Sometimes, however, you may wish to adopt a systematic order by subcategories the link types, especially if the lists are very long.

## 12.11 Summary

The thesaurus is one of the Categories of Controlled Vocabularies. The unit introduces you the tool for vocabulary control. It discusses the compounds covered in a thesaurus viz preferred terms, semantic relations between terms and the like. It describes the sources from which terms can be collected, standardizing the form of words, relationship between BT and NT and semantic categories of RT references.

It discusses the scope notes which give definitions, indicate which concepts are included or excluded, refer to other terms and provide additional instructions. Finally the unit explains the display of thesaurus.

## 12.12 Exercise

1. How is a thesaurus designed for indexing and searching in a specific subject area?
2. How is a thesaurus developed? Discuss the stalagmitic and statistics approaches.

3. Indicate the sources from which terms can be collected tor the construction of a thesaurus. What kinds of terms should you collect?
4. Discuss the principles for choosing preferred terms in a thesaurus.
5. What is the relationship between BT and NT? How many BT references can a term have?
6. Illustrate the semantic categories of RT references.
7. Which types of terms will have entries in a thesaurus?

## 12.13 Glossary

**Broader term**

The superordinate term in a **hierarchical relation.**

**BT**

A symbol used in a *thesaurus* to identify the following terms as broader terms to the heading term.

**Chain**

A sequence of terms in which the class represented by term includes all the classes represented by the terms that follow that term.

**Controlled indexing**

Indexing with terms from a controlled vocabulary, such as a *thesaurus.*

**Enumerative** Of an indexing or classification scheme, listing or enumerating terms explicitly, rather than making provision for synthesizing them.

**Equivalent term** A term in a controlled vocabulary, such as a *thesaurus,* that is treated as if it means the same thing as another term.

**Extraction indexing** Indexing with terms from the text or title of the item indexed.

**Hierarchical relation** A *semantic relation* in which one term is strictly

subordinate to the other; for example, a genus / species relation.

**Homograph** A term with the same spelling as, but a different meaning from, another term.

**Narrower term** The subordinate term in a *hierarchical relation.*

**Non-preferred term** A term in a controlled vocabulary, such as a *thesaurus,* that may not be used as *controlled indexing* term, but may be looked up.

**NT** A symbol used in a *thesaurus* to identify the following terms as *narrower terms* to the heading term.

**Postchlorination** The combination of terms at the time of retrieval to form a compound search specification that corresponds to no single index term used in indexing.

**Precision ratio** The ratio of the number of *relevant* items retrived to the total number of items retrieved.

**Procrastination** The combination of terms or other components to form compound terms at or before the time of indexing.

**Preferred term** A term in a controlled vocabulary, such as a *thesaurus* that may be used as a *controlled indexing* term.

**Quasi-synonym** An *equivalent term* that is not a synonym.

**Recall ratio** The ratio of the number of *relevant* items retrieved to the total number of *relevant* items indexed.

**Related term** A term in a *semantic relation,* but not in a *hierarchical relation,* to another term.

**Relevant** Appropriately retrieved in response to given query.

**RT** A symbol used in a *thesaurus* to indicate that the following terms are *related terms* to the heading term.

**Scope note** A note attached to a term in a controlled vocabulary, such as a thesaurus, that gives guidance on how to use the term.

**Semantic relation** A relation between terms that is true as a matter of

general knowledge, rather than depending on what the terms to in some particular document.

**SN** A symbol used in a *thesaurus* to indicate that what follows is a *scope* note to the heading term.

**Syntactic relation** A relation between terms that depends on what the terms refer to in some particular document, rather than being true as a matter of general knowledge.

**Synthetic** Of an indexing or classification scheme, making provision for the synthesis of terms out of components, rather than listing or enumerating them explicitly.

**Thesaurus** A device for vocabulary control, usually for a specific subject area, indicating preferred terms, *non-preferred terms*, and *semantic relations* between terms; the terms are in ordinary human language.

**UF** A symbol used in a *thesaurus* to indicate that the following term or terms are not used in indexing and that the heading term is used instead.

**USE** A symbol used in a *thesaurus* to indicate that the heading term is not used in indexing and that the following term or terms are used instead.

## 12.14 References and Further Reading


1. Chakraborty, A. R. and Chakraborti, B. : Indexing : principles, processes and products, World Press, 1984.
2. Mahapatra, P. K. and Chakrabarti, B : Knowledge managements in libraries. Ess Ess, 2002.
3. Svenonius, E : 'Unaanswered questions in the design of controlled vocabularies'. JASIS 1986, 37, 330-40.
4. Taylor, Airline, G. Ed : Wynar's Introduction to cataloging and classification. 9th ed., Libraries Unlimited, 2000.
5. Taylor, Airline G : The organization of information. Libraries Unlimited, 1999.

# Unit 13 ❐ Trends in Automatic Indexing

Structure



## 13.0 Objectives

Information, recorded formally in different types of documents needs to be represented before it can retrieved. After reading this unit you will able to :

(a) understand the developments in automatic indexing

(b) understand the extraction indexing

(c) understand assignment indexing in which terms used in the index are not necessarily those found in the text.

(d) know the earlier crude form of automatic indexing KWIC.

(e) gain knowledge about different types of automatic indexing,

## 13.1 Introduction

The retrieval activity takes place with document surrogates in the form of indexes, abstracts, summaries, and the like. Ideally the representation process should be conducted simply and efficiently, As Lesk, Michael noted in 1997 :

"If we had a single knowledge representation scheme that let us put each idea in one place, and if the users knew this scheme and could place each of their queries in it, subject retrieval would be straightforward .... In practice, it seems unlikely that any single knowledge representation scheme will serve all purposes."

Furthermore, how the representation scheme could be applied consistently and precisely still poses challenges to information professionals. Indexers like other human beings, are fallible, often are inconsistent, are subject to extraneous influences on their work, operate at a slow pace, and are therefore the most expensive component of an indexing operation. The idea of replacing human indexers by feeding part or all of a text into a machine that would assign index terms automatically, impartially, and with unfailing inconsistency and accuracy arose, therefore, quite early in the computer age.

## 13.2 Indexing

Indexing has been a widely adopted method for information representation. It uses terms (that is, words or phrases), either derived or assigned to represent important facets of the original document. Conceptual analysis and translation of the document to be indexed constitute the intellectual portion of the indexing practice. To be more precise, conceptual analysis in indexing involves the identification of key concepts covered in the document whereas the translation phase actually converts a chosen concept to index terms based on preselected indexing language.

### 13.2.1 Types of Indexing

Indexing type is normally dependent on how the terms are obtained. If the terms are extracted from the original text of a document, it is called derivative indexing. On the other hand, if the terms are assigned to a document, it is labeled as assignment indexing. Derivative indexing can also be treated as a synonym for keyword indexing because index terms are selected directly from keywords in the text and no controlled vocabulary is consulted. By comparison, a controlled vocabulary must be used in assigned indexing for choosing appropriate terms.

Indexing terms assigned from a controlled vocabulary are customarily called descriptors even though the controlled vocabulary used might not be a thesaurus.

### 13.2.2 Automated and Automatic Indexing

All the activities involved in indexing can be categorised into two types : intellectual and mechanical while the intellectual portion of indexing remains to be accomplished by human beings most of the time, indexing operations of mechanical nature can be completed satisfactorily by computers. If computers are used to handle both the mechanical and intellectual parts of indexing, it is automatic indexing. If computers are used only for mechanical operations of indexing and human indexers are responsible for the intellectual portion of indexing, it is automated indexing.

Automated indexing can rescue human indexers from the tedious and the repetitive indexing tasks so that they can devote more to the intellectual undertakings of indexing. On the other hand, the automatic indexing can cope with the situation when the amount of information is gigantic.

### 13.2.3 Indexing in the Hyperstructure Environment

More and more information becomes available in the hyperstructure environment, which is symbolised by the World Wide Web. For Web-based information, index terms are most naturally presented as hyperlinks that

embody both the index term and the locator mechanism. In other words, hyperlink names serve as index terms whereas the hyperlink michanism seamlessly leads the user to where the index terms are pointing. This setting is unique in the following aspects.

1. Index terms in the hyperstructure environment are embedded within the document itself rather than separate entities outside the context.
2. Index terms locators are merged into one unit.
3. The subject hierarchy in traditional indexing is not exactly reflected in the hyperstructure environment.
4. Only content-based links in the hyperstructure can be regarded as index terms.
5. Authors assume the role of indexing when they prepare hyperstructure documents and indexing is done simultaneously, if not before, as the document is created.
6. Less discrepancy will occur between the original document and the index terms in this process of information representation as the author decides which terms should become links while the document is being written, unlike in traditional indexing where author writes a document and then indexer analyses it for representation purpose.

## 13.3 KWIC and KWOC Indexing

Research into automatic indexing has been progressing since the late 1950's. The earliest automatic indexing method relying on the power of computers to perform repetitive tasks at high speed was invented by Hans Peter Luhn, an IBM Engineer, who in 1958 produced what became known as KWIC (Key Word in Context) indexing. Luhn reported his system in 1960. On the assumption that titles of scientific and technical articles generally include words indicating the most significant concepts dealt with, he wrote a program that printed strings of title words, each word appearing once in alphabetical order in the centre of a page, with all other words to the left or right of the centre word printed in the order in which they appeared in the title; when the right—hand margin was reached, the rest of the title (if any) was "wrapped around" to the left-hand margin and continued inward. A user had only to scan the left-justified middle column for a desired keyword

and could, when the word was found, read the rest of the title "in context". Most KWIC programs employ stop lists to eliminate common words such as articles prepositions, and conjunctions from the middle column.

An adaptation of the KWIC method, known as KWOC (Keyword Out of Context) simply prints the sought words in the left-hand margin instead of in the middle of the page, the rest of the title (or the entire title) being printed to the right or beneath the keyword.

### 13.3.1 Extraction of Words

KWIC indexing was only the first of the so-called derivative indexing methods, all of which are based on the principle of extracting words from machine-readable text—a title, an abstract, or even the full text of a document. Automatic extraction of words is generally complied with truncation in searching, that is, the possibility of searching for a word stem without regard to its prefixes or suffixes, in order to retrieve a maximum of potentially useful occurrences of that word. Thus, a physicist looking for the presence of the concept "pressure" may search for *PRESS* (the asterisks indicating that prefixes and suffixes are also to be searched), which may give :

COMPRESS
COMPRESSION
IMPRESSION
SUPPRESSION
PRESS
**PRESSER**
**PRESSES**
PRESSURE
PRESSURIZE
PRESSURIZATION
PRESSWORK

While truncation (or "stemming") does not increase recall, it lowers precision because it may result in unwanted and irrelevant items being retrieved (the latter known as "false drops"). At least two conditions dictate this. **First homonyms** cannot **be detected** by mere extraction methods—that is, in above-noted example, **IMPRESSION, SUPPRESSION and PRESSWORK** do **not** pertain to "pressure" in **the** physical sense, **while**

PRESS may pertain both to mechanical equipment and to newspapers, the latter being of no interest to a physicist. Second, the elimination of common words by a stop list may also result in false, drops whenever relationships are of importance, for example, a Boolean search (and, or, not) of TEACHERS and STUDENTS and EVALUATION will retrieve both teacher's evaluation of students and students' evaluation of teachers, because the elimination of the crucial words "of" and "by" makes it impossible to know who does what to whom.

For a time there was attempt to correct the lack of indicators of relationships in derivative indexing by links and roles, the former making explicit which words were linked to each other in a relationship, while the latter indicated functions (for example, acting "as" or "for" something). These devices assured higher precision, but they had to be assigned at the input stage by human beings. Finally, this method was abandoned as that greatly diminished any gains made by automatic extraction of terms.

## 13.4 Term Frequency Methods

On the assumption that terms (other than common words) to be indexed are those occurring either very frequently in a text or very seldom, methods were designed to perform automatic indexing on the basis of frequency of occurrence and co-occurrence of terms, using probabilistic models. Some investigators tried to couple such methods of determining how often a term is used (term frequency methods) with term weighting that is, assigning different degrees of importance to terms on the basis of what terms are used in a search request or on the basis of where and how terms appear (for example, in the title, in an abstract, or in the first or last paragraph of a text, and whether they are italicized or capitalized), all of which can to some extent be determined automatically. But these methods have not yet found any practical large scale application.

## 13.5 Linguistic Methods

A quite different approach to automatic indexing is by syntactic and semantic analysis. The former is concerned with the automatic recognition

of significant word order in a phrase or sentence and with inflections, prefixes and suffixes that indicate grammatical relationships, while the latter approach seeks to analyse noun phrases automatically with the aid of stored dictionaries and other linguistic aids. Both the methods are often used in conjunction. Research in the field of Natural Language Processing has proliferated in recent years, progress has been achieved. But the dream of information retrieval systems that can respond to a user's need through the use of natural language queries in full text databases analyzed only by computer remains unrealized.

## 13.6 Impact of automatic methods on professionals

As computer programs become more sophisticated, and more information appears in electronic form, there will eventually be less 'traditional' indexing work available. This loss may be balanced in the short—term by an increase in the number of indexing projects attempted. The proportion of freelance versus in-house work may also change. Humans should still be used for important works, which perhaps can be identified by studying usage and citation patterns. Indexers will have to become more selective, and decide on the quality of the works they might index as well as the subject content.

If we remain better than computers we must show this, and indicate that there are economic returns and academic returns from a quality index. On the positive side, indexing skills will be needed in the development of computer systems, and to check the output from computers. Indexers will be needed to set up and maintain thesauri and to train writers as 'bottom-up indexers' so that their work is readily retrievable.

Indexers will have to become entrepreneurial and computer literate. Indexers with skills in the related areas of computing, editing, librarianship and bibliography may be best suited to take advantage of new opportunities. We will have to be able to identify gaps in a commercially effective way.

To do this we will have to be computer literate. Not only will we have to know how to use various computer tools for indexing; we will also have to know how information in organised and used electronically, so that we can best understand the needs and make our own contributions.

## 13.7 Summary

Here we have explained how representation process could be efficiently made. Derivative and assigned indexing have been explained. Activities in indexing can be categorised as automated and automatic types. Indexing in the hyperstructure environment has been discussed. KWIC and KWOC indexing processes are highlighted. KWIC indexing method is based on the principle of extracting words. Other methods like 'term frequency methods' and linguistic methods' have also been noted. Finally impact of automatic indexing on professionals is worthy of note lo make them computer literate.

## 13.8 Keywords

1. Indexing : The intellectual analysis of the subject matter *of* a document to identify the concepts represented in the document and the allocation of descriptors to allow these concepts to be retrieved.
2. Keyword : A synonym for index term.
3. Stemming : It is used to conflate, or reduce, morphologic variants of a word to a single index term.

## 13.9 Exercise

1. What is understood by 'Automated' and 'automatic' indexing?
2. Who do you mean by 'hyperstructure environment'?
3. Explain KWIC indexing method.
4. Why does truncation lower precision?
5. Explain term frequency methods in indexing.

## 13.10 References and Further Reading

1. Chu, Heting : Information representation and retrieval in the digital age. Thomas H. Hogan Sr. 2003.

2. Foskett, A. C. : The subject approach to information 5th ed., London, Library Association, 1996.
3. Luhn, Hans Peter : "Keyword in Context Index for Technical Literature (KWIC Index)", **American Documentation,** 1960, 11, 288-295.
4. Taylor, Arlene G : The organisation of information. Libraries Unlimited, 1999. pp 158-166.

# Unit 14 ❐ IR Models

Structure

## 14.0 Objectives

After reading this unit you will be able to :—

1. understand the bases for modeling IR activities.
2. classify the IR models.
3. understand the relationship between IR models and retrieval techniques.
4. choose the right system for retrieval tasks.

## 14.1 Introduction

A model can be defined as "a tentative description of a theory or system that accounts for all of its known properties." (Soukhanov, A. H., et al Eds. 1984). There are varieties of models developed in information retrieval over the years. In this unit we shall explore some common IR models in order to facilitate the understanding and appreciation of the foundation of IR practices. There are many levels at which IR procedures can be modeled. Theories and notions from other disciplines for example, Boolean logic, vector space, and probability models, are drawn to form the bases for modeling IR activities. Various schemes have been evolved to classify all the IR models developed so far. The authorities like Baeza—Yates & Ribeiro-Neto, Belkin & Croft, Sparck Jone & Willett may be worthy of note in this context. This unit will focus on discussing system-oriented models such as Boolean logic, vector space and probability, truncation & Fuzzy Set.

## 14.2 Foundation of IR Models : Matching

Matching is the fundamental mechanism in information retrieval whether a search will be successful or not depends only if a match is found between the represented information in the system and the query submitted by the user. But it should be noted that matching is not a model in information retrieval. Matching can be made between terms or between similarity measurements such as distance and term frequency. Term matching is performed directly on terms derived from or assigned to documents, or

their representations while similarity measurement matching is conducted indirectly on measurements abtained by calculating, for example, between vectors as in the vector space model. These two kinds of matching will be discussed in the following two subsections.

### 14.2.1 Term Matching

Terms used in Information Representation and Retrieval can be keywords, descriptors or identifiers. Terms also include words, phrases, or other kinds of expressions. In addition, term matching can be done in one of the following four ways : exact matching, partial matching, positional matching and range match.

Exact match retrieval techniques are those that require that the request model be contained, precisely as represented in the query formulation, within the text representation. Case sensitivity search and phrase search are examples of the exact matching. For example, *web filtering* is a term in a query and the some phrase also appears in the system to the searched. Then an, exact match is obtained as a result of searching.

The first distinction that we make among retrieval techniques is whether the set of retrieved documents contains only documents whose representations are an exact match with the query or a partial match with the query. For a partial match, the set retrieved documents will include also those that are an exact match with the query. Unlike exact matching, partial match only has part of the term being matched with the doucment representation in the IR (Information Retrieval) system. Truncation in searching is a typical example of partial matching. For instance, a search query of *information technology** (where* is the truncation symbol) would retrieve documents containing *information technology, information technologies, information technologists* in a result of partial matching. Positional match is done by taking into consideration the positional informations of what is being matched in the process. If a search query reads used "WITH Store, the retrieved results would include document containing phrases such as used book store, used cloth store, used furniture store. Here the matching is only done with the first and last words given between the query and element representation while the word in the middle position is overlooked during the matching process.

Range match is applicable to numeric expressions (for example, sales amount) or expressions with a natural order (for example, January, Febuary ... December).

What is being matched in range matching is the upper limit of the range (Publications before your 2006), lower limit of the range (for example,publications after 2000), or both (for example, publication between 1998 and 2003). Numeric databases and publication data are conventional examples of range matching.

All these four different types of matching deal with the original queries and document representations without any further calculation or conversion. Term matching is frequently found in the Boolean logic model. In other IR models (for example, vector space and probability), terms in query and document representations are not directly matched Rather, they would be trausformed as similarity measurements before being matched.

### 14.2.2 Similarity Measurement Matching

Similarity measurement matching is done in a number of different ways. For example, matching is based on the distance between vectors or degree of vector angle in the vector space model : The smaller the vector angle, the higher degree of minilarities between queries and documents. In the probabilistic model, similarity can be calculated based on term frequency to determine the probability of relevance between queries and documents.

In matching of this kind, a similarity measure other than terms themselves should be chosen. It is on this measurement that matching is finally done. On the hand, the similarity measurement matching provides additional approaches to accomplishing the retrieval task. On the other hand, such practice can introduce mistakes and noises, particularly in the process of obtaining similarity measurement.

Regardless of the matching type, matching is the essential mechanism in information retrieval. Here we shall discuss how matchingis done in different circumstances along with their respective features, advantages and limitations.

## 14.3 The Boolean Logic Model

The Boolean logic model is named after George Boole who proposed Boolean logic in the mid-19th century. Boolean logic consists of three logical operations : the logical product (x). the logical sum (+) and the logical difference (-). These corresponding operators AND, OR, NOT are used to express the logical operations in information retrieval.

The logical product, or AND operator, combines two or more terms in a search statement and requires all the terms be present in the documents to be retrieved. The logical sum, or OR operator, iinks two to more synonyms or related terms in a search statement. Documents containing any one of the terms specified in the search statement would be regarded a hits or anticipated resulted results. The logical difference, or NOT operator, limits searches by excluding terms listed after the NOT operator in queries.

### 14.3.1 Strengths

The extensive use of the Boolean logic model in information retrieval proves the value of the model. First, it supports the manipulation of different facets obtained by decomposing a query or document. The AND operator combines two simple facets to form a complex one, thus narrowing down the search. The OR operator allows the specification of alternative facets of a query or document, thus broadening a search. The NOT operator can separate facets into individual simple ones thus excluding unwanted facets from appearing in final search results. Such manipulations. if applied appropriate!, can bring flexibility, and effectivencess to information retrieval at a level with which no other existing IR models can compete.

According to Frants Valery I, et al, Boolean logic IR systems are cost effective and already indispensable to the user. Thousands of Boolean logic IR systems are operational. The end user can make of the Boolean operators to broaden, narrow, or eliminate certain results from a search. In that sense, the Boolean logic models well established although Belkin, Nicholas and Croft, W. Bruce commented that the model became established more through practice than theory.

Sparck Jones and Willett suggest that the Boolean logic model is well understood even though there is much less discussion about what the model can accomplish than what the model is unasble to do, which might be a result of two factors. One is that the Boolean logic model is the oldest among all IR models. It is assumed that its strengths have been will appreciated so no further elaboration is expected. The other factor is that Boolean logic, as the oldest IR model, will have to criticized whenever a new IR model is presented.

Finally, Boolean IR systems are relatively easy to build, as the algorithms involved appear simpler to implement than those based on other IR models. That may also contribute in part to the wide implementation of Boolean IR systems.

### 14.3.2 Limitations

It is difficult for the user to conduct Boolean searches without a fair degree of training and practice. The limitations of this model have been explored by many authors viz korfhage, Sparck Jones & Willett and others. According to them, the difficulty lies in the following two aspects :

(a) It is difficult for the user to choose the right Boolean operator. Usually there exists confusion about the AND and OR operators among users as these words have different meanings in conventional sense. The word And traditionally means 'plus' as in, for example, "They will search for this topic at AltaVista *and* Google," while the word OR implies "either one as in, for example. "They will search for this topic at AltaVista or Google." Literally, more sites will be searched in the former cases than in the latter one. Some users customarily think the same when they form Boolean queries that is, use AND operator when they require more results and use the OR operator when they desire less. Boolean logic obviousit does not work that way.

(b) It is difficult for the user to correctly employ the order of processing in compound Boolean searching. Compound Boolean searching involves more than one type of Boolean operator whose natural order of Processing is defined as : the NOT operator first, the AND

operator seeond. and the OR operator last. However, the natural order of processing for compound Boolean searching can be altered by applying parentheses when needed. Such alteration is, nevertheless, complicated for novices. It is difficult to express relationships other than the Boolean among terms (for example, casual relationship) because such mechanism is not provided in the model. Suppose a user likes to find some information about "the application of computers in education." A search query may than be formed using the Boolean operator as **Computers** AND **education.** The term application is not included in the query because this relationship is supposed to be expressed Via Boolean operators. Consequently, the used would find information on not only the use of computers in education" but also "education about computers" by conducting the search.

The Boolean logic model does not have any weighting the relative importance of different concepts in a search query. All the terms or concepts in Boolean queries are assumed to have the same degree of significance, which is not always true in information retrieval.

The user may get null output overload when performing Boolean searching. Null output is possible if the search query is very restrictive when, for example, several terms are connected together by the AND operator. Output overload may occur if the search query is broad when, for example, several terms are Jinked together by the OR operator.

In order to cope with the limitations of the Boolean logic model W. S. Cooper suggested some possible solutions, such as formulating queries free of Boolean operators to manage the unfriendliuess of Boolean queries. A number of methods including IR models, have been developed for providing ranked outouts, weighted outputs and the like.

## 14.4 Vector Space Model

The vector space model has been developed by Salton and his colleagues who built the SMART system to serve as the base for conducting extended series of IR experiments. A series of IR techniques for example, term weighting, ranked output and relevance feedback, have been devised in the process of

building the model. In the vector space model, each term is defined as a dimension while each query or document is expressed as a vector. A vector actually consists of a list of term values representing an item that is, a query of document. Term values in a vector could be either binary or weighted. The binary values could be either one or Zero, with one indicating the existence of the term in the item. The weighted values include real positive numbers for example, 1.5, 0.4, 2.5, 6.5. The weighted value for each term corresponds to the relative importance of that terms in representing the item.

Conducting a search in the vector space IR system means to check the distance, shown as an angle between the query and document vectors in the space. The vector space model judges the similarity between a document and a query or between any two documents by comparing their corresponding dimensions. If the angle between their vectors is small, the query and the document will be on a similar topic. If a query and a document are on different topics, the angle between their vectors should be large.

### 14.4.1 Strengths

When conducting searches in a vector space IR system the user is not required to understand and apply Boolean logic. The user only has to select several terms based on his or her information need.

Second, the terms or concepts chosen to represent the query or document can be weighted to indicate their relative importance in the vector.

Third, the output of searches using vector space IR system can be ranked in dexreasing relevance.

Fourth, relevance feedback mechanism can be implemented in this model to improve retrieval performance. It is worth to note that the vector space model can overcome some limitations associated with the Boolean logic model. But it has also introduced some new problems in the field of information, retrieval.

### 14.4.2 Limitations

The first limitation of this model refers to the assumptions of independence between terms selected for describing a vector.

The second limitation is the difficulty in specifying explicitly synonymous or phrasal relationships owing to the absence of the Boolean operators. For example, the OR operators cannot be used to indicate synonyms for example, cars and automobiles and the WITH operator cannot be used to form phrases, for example, Information WITH Retrieval in a vector.

The third limitation of this model relates to its weighting mechanism, which can be subjective and complex.

Apart from three limitations as noted above, the vector space model also needs several terms to represent a query or document so the vector would be dicriminating enough for good retrieval performance.

The vector space model has not appled properly unit the advent of Internet retrieval systems. The SMART system actually enables the model to grow and develop.

## 14.5 Probability Model

M.E. Maron and Juhns, J. L. introduced the probability model in 1960 and further it was developed by Robertson and other researchers. The model applies the theory of probability, for example, an event has a possibility from 0 percent to 100 percent of occurring, to information retrieval. It takes into consideration the uncertainty element in the IR process, that is, uncertainty about whether documents retrieved by the system are relevant to a given query.

The model adopts various methods to determine the probability of relevance between queries and documents. Relevance is judged according to the similarity between queries and documents. The similarity judgement is further dependent on term frequency. Generally speaking, the more similarity exists and documents, the higher probability that those documents are relevant to the queries.

### 14.5,1 Strengths

The retrieval processes are characterized with a degree of uncertainty when relevance between queries and documents is judged. The query-document similarity measurement is determined by the model itself instead

of some arbitrary decision as in vector space model.

The model weights query terms and association between queries and documents so that users can specify the relative importance of the term or association in the retrieval task. Ranked output is also provided since the model assumes that the principal function of an IR system is to rank the documents in a collection in order of decreasing probability of relevance to a users information need. Thus both the weighting and ranking are expressed in this model. Third, the model can take advantage of feedback information to develop well-founded methods. Moreover, self-improving feature adds another positive element to this model.

Fourth, IR systems based on this model are more user-friendly than Boolean system.

### 14.5.2 Limitations

First although the relevance value in this model is continuous instead of the dichotomous zero or one as in the Boolean model, the probability model assumes that relevance has a binary property :

$$\underset{\text{(non-relevance)}}{Pr} = \underset{\text{(relevance)}}{I - Pr}$$

The value for the probability of non-relevance is fixed once the probability of relevance is computed, thus eliminating the intrinsic uncertainty factor from the IR process.

Second, the probability model cannot improve retrieved effectiveness. Apart from these two limitations, the model has other weaknesses. For example, sophistication in mathematics is necessary to understand and use the probability theory.

## 14.6 Fuzzy-Set Model

Zadeh, L. A. proposed in 1965 the fuzzy set model to overcome one limitation of the Booleam logic model by generalising the traditional set theory. In a traditional set object is either in the set or not in the set. Linewise, a document is either relevant or not relevant to a given query in the Boolean

system. But such a clearcut boundary does not exist in information retrieval. There may be a partial relevance factor. The model assumes that fuzzy documents and query do not exist but fuzzy judgment can be made. The boundary that divides members and nonmembers of a set thus becomes fuzzy.

A membership grade can be assigned to a document to indicate how close it belongs to set of relevant documents. Membership grades for the fuzzy set of documents are determined by the indexer during the indexing process.

### 14.6.1 Strengths

The strengths of this model include the relaxation of restriction in the Boolean logic model, namely, documents are either relevant or non-relevant to a query and no partial relevance is allowed. In assigning the membership grade, this model can provide ranked output in decreasing order of relevance. In addition, the Boolean query structure is preserved in the model for expressing logical relationships.

### 14.6.2 Limitations

The fuzzy set IR system does assign weights to query terms as opposed to document terms. As such the system is not as flexible as desired. It provides no mechanism for query expansion if compared with vector space model. In contrast to the probability model, the fuzzy set model is less robust at least theoretically.

## 14.7 Summary

We have discussed different IR models along with their pros and cons. IR systems applying these models consequently can only perform certain retrieval functions. In order to take advantage in the retrieval techniques different IR models have been described along with their strenghts and limitations. The Boolean logic model seems to be the weakest among all. But the model is heavily used and applied widely in information retrieval. The

criteria used by the vector space and probability models for weighting terms, ranking outputs and measuring similarity are nevertheless dissimilar. In addition relevance feedback is introduced as a unique retrieval capability in vector space model.

## 14.8 Exercise

1. What is a model? Describe the major models in information retrieval.
2. What do you understand by 'Term matching'? Discuss various types of term matching.
3. Discuss the extensive application of Boolean logic model.
4. Discuss strengths and limitations of vector space model.
5. What is rationale for introducting probability model?
6. How does fuzzy set **model overcome** the limitation of **the** Boolean logic model?

## 14.9 Key Words

1. **Feedback** : They are used to refune the request model, which is then used for another search. Feedback techniques are an extremely important part of ensuring that a document retrieval system will be effective.
2. **Fuzzy Set** : This is an integration of Boolean queries with ranking techniques. This integration is limited, however, when compared with extended Boolean retrieval based on the vector space model or the use of term dependencies in probabilistic model.
3. **Probabilistic model** : The basic aim is to retrieve documents in order of their probability of relevance to the query.
4. **SMART** : System for the Manipulation and Retrieval of Texts which is humorously known as salton's Magical Automatic Retriever of Text,
5. **Vector Space model** : Here documents and queries are vectors in an n-dimensional space, where each dimension corresponds to an index term.

## 14.10 References and Further Reading


1. Belkin, Nicholas J. and Croft, W. Bruce : Retrieval techniques. Annual Review of Information Science and Technology. 22, 1987, 109-145.
2. Cooper, W. S. : Getting beyond Booie. **Information. Processing and Management,** 1988, 24, 243-248.
3. Chu, Heting : Information representation and retrieval, in the digital age. Thomas H. Hogan Sr. 2003.
4. Korfhage, Robert R. : Information storage and retrieval. New Yourk, John Wiley & Sons, 1997.
5. **Maron, M. E. and Kuhns, J. L. : On** relevance, probabilistic indexing and information retrieval. **Journal of the ACM,** 1960, 7.216-244.
6. Robertson, S. E. : The probability ranking principle in IR J. **Doc,** 1977, 33, 294-304
7. Sparck Jones, Karen and Willett, Peter (Eds). Readings in information retrieval. San Francisco, Morgan Kunfmann, 1997.
8. Zadeh, L. A. Fuzzy sets. **Information and Control,** 1965. 8, 338-353.

# Unit 15 ❐ Retrieval Techniques

Structure

## 15.0 Objectives

The main objective of this unit is to demonstrate the role of different retrieval techniques for developing truly effective information systems. As a memorable aphorism preface his novel Howard's End. E. M. Froster gave simply "Only connect". It could claim to be the finest, even though briefest, definition of intelligence we have. To understand anything, whether it is the operation of a complicated mechanism or the complex social factors that underlie almost any human situation, understanding it means seeing the connection.

## 15.1 Introduction

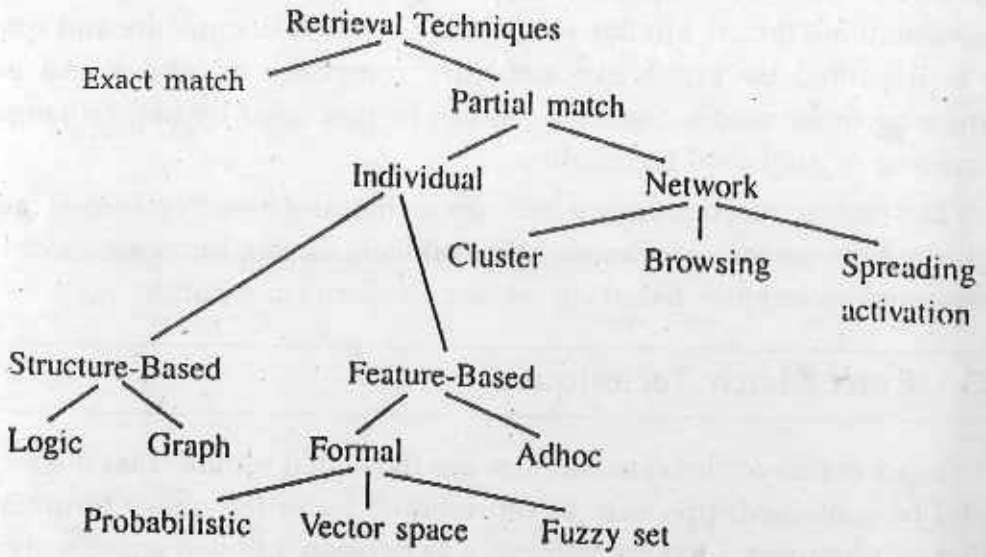
An IR system comprises the people, activities and equipment concerned with the acquistion, organisation and retrieval of information. Discussion of IR usually assumes that an IR system is computer-based. Whist this is increasingly the case, IR systems can be manual and the definition given here would include all manually searched library catalogues as well as bibliographies, indexes and abstracting publications. Nevertheless IR more commonly means retrieval from a computer system, whether the information is held on a local system increasingly CD-ROM form, or on a remote system accessed by a telecommunications network. The queries put to IR systems are one of two types : the search is either for a known item or for items on a particular subject.

In responding to queries, IR systems must achieve a balance between speed, accuracy, cost and retrieval effectiveness in revealing and the existence of information items and displaying surrogates (representation) or the original items. Whilst there are increasing number of IR systems which are full text, the principles and practices of IR developed with bibliographic systems in which the full text is represented by a surrogate. At the heart of a bibliographic IR system is a database of document representations.

## 15.2 A Classification of Retrieval Techniques

A retrieval technique is a technique for comparing the query with the

document representations. We can classify retrieval techniques in terms of the characteristics of the retrieved set of documents and the representations that are used Some techniques do not fall naturally into a single category in this classification, and others are hybrids of techniques from different categories, but the scheme is useful for discussing the broad distinctions among retrieval techniques.




A classification of retrieval techniques

In the above figure the first distinction that we make among retrieval techniques is whether the set of retrieved documents contains only documents whose representations are an exact match with the query or a partial match with the query.

The next level of the classification distinguishes between retrieval techniques that compare the query with individual document representation and techniques that use a representation of documents that emphasizes connections to other documents in a network. In this category individual documents are retrieved, but the retrieval is based on connections to other documents and not on the contents of an individual document. In the network category, we identify the subcategories ol cluster-based searches, searches based on browsing a network of documents, anc spreading activation searches.

The individual category breaks down into retrieval techniques that use a feature-based representation of queries and documents and techniques that use a structure-based representation. In a feature-based representation, queries and documents are represented as sets of features such as index terms. Features can be weighted and can represent more complex entities in the text than single words. The structure-based category is divided into representations based on logic, that is, those in which the meaning of queries and documents are represented using some formal logic, and on representations tht are similar to graphs, in which documents and queries are represented by graph-like structure composed of nodes and edges connecting those modes. Such graphs can be produced by natural language processing or statistical techniques.

The feature-based category includes techniques based on formal models (including the vector space model, probabilistic model, fuzzy set model and others) and techniques based on ad-hoc similarity measures.

## 15.3 Exact Match Techniques

Exact match retrieval techniques are those that require that the request model be contained, precisely as represented by in the query formulation, within text representation. Implemented as Boolean, full-text string searching, this is the current retrieval technique in current use in most of the large operational IR systems.

### 15.3.1 Disadvantages

In the simple case exact match searching

(a) misses many relevant texts whose representations match the query only partially; (b) does not rank retrieved texts; (c) cannot take into account the relative importance of concepts either within the query or within the text; (d) requires complicated query logic formulation; and (c) depends on the two representations being compared having been drawn from the same vocabulary.

Given its many and obvious objections, the exact match searching

remains the paradigm for operational systems. The traditional answers are that the investment in current systems is so great that changing them is economically unfeasible, that alternative techniques are untested in large-scale environments and that the results of alternative techniques are not sufficiently better even in experimental environment to justify any changes.

## 15.4 Partial Match Techniques

### 15.4.1 Individual, Feature-Based Techniques

Techniques in this category are used to compare queries with document represented as sets of features or index terms. The document representatives are derived from the text of the document either by manual or automated indexing Similarly, the query terms can either be derived from a query expresent in natura language, or an indexing vocabulary can be used directly for specifying queries.

Features can represent single words, phrases or concepts and can have weight: associated with them.

#### 15.4.1.1 Formal

These retrieval techniques are based on formal models of document retrieva and indexing. We concentrate here.on the major modeling approaches that have beer used for information retrieval : vector space, probabilistic and fuzzy set.

##### 15.4.1.1.1 Vector space model

In the vector space model, documents and queries are vectors in an dimensional space, where each dimension corresponds to an index term. The mode has intnitive appeal and has formed the basis of a large part of IR research, including the SMART system. The vector space model makes the following assumptions :

(a) the more similar a document vector is to a guery vector, the more

likely it is that the document is relevant to that query, (b) the words used to define ths dimensions of the space are independent.

#### 15.4.1.1.2 Probabilistic model

The retrieval techniques based on the probabilistic model are very similar to those developed from the vector space model. The basic aim is to retrieve document; in order of their probability of relevance to the query. 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.

#### 15.4.1.1.3 Fuzzy Set

In fuzzy set theory, a element has a varying degree of membership to a set instead of the traditional binary membership choice. The right of an index term for a given document reflects the degree to which this term describes the content of a document. Hence the right 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 to two fuzzy sets is equal to the maximum and minimum, respectively, of the degrees of membership of the elements of the two sets. A fuzzy set approach to information retrieval has been discussed in many papers. The main contribution of this work in terms of retrieval techniques has been an integration of Boolean queries with ranking techniques.

### 15.4.1.2 Adhoc

A number of similarity measures for comparing queries and documents have been proposed in the literature. Many of them were developed in the context of numerical taxonomy Similarity measures typicaly consist of a measure of the overlap of the query and document sets of terms normalized by the size of the sets involved.

#### 15.4.1.2.1 Individual, Structure Based

In this category of retrieval techniques, either the query of the documents or both are represented by more complicated structures than the sets of terms used in feature-based techniques.

##### 15.4.1.2.1.1 Logic

It is theoretically possible to represent the information conveyed by the text in documents as sentences in a formal logic. For example, the statement "DEC sells computers", could be considered in first-order predicate calculus us )Sells Dec computer)

##### 15.4.1.2.1.2 Graph

The general characteristic of a graph-like representation is a set of nodes and edges (or links) connecting these nodes. Specific examples include semantic nets and frames, which are produced by natural language processing. Simpler networks structures can be produced by statistical techniques.

### 15.4.2 Network

Network comprises Cluster, Browsing and Spreading activation, techniques.

#### 15.4.2.1 Cluster

A cluster is a group of documents whose contents are similar. A particular clustering method goves a more detailed definition of cluster and provides a technique for generating them. The use of clustering for information retrieval was a major topic in the SMART project. The approach used was to form a cluster hierarch using an adhoc clustering technique. The cluster

lierarchy was formed by dividing documents into a few large clusters, dividing these clusters into smaller clusters, and so on. A top-down search of the cluster hierarchy is performed by comparing (using a similarity measure) the query to cluster representatives of the top-level clusters, choosing the best clusters, comparing the query with the representatives of lower-level clusters within these clusters, and so on until a ranked list of lowest-lead clusters is produced. The documents in the top-ranked clusters are the ranked individually for presentation to the user.

#### 15.4.2.2 Browsing

If the documents, terms, and other bibliographic information are represented in the system as a network of nodes and connections, the user can browse through this network with system assistance. Browsing is an interesting retrieval technique in that it places less emphasis on query formulation than do other techniques. It relise heavily on the immediate feedback provided by user browsing decisions.

#### 15.4.2.3 Spreading activation

Spreading activation is a retrieval techniques that has some similarities to browsing. A query is used to "activate" parts of a network that diatribes the contents of documents, and how they are related to each other. In the simplest case, the query would activate index term nodes that are connected to document nodes and other terms. In more knowledge—intensive networks, the links and nodes represent concepts from the subject domain and how they relate to each other as will as the documents that contain those concepts. From the "start nodes" provided by the query, other nodes connected to those nodes are in turn "activated" (hence the term "spreading activation".

## 15.5 Feedback Methods

Relevant feedback techniques are not considered retrieval techniques by our criteria. Rather they are used to refine the request model, which is than used for another search. Feedback techniques are, however, an extremely

important part of ensuring that a document retrieval system will be effective.

These techniques were primarily developed in the context of feature-based retrieval although the principles apply to any retrieval technique. The main part of relevance feedback is the adjustement of weights associated with query terms. This adjustment is done on the basis of term occurrence in the documents identified by the user as relevant. The query (or request model) can also be changed by the addition of new terms from relevant documents. Some control is needed over the number of new terms added, and it seems that the most reliable method is to have users identify interesting terms in relevant documents.

Other types of modifications based on feedback are possible, such as adding term dependencies identified in relevant documents or modifying the "document space" (document indexing) to make relevant documents more similar to queries.

## 15.6 Current Status

In the past, most research in retrieval techniques are rather far removed from operational environments and to some extent from operational constraints. The past several years seem to show movement in this research in three directions; two respond to some extent to what may have been problems in acceptance of research results; the third seems to offer promise in greatly improving relevance performance.

First there has been a good deal of work on relating partial match techniques to exact match, as in extended Boolean searching and the use of Boolean-derived dependencies in probabilistic searching. This can be seen as a response to the demands of the operational environment.

The second response seems to be the realization that no one technique will be adequate for all purposes and that either a mix of techniques or a principled choice of techniques is required to improve IR system of performance.

The their new direction we note is inerasingly complex representations of the reuquest or user's problem. Although retrieval techniques are not the same as representations, the techniques one can use are determined by the

representation. The more complex the representation, one might think, the more kind; of retrieval techniques are passible.

## 15.7 Summary

This unit will make you aware of what a retrieval technique means. You will know the distinction between exact match and partial match. Why does exact match searching remain the paradigm for operational system? The unit gives a classification of retrieval techniques and describes each technique. This unit highlights feedback methods and current status of IR techniques

## 15.8 Exercise

1. What is a retrieval technique? What are the disadvantages of exact match technique?
2. Bring out the salicut features of the following techniques spreading activation, Feature—based and Fuzzy set.
3. Give a diagrammatic view of the classification of retrieval technique.
4. Describe briefly the exact match technique of information retrieval.
5. Why does exact match searching remain the paradigm of operational systems?

## 15.9 References and Further Reading


1. Lancaster, F. W. : Indexing and abstracting in theory and practice. Graduate School of Library and Information Science, 1991.
2. Lancaster, F. W. and Warmer, A : Information retrieval today. Information Resources Press, 1988.
3. Nicholas, J Belkin and Croft, W. Bruce : Retreval techniques. ARIST 24, 1991.
4. Sparck Jones, Karen and Willett, Peter Eds. Readings in information retrieval. Morgan Karufmann, 1977.

# Unit 16 ❐ Evaluation of IR Systems

Structure

## 16.0 Objectives

By an IR system we mean a system that retrieves documents or references to documents, as opposed to data. In this unit you will have an idea about :

1. traditional cranfieid model for IR evaluation
2. the problems with the model
3. the emerging themes in IR evaluation

## 16.1 Introduction

Evaluation takes many forms in many contexts. Peter Hernon and Charles R. McCiure define it as "the process of identifying and collecting data about specific services or activities, establishing criteria by which their success can be assessed, and determining both the quality of the service or activity and the degree to which the service or activity accomplishes stated goals and objectives." We adopt a similar definition. Here we concentrate on works that are primarily about IR systems from a basic research perspective.

Evaluation of information retrieval is a subject that has been of interest to many great minds in the field for more than 40 years. Nevertheless, there exists much controversy about evaluation measures and methodologies. In this unit, three major topics will be discussed : evaluation measures (for example, precision and recall) and retrieval performance; evaluation criteria for different types of IR systems; three IR evaluation projects : cranfieid tests. The STAIR evaluation and TREC series.

## 16.2 Evaluation meansures for IR

Information retrieval is both performance-oriented and process-oriented, especially end-users, are typically more interested in what has been retrieved than what has been done during the process. A better understanding of the retrieval process would, however, help improve retrieval performance. For the reason, evaluation measures developed for information retrieval can be approximately categorised into two kinds, one concentrating on performance and the other on process. Among the measures for evaluating retrieval performance, recall and precision, Salton suggests that recall and

precision are the two most well known and widely adopted criteria.

### 16.2.1 Recall and Precision

| Judgment / Result | Relevant | Not Relevant | Total |
|---|---|---|---|
| Retrieved | a (hits) | b (noise) | a + b (all retrieved) |
| Not retrieved | c (misses) | d (rejects) | c + d (all non retrieved) |
| Total | a + c (all relevant) | b + d (all non relevant) | a + b + c + d (total in the system) |

The above table presents all the possible outcomes for a retrieval task, using the notation and terminology seen in IR publications. For instance, "hits" mean retrieved documents that are relevant. "Misses" are defined as documents that are relevant but not retrieved. Based on these possible retrieval outcomes, a variety of evaluation measures are dirived. Recall and precision, initially proposed by Allen Kent and his colleagues in 1955 as "pertinency factor" and "recall factor", are two such measures.

As noted in earlier unit, the relationship between recall and precision tends to be inverse although Robert Figmann Challenged the statement with several examples. Fugmenrfound that an increase in precision is not always accompanied by a corresponding decrease in recall, and an increase in recall, and an increase in recall is by no means observed to have always in its wake a decrease in precision. A partial explanation to Fugmann's is that retrieval of relevant documents affects both recall and precision while retrieval of nonrelevant documents affects only precision.

### 16.2.2 Fallout

John A. Swets defined fallout as the ratio between non-relevant documents retrieved and all nonrelevant documents in a system database.

Using the notation in Table 14.2.1.

$$F = \frac{b}{b+d}$$

Fallout measures the inability of an 1R system in excluding nonrelevant document from retrieval resules, which Robertion called noise factor. The smaller the fallout value the better the IR system is from the evaluation viewpoint.

### 16.2.3 Generality

Generality is defined as the proportion of documents in a system database that is relevant to a particular topic According to the notation given in Table 14.2.1.

$$G = \frac{a+c}{a+b+c+d}$$

The higher the generality number or the greater the density of relevant items in the database, the easier the search tends to be. But generality is, strictly speaking, a measure for the database quality from the perspective of relevance rather than for retrieval performance directly.

## 16.3 Evaluation Criteria for IR Systems

Recall, precision and other measures are basically for evaluating retrieval performance. Retrieval performance apparently only represents one aspect of information retrieval even though the end-user pays more attention to it. In order to get a complete view of how well and IR system works, measures other than those for evaluating IR performance should be considered. Different measures should be used for different types of IR systems.

### 16.3.1 Online Systems

Evaluations of online systems can possibly be done at three levels : evaluation of effectiveness, evaluation of cost-effectiveness and cost-benefit evaluation. Although various criteria have been established for evaluating

online systems, a consensus seems to have been reached about what aspects should be inspected. Lancaster and Fayen recapitulated them in the form of six criteria : coverage, recall, precision, response time, user effort, and form of output.

### 16.3.2 CD-ROM Systems

Although there exists no common list of evaluation criteria for CD-ROM systems, similarities among the different sets of measures for evaluating CD-ROM systems can be easily discerned with respect to hardware and software considerations, database, searching facilities, user friendliness, and output features. Recall and precision are not explicitly listed as evaluation measures for CD-ROM retrieval. One possible explanations could be that some CD-ROM systems are built for retrieving known items rather than for locating information on a subject. Within that context, the calculation of recall and precision becomes meaningless.

### 16.3.3 Internet

Relatively a few evaluation projects have addressed Internet retrieval systems, The composition of indexes in an 1R system is an important component for evaluation. Search capability consitutes another criterion for evaluating Internet retrieval systems. Output, as an evaluation criterion for Internet retrieval, systems. should be examined from three perspectives : accessibility, content, and format.

## 16.4 Major Evaluation Projects for IR

The evaluation study conducted by Documentation Inc. in 1953 was identified by Cleverdon as the first test of any significance. This was essentially a comparison between uniterm system and an alphabetical subject catalogue prepared by the Armed Services Technical Information Agency (ASTIA) of USA 15000 documents were indexed by both the methods and 98 queries submitted by ASTIA users were utilised to measure the performance of each system. However, results of this study were incoclusive. In 1954 Cleverdon

and Throne conducted a small study on Uniterm method, which though inconclusive had an impact in later days on Cranfield tests. Leaving apart these stray beginnings, definite trends are observed for each of the succeeding decades since the year 1958.

Numerous projects have been carried out to evaluate information retrieval since that time. The Cranfield tests (for example, Cleverdon, 1962), the MEDLARS (Medical Literature Analysis and Retrieval System) project (for example, Lancaster, 1968), the SMART experiment (for example, Salton, 1981), the STAIRS (S Torage And Information Retrieval System) study, (for example, Blair & Moron, 1985) and the TREC series (for example, Harman, 1993; Voorhees & Harman, 2000) are some of major projects is this category.

### 16.4.1 The Cranfield Tests

**16.4.1.1** The Cranfield tests were carried out in two phases, named Cranfield I and Cranfield II respectively, between 1957 and 1967. A grant was received by ASLB in 1957 from the National Science Foundation to investigate the comparative performance of different indexing systems. The test was carried out at the College of Aeronautics, Cranfield under the supervision of Cyril Cleverdon to study the following four indexing systems :

(i) An Alphabetical Subject Catalogue

(ii) A faceted classification scheme

(iii) A Uniterm system of coordinate indexing

(iv) Universal Decimal Classification.

The project involved indexing of 18000 documents in the filed of aeronantics by each of the four systems and subsequent search in response to queries formulated for this purpose. The searching was done by the project staff. To carry out the search 1400 questions were initially framed by users from different organisations, and these were serutinised by a panel of three experts to select only 400 questionsfor this test. The questions were then put to four index files.

#### 16.4.1.2 Test Findings

The most important finding in Cornfield I is that the Uniterm system

gave the best recall among the four different indexing systems.

| Indexing system | Recall % |
|---|---|
| Uniterm | 82.0 |
| Alphabetical subject | 81.5 |
| UDC | 75.6 |
| Faceted classification | 73.8 |

### 16.4.2 Cranfield II

Cranfield II was designed to assess the effects of different indexing devices on retrieval performance. Indexing devices are measures employed by the indexer to improve either recall or precision. Examples of indexing devices include synonyms, generic relations, coordination, links, and roles.

. The test collection of 1400 documents is mainly in the field of high-speed aerodynamics and aircraft structures. Each document was "indexed" in three ways. First, the most important concepts were selected and recorded in natural language of the document. Second, the single words in each of the chosen concepts were listed. Third, the concepts were combined in different ways to form the main themes of the documents. Each term was given weight that is, 1, 2 or 3 with I being the most important at the time of indexing to indicate its relative importance Various indexing devices were applied as well during the indexing process in order to probe their effects on retrieval performance.

Searching was done by making use of the various indexing devices implemented in three major types of indexing languages : single term, simple concept and controlled term index languages. To conduct searches 221 questions were used. There were generated by the suthors of research papers. Relevancy of document was ascertained in respect of questions. Relevancy thus obtained was graded from 1 to 4 in the following manner :

(a) Complete answer to the question
(b) High degree of relevance
(c) Useful, providing general background of the work or dealing with specific area

(d) Minimum interest, providing information like historical viewpoint.

For the assessment of a single performance measure, called normalized recall, was introduced. This is a ratio of cumulated recall ratio and number of search stages involving document output cut-off groups.

### 16.4.1.2 Findings of Cranfield II

Other than corroborating inverse relationship of recall and precision, the results of the test was somewhat surprising. The test reported the following findings :

Single term indexing languags were superior to any other type in terms of performance.

When single terms were used for indexing, the inclusion of collateral classes and in particular, quasi-synonyms worsened the performance.

When concepts were used for indexing, the performance worsened with the inclusion of superordinate, subordinate, and collateral classes along with original concepts.

When controlled terms were used for indexing, the inclusion of narrower and broader terms worsened the performance. Indexing languages formed out of titles performed better than those formed out of abstracts.

#### 16.4.1.2.1 Criticisms of Cranfield II

Problems with the Cranfield model have been classified into four major categories of issues : validity and reliability, generalizabiiity, usefulness and conceptual.

Validity and reliability issues : The omission of the user from the traditional IR model, whether it is made explicit or not, stems directly from the user's absence from the craufield instrument. The model assumes that the user, hairing recognised an information need, comes to an information system for help, with a query or request derived from the information need. The query is the input into the IR system, which takes the request and matches it against document. Susogates in the database with a goal of presenting to the user the texts most likely to satisfy the information need.

In deciding how to evaluate an IR system, it seems crucial to understand how the users of the system themselves evaluate performance. It is possible that different kinds of operational systems and different kinds of users require different evaluation criteria.

Because all relevant documents must be known in advance, recall is very difficult and expensive to calculate, even in very small experimental collections.

Generalizabiiity issues : Cornfield experiments employ a variety of human subjects, information needs, questions, documents and relevance judgments. Yet random sampling from these populations has rarely been done; instead, smaples have typically been based on convenience. Because of this known statistical techniques are not applicable to many retrieval experiments, and in the few cases where they are, statistical significance has only rarely been achieved. Thus the reproducibility and experiments is called into question, as will as external validity and generalizability of findings.

Usefulness issues : The experimental approach to IR typified by the Grainfield steadies cannot be used in real, operational systems with real users.

Conceptual issues : Relevance is only one of the IR Concepts that are both ubiquitous and slippery—concepts that have thus for eluded definition, either operationally or conceptually. Nicholas J. Belkin discussed other "ineffable concepts'[1] which include information need, desire, information, aboutness, meaning, satisfaction, effectiveness and synthema. Befking shows that these concepts are central to IR and as such must be more fully understood in IR own right in order to provide a sounder basis for IR design and evaluation. Vickery agreed that dinfings of cranfield II are valuable explanation of the retrieval process. But in the same breath he said "They give no final answers, and their concussions must be treated with caution."

### 16.4.2 The STAIRS Evaluation

The STAIRS project was a rare large-scale evaluation of an operational IR system. It was conducted by IBM's STAIRS system with a huge database of legal materials. The STAIRS project was unusual in several respects. Although it was based on a cranfields design, the investigators made great efforts to ensure a higher defree of validity than in previous studies. It also

addressed several objections that had been raised about previous cranfields instruments : small size, lack of inclusion of full texts, and failure to find significant numbers of relevant documents which to base recall estimates. The stairs project stands as the first and most comprehensive evaluation of full-text system. Its database was comparatively large (about 350,000 pages of online instruments). It remains one of the few large-scale studies of a commercial operational system. It also employed a unique method to make accurate assessments of recall (or more accurately, maximum recall), by sampling subsets of documents likely to be rich in relevant documents. Finally, David C. Blair noted in 1996 that it is one of the few published IR evaluations that was undertaken in an environment innlving political, legal and ethical issues that were contended with and discussed fully by the researchers.

### 16.4.3 The TREC Series

The Text Retrieval Conference (TREC) is a major research initiative in IR evaluation, coordinated by the National Institute of Standards and Technology (NIST) and sponsored by the Advanced Research Projects Agency (ARPA) of the United States Department of Defense. The TREC series constitutes a succession of sonferences that have been held in the United States annually since 1992 (that is, TREC-I) with the intention to meet the following goals :

(a) To encourage reasearch in text retrievaJ based on large text data collections;

(b) To increase communication among industry, academic, and government by creating an open forum for the exchange of research ideas;

(c) To speed the transfer of technology from research laboratories into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems;

(d) To increase the availability of appropriate evaluation techniques for useby industry and academic, including developments of new evaluation techniques more applicable to current systems.

The invited participants in TREC test their IR system on large heterogeneous test collections of full texts. Twenty-five research groups took part in TREC-I. Thirty six groups (14 companies and 22 universities participated in TREC-4. Although TREC differs in some respects from earlier IR experimentation, it is at heart, a Cornfield design.

TREC is central on two tasks routing and ad hoc. Routing (or filtering, profiling) tasks assume that the same questions are asked but the new documents are continually being searched. In the routing tasks, topies and relevant documents are known and new data are used for testing. Ad hoc tasks consist of new (library, reference type) questions continually being posed to a static set of data.

TREC is a major, significant contribution to IR evaluation in several ways. It is resulted in a collaborative environment for IR testing and evaluation that is unprecedented. The number of participating research teams is also large, and because a common approach and format to performance evaluation is used, in addition to questionnaires and workshop that follow each phase, there is rapid dissemination of findings and continued evaluation of the TREC approach.

#### 16.4.3.1 Findings of TREC

The findings in TREC are too many and divesified. Therefore, only the general ones are noted below :

Many different approaches give similar performance, indicating that "All roads lead to Rome" and different methods can be sought to achieve the same objective.

Automatic processing (that is, query construction or retrieval) is as good as manual processing but may not be quite so convincing for poor queries.

Statistically based indexing and searching techniques are cheap and competitive, particularly for large collections, or datalsases.

Simple phrases along with single terms contribute something, implying additional promise for automatic methods in the field.

Relevance feedback is valuable, and moderate guery expansion seems helpful. A system still cannot automatically analyses a request for the

implications of its term makeup and choose one particular query formation strategy that is likely to give best results.

#### 16.4.3.2 Criticism to TREC

The main criticism to TREC is that it basically follows the footsteps of the cranfields tests by using an artificial test environment, and choosing recall and precision as retrieval performance measures :

The artificiality of the TREC series is threefold. First, the test documents of mainly newspapers, are gathered for the project rather than a collection naturally built over the years. Second, TREC does not have genuine and hands-on users. Third, the tests are done in the traditional laboratory environment with certain degree of control and manipulation. It is hard to relate such an environment to the real, operational system.

#### 16.4.3.3 Significance of TREC

With many of its unique features (for example, test document size and intrcasing number of participating teams) TREC holds a significant place in the history of IR evaluation. Its significance has been reflected in the following aspects.

It provides an open forum for evaluating informations retrieval. By being an open forum, it has encouraged participation by many institutions and organisations from home and abroad.

Many different approaches have been tried in TREC, which include passage retrieval, data fusion, statistical and probabilistic indexing and term weighting strategies.

TREC is a continuing research effort. It attempts to answer many questions, which can be put into such broad categories as indexing and retrieval modeals, indexing vocabulary document descriptions, indexing sources, queries and query sources, search strategy, scoring criteria and learning ability of IR systems.

## 16.5 Conclusions

Proposals and suggestions have been made to advance research in evaluating information retrieval by going beyond the existing models and

implementing methodology pluralism. Future evaluation projects should be done in a real-life environment and use measures other than recall, precision and fallout to overcome the two most notorious limitations of previous assessments in the field, namely, subjectivity in judging relevance and inability to determine the total number of relevant documents in a system.

## 16.6 Summary

This unit introduces the evaluation of information retrieval. It explains the different measures of evaluation. It sums up the evaluation criteria for different types of IR systems. It describes numerous projects viz. Cranfields tests, STAIRS, and TREC series alongwith problems and significances and test findings. It has also suggested evaluation of IR by going beyond existing models and methodology pluralism.

## 6.7 Exercise

1. Describe the widely adopted criteria for evaluating retrieval performance.
2. Elucidate the relationship between recall and precision alson with the Fugmann's argument in this regard.
3. Discus the significance of the following projects (a) Cranfields II, STAIRS and TREC Series.

## 16.8 References and Further Reading


1. Balkin, Nicholas J : Ineffable concepts in Infotmation Retrieval. In Sparck Jones, Karen ed. Information Retrieval Experiment. London, Bulterworths, 1981.
2. Blair, David C : STAIRS Redux : thoughts on the STAIRS Evaluation, ten years after. **JASIS** 1996, 47, 4—22.
3. Chu, Hating : Information representation and retrieval in the digital age. Thomas J. Hogan Sr. 2003.
4. Vickery, B. C. : Reviews on the Cranfields 2. report. J. **Doc** 1967, 23(4), 338-40.
5. Williams, Martha E Ed : ARIST vol 32, 1997, 3-15.

মানুষের জ্ঞান ও ভাবকে বইয়ের মধ্যে সঞ্চিত করিবার যে একটা প্রচুর সুবিধা আছে, সে কথা কেহই অস্বীকার করিতে পারে না। কিন্তু সেই সুবিধার দ্বারা মনের স্বাভাবিক শক্তিকে একেবারে আচ্ছন্ন করিয়া ফেলিলে বুদ্ধিকে বাবু করিয়া তোলা হয়।

*— রবীন্দ্রনাথ ঠাকুর*

ভারতের একটা mission আছে, একটা গৌরবময় ভবিষ্যৎ আছে, সেই ভবিষ্যৎ ভারতের উত্তরাধিকারী আমরাই। নূতন ভারতের মুক্তির ইতিহাস আমরাই রচনা করছি এবং করব। এই বিশ্বাস আছে বলেই আমরা সব দুঃখ কষ্ট সহ্য করতে পারি, অন্ধকারময় বর্তমানকে অগ্রাহ্য করতে পারি, বাস্তবের নিষ্ঠুর সত্যগুলি আদর্শের কঠিন আঘাতে ধূলিসাৎ করতে পারি।

*— সুভাষচন্দ্র বসু*

Any system of education which ignores Indian conditions, requirements, history and sociology is too unscientific to commend itself to any rational support.

*—Subhas Chandra Bose*

**Price : ₹ 150.00**
(Not for sale to the Students of NSOU)

Published by : Netaji Subhas Open University, DD-26, Sector-I, Salt Lake, Kolkata-700 064 &
Printed at : The Saraswati Printing Works, 2, Guru Prosad Chowdhury Lane, Kolkata 700 006