

# Information Storage and Retrieval

DLIS405



**L**OVELY  
**P**ROFESSIONAL  
**U**NIVERSITY



# **INFORMATION STORAGE AND RETRIEVAL**

Copyright © 2012 Wasim Ul Haque  
All rights reserved

Produced & Printed by  
**LAXMI PUBLICATIONS (P) LTD.**  
113, Golden House, Daryaganj,  
New Delhi-110002  
for  
Lovely Professional University  
Phagwara

# SYLLABUS

## Information Storage and Retrieval

### Objectives:

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

| Sr. No. | Content  |
|---------|--|
| 1       | <b>Classification:</b> Development and trends in general classification Schemes, CC, DDC, UDC, LC, BC.   |
| 2       | Activities of organization in Classification research. DRTC, ISKO, CRG   |
| 3       | <b>Cataloguing:</b> Development and trends in Cataloguing, ISBD, MARC, CCF, OPAC.  |
| 4       | Subject Cataloguing and concept indexing for online research.  |
| 5       | <b>Indexing:</b> Developments and trends in indexing. Derived indexing; Assigned indexing, Alphabetical indexing, Keyword indexing. Pre and post coordinate indexing systems, citation indexing. |
| 6       | Features of Information storage and retrieval system – precision, recall relevance.  |
| 7       | Online searching and search strategies.  |
| 8       | <b>Vocabulary control:</b> Methodology current trends and development  |
| 9       | Sears list of subject Headings, Library of congress subject Headings. Medical Subject Headings (MeSH).   |
| 10      | Thesaurus of ERIC descriptors, Thesauro – facet.   |

## CONTENTS

|                 |  |     |
|-----------------|--|-----|
| <b>Unit 1:</b>  | Introduction to Library Science          | 1   |
| <b>Unit 2:</b>  | Library Classification                   | 13  |
| <b>Unit 3:</b>  | Organization in Classification Research  | 30  |
| <b>Unit 4:</b>  | Cataloguing-Development and Trends       | 48  |
| <b>Unit 5:</b>  | MAchine-Readable Cataloguing and Online  | 52  |
| <b>Unit 6:</b>  | Cataloguing                              | 58  |
| <b>Unit 7:</b>  | Sorting and Indexing                     | 64  |
| <b>Unit 8:</b>  | Indexing                                 | 73  |
| <b>Unit 9:</b>  | Trends in Indexing                       | 84  |
| <b>Unit 10:</b> | Information Storage and Retrieval System | 96  |
| <b>Unit 11:</b> | Online Searching: Library Databases      | 107 |
| <b>Unit 12:</b> | Vocabulary Control                       | 111 |
| <b>Unit 13:</b> | Subject Headings                         | 120 |
| <b>Unit 14:</b> | ERIC and Thesaurofacet                   | 163 |

## Unit 1: Introduction to Library Science

### CONTENTS

Objectives

Introduction

- 1.1 Development of Library Science
  - 1.1.1 Ancient Information Retrieval
  - 1.1.2 Education and Training
- 1.2 Librarians in Different Types of Libraries
- 1.3 Development of Library Science Literature
- 1.4 Geographic Distribution of Library and Information Science Literature
- 1.5 Summary
- 1.6 Keywords
- 1.7 Review Questions
- 1.8 Further Readings

### Objectives

After studying this unit, you will be able to:

- Define development of library science
- Describe librarians in different types of libraries
- Explain development of library science literature
- Describe geographic distribution of library and information science literature.

### Introduction

Library science (or Library and Information science) is an interdisciplinary field that applies the practices, perspectives, and tools of management, information technology, education, and other areas to libraries; the collection, organization, preservation, and dissemination of information resources; and the political economy of information. The first school for library science was founded by Melvil Dewey at Columbia University in 1887.

Historically, library science has also included archival science. This includes how information resources are organized to serve the needs of select user groups, how people interact with classification systems and technology, how information is acquired, evaluated and applied by people

**Notes**

in and outside of libraries as well as cross-culturally, how people are trained and educated for careers in libraries, the ethics that guide library service and organization, the legal status of libraries and information resources, and the applied science of computer technology used in documentation and records management.

The term library and information science (LIS) is most often used; most librarians consider it as only a terminological variation, intended to emphasize the scientific and technical foundations of the subject and its relationship with information science. LIS should not be confused with information theory, the mathematical study of the concept of information.

LIS can also be seen as an integration of the two fields' library science and information science, which were separate at one point. Library philosophy has been contrasted with library science as the study of the aims and justifications of librarianship as opposed to the development and refinement of techniques.

### **1.1 Development of Library Science**

The history of the library, it may be argued, began with the first effort to organize a collection of information and provide access to that information.

At Ugarit in Syria excavations have revealed a palace library, temple library, and two private libraries which date back to around 1200 BC, containing diplomatic texts as well as poetry and other literary forms. In the 7th century, King Ashurbanipal of Assyria assembled what is considered the first systematically collected library at Nineveh; previous collections functioned more as passive archives. The legendary Library of Alexandria is perhaps the best known example of an early library, flourishing in the 3rd century BC and possibly inspired by Demetrius Phalereus.



*Task*

Write a note on development of Library Science.

#### **1.1.1 Ancient Information Retrieval**

One of the curators of the imperial library in the Han Dynasty is believed to have been the first to establish a library classification system and the first book notation system. At this time the library catalog was written on scrolls of fine silk and stored in silk bags.

#### **19th Century**

Thomas Jefferson, whose library at Monticello consisted of thousands of books, devised a classification system inspired by the Baconian method, which grouped books more or less by subject rather than alphabetically, as it was previously done. Jefferson's collection became the nucleus of the first national collection of the United States when it was transferred to Congress after a fire destroyed the Congressional Library during the War of 1812. The Jefferson collection was the start of what we now know as the Library of Congress.



*Did u know?*

The first textbook on library science was published in 1808 by Martin Schrettinger, followed by books of Johann George Seizinger and others.

## 20th Century

## Notes

In the English speaking world the term “library science” seems to have been used for the first time in a book in 1916 in the “Panjab Library Primer” written by Asa Don Dickinson and published by the University of the Punjab, Lahore, Pakistan. This university was the first in Asia to begin teaching ‘library science’. The “Panjab Library Primer” was the first textbook on library science published in English anywhere in the world.



*Notes* The first textbook in the United States was the “Manual of Library Economy” which was published in 1929.

Later, the term was used in the title of S. R. Ranganathan’s *The Five Laws of Library Science*, published in 1931, and in the title of Lee Pierce Butler’s 1933 book, *An introduction to library science* (University of Chicago Press).

Shiyali Ramamrita Ranganathan was a mathematician and librarian from India. His most notable contributions to the field were his five laws of library science and the development of the first major analytico-synthetic classification system, the colon classification. He is considered to be the father of library science, documentation, and information science in India and is widely known throughout the rest of the world for his fundamental thinking in the field.

In more recent years, with the growth of digital technology, the field has been greatly influenced by information science concepts. Although a basic understanding is critical to both library research and practical work, the area of information science has remained largely distinct both in training and in research interests.

### 1.1.2 Education and Training

Academic courses in library science include collection management, information systems and technology, research methods, cataloging and classification, preservation, reference, statistics and management. Library science is constantly evolving, incorporating new topics like database management, information architecture and knowledge management, among others.

Most professional library jobs require a professional post-baccalaureate degree in library science, or one of its equivalent terms, library and information science as a basic credential. In the United States and Canada the certification usually comes from a master’s degree granted by an ALA-accredited institution, so even non-scholarly librarians have an originally academic background. In the United Kingdom, however, there have been moves to broaden the entry requirements to professional library posts, such that qualifications in, or experience of, a number of other disciplines have become more acceptable. In Australia, a number of institutions offer degrees accepted by the ALIA (Australian Library and Information Association).

## 1.2 Librarians in Different Types of Libraries

### Public

The study of librarianship for public libraries covers issues such as cataloging; collection development for a diverse community; information literacy; readers’ advisory; community standards; public services-focused librarianship; serving a diverse community of adults, children, and teens; intellectual freedom; censorship; and legal and budgeting issues. The public library as a commons or public sphere based on the work of Jürgen Habermas has become a central metaphor in the 21st century.



**Notes**

**School**

The study of school librarianship covers library services for children in schools through secondary school. In some regions, the local government may have stricter standards for the education and certification of school librarians (who are often considered a special case of teacher), than for other librarians, and the educational program will include those local criteria. School librarianship may also include issues of intellectual freedom, pedagogy, and how to build a cooperative curriculum with the teaching staff.

**Academic**

The study of academic librarianship covers library services for colleges and universities. Issues of special importance to the field may include copyright; technology, digital libraries, and digital repositories; academic freedom; open access to scholarly works; as well as specialized knowledge of subject areas important to the institution and the relevant reference works.

Some academic librarians are considered faculty, and hold similar academic ranks as professors, while others are not. In either case, the minimal qualification is a Master's degree in Library Studies or Library Science, and, in some cases, a Master's degree in another field.

**Archives**

The study of archives includes the training of archivists, librarians specially trained to maintain and build archives of records intended for historical preservation. Special issues include physical preservation of materials and mass deacidification; specialist catalogs; solo work; access; and appraisal. Many archivists are also trained historians specializing in the period covered by the archive.

**Special**

Special librarians include almost any other form of librarianship, including those who serve in medical libraries (and hospitals or medical schools), corporations, news agencies, government organizations, or other special collections. The issues at these libraries will be specific to the industries they inhabit, but may include solo work; corporate financing; specialized collection development; and extensive self-promotion to potential patrons.

**Preservation**

Preservation librarians most often work in academic libraries. Their focus is on the management of preservation activities that seek to maintain access to content within books, manuscripts, archival materials, and other library resources. Examples of activities managed by preservation librarians includes binding, conservation, digital and analog reformatting, digital preservation, and environmental monitoring.

**Theory and Practice**

Many practicing librarians do not contribute to LIS scholarship, but focus on daily operations within their own libraries or library systems. Other practicing librarians, particularly in academic libraries, do perform original scholarly LIS research and contribute to the academic end of the field.

On this basis, it has sometimes been proposed that LIS is distinct from librarianship, in a way analogous to the difference between medicine and doctoring. In this view, librarianship, the application of library science, would comprise the practical services rendered by librarians in their day-to-day attempts to meet the needs of library patrons.

Whether or not individual professional librarians contribute to scholarly research and publication, many are involved with and contribute to the advancement of the profession and of library science and information science through local, state, regional, national and international library or information organizations.

### **1.3 Development of Library Science Literature**

Development of library and information science literature is mapped on the literary outputs available through Library, Information Science and Technology Abstracts (LISTA). The findings vividly indicate that the growth of literature in library and information science is on increase. Professionals and researchers all over the world have embraced the scholarly publication revolution from various disciplines, to which library and information science is no exception.

The study provides a detailed description of library and information science literature published in various formats. In country-wise output, U.S.A. ranks first with 301 (37.76%) publications. It also gives a thorough insight of the growth and development of library literature in a chronological order. It is surprising to know that the decade of 1980s ranks first with 155 (19.44%) publications. Furthermore, to specify the growth and development of library and information science literature, coverage policy and source type are also traced out. Core publications rank first with a literary output of 485. Among the literary wealth, academic journals hold the first place harvesting 409 publications and forming 51.31% of the total publications.

A number of studies have been conducted on the growth of library and information science literature. A study conducted by Bottle and Efthimiadis (1984) investigated the sampling issues of LISA (Library and Information Science Abstract), ISA (Information Science Abstract), RZI (Referativnyi Zhurnal Informatics Abstract), BS (Bulletin Signaletique) and CCA (Computer & Control Abstracts) for the year 1983. Journals (71%) are the dominant format. The study clearly indicates that literature coverage in the field of library and information science has increased dramatically. 1391 distinct journal titles were identified in the coverage of LISA, ISA, RZI and BS.

Most of the literature originated from North America (38%) and Western Europe (34%). 1545 journals were identified from Ulrich's Guides. Ali (1985) provides an overall picture of growth of librarianship and information science literature (academic research and practitioner's literature) and various outlets available for reporting research findings, with special emphasis on United States and Great Britain.

During the last ninety-two years, many journals have been published in Asia, but many have ceased publication for various reasons. At present, over two hundred journals in library and information science are published in Asian countries.

All these studies make it clear that the growth is multidimensional. The present study is to gauge once again the growth of LIS literature in the revolutionary era of ICT (Information Communication Technology) when the publication channels have been tremendously increased and the sharing of ideas and findings have become much more easier and affordable.

### **Objectives**

The study was initiated with the following objectives:

- To trace the intensification of literature chronologically. The study attempts to delve deep to discover the growth of library literature published worldwide as per LISTA indexing and abstracting service in different communication channels.
- To map the geographic distribution of the literature. The study makes an effort to identify the various world regions from where library and information science literature is being published.
- To outline the coverage policy and source type of the published literature. Since the library and information science literature is being published in different formats, the present study will identify them and also show the amount of literature being published through all these sources.

Notes

**Methodology**

LISTA (Library, Information Science and Technology Abstracts), a world class bibliographic indexing and abstracting database that provides coverage on subjects such as librarianship, classification, cataloguing, bibliometrics, online information retrieval, information management and more, was selected for analyzing library and information science literature. LISTA indexes nearly 700 periodicals plus books, research reports and proceedings, which makes it an ideal source for study. The data was carefully analyzed and interpreted. The geographic output and the date of the first issue was extracted from Ulrichsweb (Online version of Ulrich’s Periodical Directory published by R.R. Bowker), one of the renowned source listing periodical publications all over the world.

**Chronological Development**

Library and information science literature grew right from the time when the first core trade publication “Bookseller” (abstracted now in LISTA) appeared in 1852. The first publication date of all the titles under study is taken from Ulrich’s online Periodical Database. From 1852 to date, there is a tremendous increase in library literature in various forms. For the sake of convenience and clarity, the study is divided into decades from 1850’s to 2000’s. In 1850’s, only one publication fell into the scope of library and information science literature.



*Notes* During 1850’s and 1860’s, the literary output phase was dormant and produced no publications.

During 1870’s, 4 (0.50%) publications came out; in 1890’s, the number rose up to 5 (0.62%); and in 1900’s, it went up to 7 (0.87%) publications. The first magazine “Author” appeared in 1890’s. The first academic journal “New Library World” was also launched in the same decade. But with the passage of time, more and more publications emerged from the field. Table 1.1 demonstrates that the 1980’s has the largest number of publications, *i.e.*, 155 (19.44%). 1970’s and 1960’s rank 2nd and 3rd with a total of 124 (15.55%) and 68 (8.53%) publications respectively. Table 1.1 is supplemented by Fig. 1.1 to better demonstrate the data. The first year of publication for 163 publications could not be ascertained from Ulrich’s online periodical directory. Therefore, they were kept under the heading “Not Traceable”.

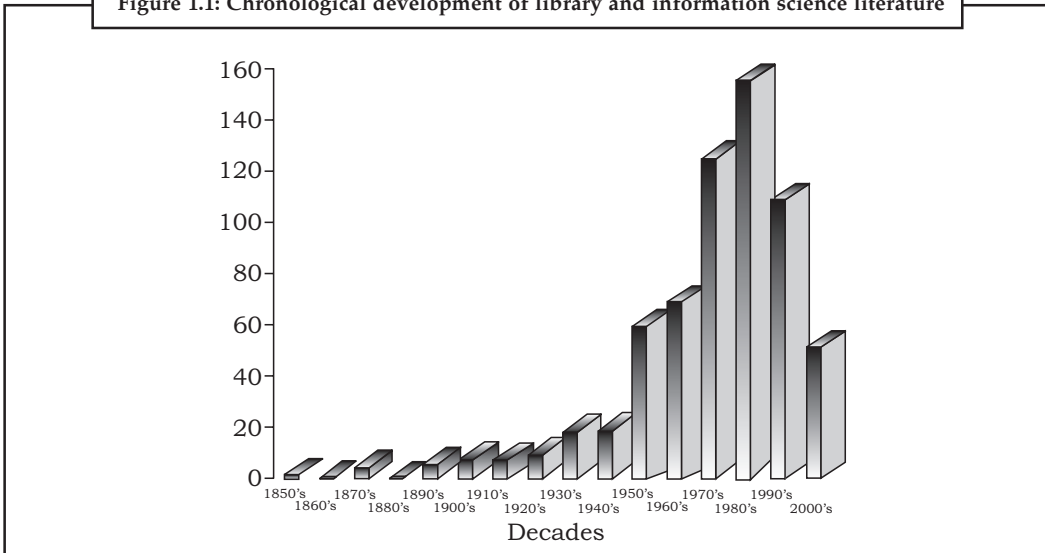
**Table 1.1**

| <i>Decades</i> | <i>Output</i> | <i>Percent</i> |
|----------------|---------------|----------------|
| 1850’s         | 1             | 0.12           |
| 1860’s         | 0             | 0.00           |
| 1870’s         | 4             | 0.50           |
| 1880’s         | 0             | 0.00           |
| 1890’s         | 5             | 0.62           |
| 1900’s         | 7             | 0.87           |
| 1910’s         | 7             | 0.87           |
| 1920’s         | 9             | 1.12           |

Notes

|               |     |       |
|---------------|-----|-------|
| 1930's        | 18  | 2.25  |
| 1940's        | 18  | 2.25  |
| 1950's        | 59  | 7.40  |
| 1960's        | 68  | 8.53  |
| 1970's        | 124 | 15.55 |
| 1980's        | 155 | 19.44 |
| 1990's        | 108 | 13.55 |
| 2000's        | 51  | 6.39  |
| Not Traceable | 163 |       |

Figure 1.1: Chronological development of library and information science literature



### Growth of Literature over Consecutive Decades

It is evident from Table 1.2 (supplemented by Fig. 1.2) that 1870's ranks first as there is an enormous increase of 400% of literature over its previous decade due to the fact that literature in the field was just starting to boom. The 2nd and 3rd ranks are occupied by 1890's (100%) and 1950's (85.50%) of literature over their previous decade.

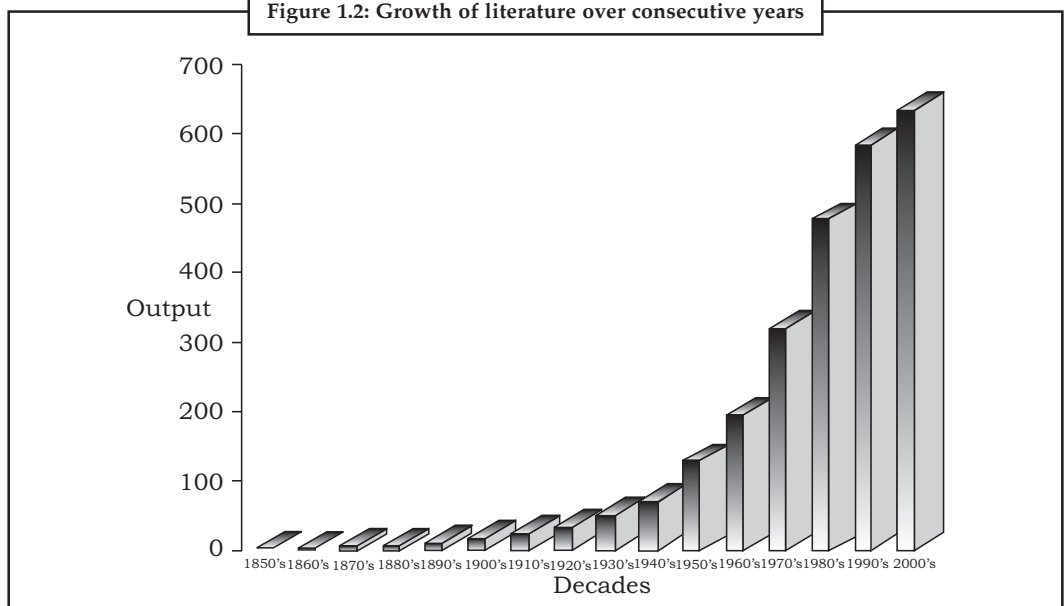
Table 1.2: Growth of literature over consecutive decades

| Decades | Output | Increase | Percent |
|---------|--------|----------|---------|
| 1850's  | 1      | -        | -       |
| 1860's  | 0      | 1        | 0.00    |
| 1870's  | 4      | 5        | 400.00  |

Notes

|               |     |     |        |
|---------------|-----|-----|--------|
| 1880's        | 0   | 5   | 0.00   |
| 1890's        | 5   | 10  | 100.00 |
| 1900's        | 7   | 17  | 70.00  |
| 1910's        | 7   | 24  | 41.17  |
| 1920's        | 9   | 33  | 37.50  |
| 1930's        | 18  | 51  | 54.54  |
| 1940's        | 18  | 69  | 35.29  |
| 1950's        | 59  | 128 | 85.50  |
| 1960's        | 68  | 196 | 53.12  |
| 1970's        | 124 | 320 | 63.26  |
| 1980's        | 155 | 475 | 48.43  |
| 1990's        | 108 | 583 | 22.73  |
| 2000's        | 51  | 634 | 8.74   |
| Not Traceable | 163 |     |        |

Figure 1.2: Growth of literature over consecutive years



Self Assessment

Multiple Choice Questions:

- Core publications rank first with a literary output of:
  - 484
  - 485
  - 486
  - 495
- ..... distinct journal titles were identified in the coverage of LISA, ISA, RZI and BS.
  - 1380
  - 1381
  - 1390
  - 1391

3. .... journals were identified from Ulrich's guides. Notes
- (a) 1525 (b) 1535
- (c) 1545 (d) 1555
4. In ..... only one publication fell into the scope of library and information science literature.
- (a) 1850's (b) 1870's
- (c) 1950's (d) 1750's

## 1.4 Geographic Distribution of Library and Information Science Literature

The regional distribution of publication channels in library and information science vividly demonstrate that the developed countries provide more publication channels. From Table 1.3, it is obvious that North America ranks first with 325 publications and accounts for 40.77% of the total. North America is followed by Europe and South America with 252 (31.61%) and 5 (0.62%) publications respectively.

The 4th, 5th, and 6th positions are attained by Asia (36, 4.51%), Australia (10, 1.25%), and Africa (7, 0.87%). Within Asia, India took the lead with 12 publications, which accounts for 1.50% of the total. Japan and Taiwan rank 2nd (10, 1.25%) and 3rd (6, 0.75) in Asia. However, Table 1.3 (supplemented by Fig. 1.3) clearly implies that the publication channels for literary wealth in library and information science are budding from other developing nations as well.

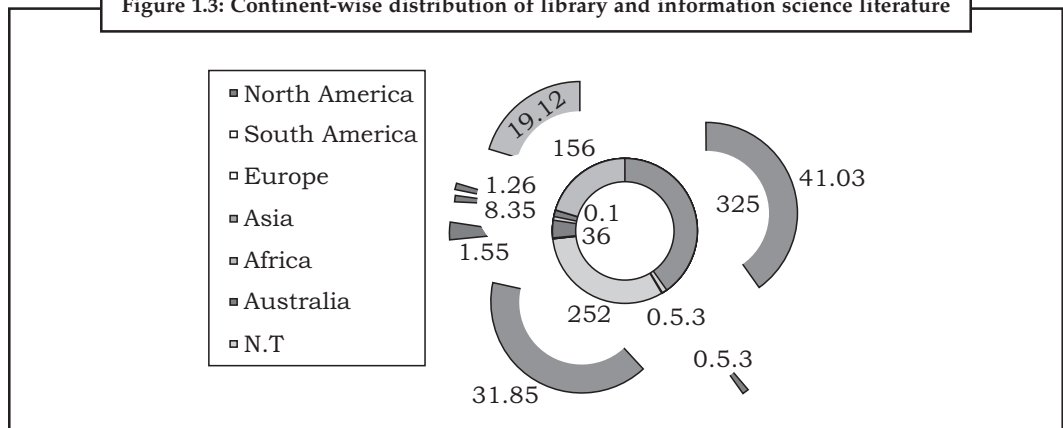
Table 1.3: Geographic output of library and information science literature

| <i>Continent</i>     | <i>Country of Publication</i> | <i>Channels of Publication</i> | <i>%</i> |
|----------------------|-------------------------------|--------------------------------|----------|
| <b>North America</b> | United States of America      | 301                            | 37.76    |
|                      | Canada                        | 22                             | 2.76     |
|                      | Costarica                     | 1                              | 0.12     |
|                      | Mexico                        | 1                              | 0.12     |
|                      | <b>Total</b>                  | <b>325</b>                     |          |
| <b>Europe</b>        | United Kingdom                | 164                            | 20.57    |
|                      | Netherlands                   | 27                             | 3.38     |
|                      | Germany                       | 19                             | 2.38     |
|                      | France                        | 7                              | 0.87     |
|                      | Spain                         | 5                              | 0.62     |
|                      | Italy                         | 4                              | 0.5      |
|                      | Poland                        | 4                              | 0.5      |
|                      | Croatia                       | 3                              | 0.37     |
|                      | Hungary                       | 3                              | 0.37     |
|                      | Denmark                       | 2                              | 0.25     |
|                      | Portugal                      | 2                              | 0.25     |
|                      | Sweden                        | 2                              | 0.25     |
|                      | Finland                       | 2                              | 0.25     |

Notes

|      |                    |            |      |
|------|--------------------|------------|------|
|      | Austria            | 2          | 0.25 |
|      | Turkey             | 2          | 0.25 |
|      | Ireland            | 1          | 0.12 |
|      | Belgium            | 1          | 0.12 |
|      | Slovakia           | 1          | 0.12 |
|      | Slovenia           | 1          | 0.12 |
|      | <b>Total</b>       | <b>252</b> |      |
| Asia | India              | 12         | 1.5  |
|      | Japan              | 10         | 1.25 |
|      | Taiwan             | 6          | 0.75 |
|      | China              | 2          | 0.25 |
|      | Malaysia           | 2          | 0.25 |
|      | Phillipines        | 1          | 0.12 |
|      | Iran               | 1          | 0.12 |
|      | Russian Federation | 1          | 0.12 |
|      | Pakistan           | 1          | 0.12 |
|      | <b>Total</b>       | <b>36</b>  |      |

Figure 1.3: Continent-wise distribution of library and information science literature



### Conclusion

The growth of library and information science literature is at a good pace. Developed countries are contributing a large chunk of literature through various types of publications. Developing countries like India have also made laudable contributions to library and information science literature. The chronological study indicates that library and information science literature has been expanding its subject boundaries. Most noticeably is its extensive coverage of IT-related services since 1980.

The maturity of the scientific aspect of the library science discipline has helped in increasing the literary output in the field of library and information science. Furthermore, the countries with a well established tradition in the field of LIS are showing an explosive growth in the LIS literature. Literature is making itself available in different forms. In the years to come, not only academic journals will be stealing the show but other forms like trade publications, monographs, and conference proceedings are also expected to reach the zenith in the field.

**Self Assessment**

Notes

Fill in the blanks:

5. The ..... was the first textbook on library science published in english anywhere in the world.
6. The first magazine ..... appeared in 1890's.
7. The first academic journal ..... was also launched in 1890's.
8. 1545 Journals were identified from ..... .
9. Shiyali Ramamrita Ranganathan was a ..... and ..... from India.

**1.5 Summary**

- Development of library and information science literature is mapped on the literary outputs available through Library, Information Science and Technology Abstracts (LISTA).
- Library and information science literature grew right from the time when the first core trade publication "Bookseller" (abstracted now in LISTA) appeared in 1852.
- The regional distribution of publication channels in library and information science vividly demonstrate that the developed countries provide more publication channels.
- Term "library science" seems to have been used for the first time in a book in 1916 in the "Panjab Library Primer" written by Asa Don Dickinson and published by the University of the Punjab, Lahore, Pakistan.
- Library science is constantly evolving, incorporating new topics like database management, information architecture and knowledge management, among others.
- Special librarians include almost any other form of librarianship, including those who serve in medical libraries (and hospitals or medical schools), corporations, news agencies, government organizations, or other special collections.
- The geographic output and the date of the first issue was extracted from Ulrichsweb One of the renowned source listing periodical publications all over the world.

**1.6 Keywords**

*Michael Gorman's Our Enduring Values* : Librarianship in the 21st Century, features his 8 principles necessary by library professionals and incorporate knowledge and information in all their forms, allowing for digital information to be considered.

*Not Traceable* : The literature whose chronology was not traceable was put into a separate category called "Not Traceable" (N.T.).

**1.7 Review Questions**

1. What do you mean by the term Library Science? When it was founded and by whom?
2. Who is written and published to the Punjab Library Primer?



Notes

3. Explain Librarians in different types of libraries
4. Write the full form of LISA, ISA, RZI, BS and CCA.
5. Explain Geographic distribution of Library and information science literature.
6. Describe briefly Development of Library Science literature.

**Answers: Self Assessment**

1. (b)
2. (d)
3. (c)
4. (a)
5. Punjab library primer
6. Author
7. New library world
8. Ulrich's guides
9. Mathematician, Librarian

**1.8 Further Readings**



*Books*

Maltby, A., ed. *Sayer's manual of classification for libraries*. 5th. Ed. London: Andre Deutsch, 1975.

Oddy, P. *Future libraries, future catalogs*. London: LA, 1996.

Deegan, M. and Simon Tanner. *Digital futures*. London. LA, 2002.



*Online links*

<http://encyclopedia.jrank.org/articles/pages/4/Librarian.html>

<http://careers.stateuniversity.com/pages/900/>

## Unit 2: Library Classification

### CONTENTS

Objectives

Introduction

- 2.1 Description of Library Classification
- 2.2 Types of Library Classification
  - 2.2.1 Comparing Classification Systems
- 2.3 Colon Classification
  - 2.3.1 Components of Ranganathan's Scheme
- 2.4 Dewey Decimal Classification
  - 2.4.1 Design
  - 2.4.2 Classes Listed
  - 2.4.3 Current Use
  - 2.4.4 Development
  - 2.4.5 Editions
  - 2.4.6 Structure and Notation
  - 2.4.7 Arrangement of the DDC
- 2.5 Universal Decimal Classification
- 2.6 Library of Congress
  - 2.6.1 Origins and Jefferson's Contribution (1800–1851)
- 2.7 Bliss Bibliographic Classification
- 2.8 Summary
- 2.9 Keywords
- 2.10 Review Questions
- 2.11 Further Readings

Notes

## Objectives

After studying this unit, you will be able to:

- Explain types of library classification
- Describe Colon and Dewey decimal classification
- Define universal decimal classification
- Explain library of congress
- Describe Bliss bibliographic classification.

## Introduction

Library classification, system of arrangement adopted by a library to enable patrons to find its materials quickly and easily. While cataloguing provides information on the physical and topical nature of the book (or other item), classification, through assignment of a call number (consisting of class designation and author representation), locates the item in its library setting and, ideally, in the realm of knowledge. Arranging similar things in some order according to some principle unites and controls information from various sources.

A library classification is a system of coding and organizing library materials (books, serials, audiovisual materials, computer files, maps, manuscripts) according to their subject and allocating a call number to that information resource. Similar to classification systems used in biology, bibliographic classification systems group entities together that are similar, typically arranged in a hierarchical tree structure. A different kind of classification system, called a faceted classification system, is also widely used which allows the assignment of multiple classifications to an object, enabling the classifications to be ordered in multiple ways.

### 2.1 Description of Library Classification

Library classification form part of the field of library and information science. It is a form of bibliographic classification (library classifications are used in library catalogs, while “bibliographic classification” also covers classification used in other kinds of bibliographic databases). It goes hand in hand with library (descriptive) cataloging under the rubric of cataloging and classification, sometimes grouped together as technical services. The library professional who engages in the process of cataloging and classifying library materials is called a cataloguer or catalog librarian. Library classification systems are one of the two tools used to facilitate subject access. The other consists of alphabetical indexing languages such as Thesauri and Subject Headings systems.

Library classification of a piece of work consists of two steps. Firstly, the “aboutness” of the material is ascertained. Next, a call number (essentially a book’s address) based on the classification system in use at the particular library will be assigned to the work using the notation of the system.

It is important to note that unlike subject heading or thesauri where multiple terms can be assigned to the same work, in library classification systems, each work can only be placed in one class. This is due to shelving purposes: A book can have only one physical place. However in classified catalogs one may have main entries as well as added entries. Most classification systems like the Dewey Decimal Classification (DDC) and Library of Congress classification also add a cutter number to each work which adds a code for the author of the work.

Some classification systems are more suitable for aiding subject access, rather than for shelf location. For example, UDC which uses a complicated notation including plus, colons are more difficult to use for the purpose of shelf arrangement but are more expressive compared to DDC in terms of

showing relationships between subjects. Similarly faceted classification schemes are more difficult to use for shelf arrangement, unless the user has knowledge of the citation order.

Depending on the size of the library collection, some libraries might use classification systems solely for one purpose or the other. In extreme cases a public library with a small collection might just use a classification system for location of resources but might not use a complicated subject classification system. Instead all resources might just be put into a couple of wide classes (Travel, Crime, Magazines etc.). This is known as a “mark and park” classification method, more formally called reader interest classification.



Task

Write a note on Library Classification and in how many types it is divided.

## 2.2 Types of Library Classification

There are many standard systems of library classification in use, and many more have been proposed over the years. However in general, Classification systems can be divided into three types depending on how they are used.

Universal schemes covering all subjects. Examples include Dewey Decimal Classification, Universal Decimal Classification and Library of Congress Classification.

Specific classification schemes for particular subjects or types of materials. Examples include Iconclass, British Catalogue of Music Classification, and Dickinson classification, or the NLM Classification for medicine.

**In terms of functionality, classification systems are often described as:**

Enumerative: produce an alphabetical list of subject headings; assign numbers to each heading in alphabetical order.

Hierarchical: divides subjects hierarchically, from most general to most specific faceted or analytico-synthetic: divides subjects into mutually exclusive orthogonal facets.

There are few completely enumerative systems or faceted systems, most systems are a blend but favouring one type or the other. The most common classification systems, LCC and DDC, are essentially enumerative, though with some hierarchical and faceted elements (more so for DDC), especially at the broadest and most general level. The first true faceted system was the Colon classification of S. R. Ranganathan.

**Universal classification systems used in the English-speaking world:**

- Dewey Decimal Classification (DDC)
- Library of Congress Classification (LCC)
- Bliss Bibliographic Classification (BC)

The above systems are the most common in the English-speaking world.

**BISAC Subject Headings:** The publishing industry standard for classification that is being adopted by some libraries.

**Harvard-Yenching Classification:** An English classification system for Chinese language materials.

A system of book classification for Chinese libraries (Liu’s Classification) or library classification for user:

- New Classification Scheme for Chinese Libraries
- Nippon Decimal Classification (NDC)
- Chinese Library Classification (CLC)
- Korean Decimal Classification (KDC)

**Notes**

- Library-Bibliographic Classification (BBK) from Russia.

Universal classification systems that rely on synthesis (faceted systems)

- Bliss Bibliographic Classification
- Colon Classification
- Cutter Expansive Classification
- Universal Decimal Classification.

Newer classification systems tend to use the principle of synthesis (combining codes from different lists to represent the different attributes of a work) heavily, which is comparatively lacking in LC or DDC.



*Notes* Colon classification of S.R.Ranganathan was the first true faceted system.

## 2.2.1 Comparing Classification Systems

As a result of differences in Notation, history, use of enumeration, hierarchy, facets, classification systems can differ in the following ways:

**Type of Notation:** Notation can be pure (consisting of only numerals, for example) or mixed (consisting of letters, numerals, and other symbols).

**Expressiveness:** This is the degree in which the notation can express relationship between concepts or structure.

**Whether they support mnemonics:** For example the number 44 in DDC notation usually means it concerns some aspect of France. For example 598.0944 concerns “Birds in France”. The 09 signifies country code, and 44 represent France.

**Hospitality:** The degree in which the system is able to accommodate new subjects.

**Brevity:** Length of the notation to express the same concept.

**Speed of updates and degree of support:** The best classification systems are constantly being reviewed and improved.

- Consistency
- Simplicity
- Usability.

## 2.3 Colon Classification

### 2.3.1 Components of Ranganathan’s Scheme

The Colon Classification uses 42 main classes that are combined with other letters, numbers and marks in a manner resembling the Library of Congress Classification to sort a publication. The Colon Classification, just as other classification schemes, starts with a number of main classes (42), which represent the fields of knowledge.



*Did u know?* Each class is analyzed and broken down into its basic elements, grouped together by common attributes, called facets.

Upon examining all the facets, Ranganathan notices that there are five main groups into which the facets fall, and he calls these the fundamental categories, represented by the mnemonic PMEST in an order of decreasing concreteness.

### Facets

CC uses five primary categories, or facets to further specify the sorting of a publication collectively called "PMEST".

Fundamental categories

, personality

; matter-property

: energy

. space

' time

Time isolates

...

'M 1800 to 1899

'N 1900 to 1999

'N5 1950's

'P 2000 to 2099

...

z Generalia

[material] , [kind] , ...

1 Universe of knowledge

2 Library science

[library] ; [material] : [problem]

2;45:6 circulation of newspapers

234:81 book selection in university library

234;45:81 newspaper selection in university library

3 Book science

4 Journalism

### Index

B Mathematics

C Physics

D Engineering

E Chemistry

F Technology

G Biology

H Geology

I Botany

J Agriculture

K Zoology

L Medicine

M Useful arts

Δ Mysticism

N Fine arts

O Literature

P Linguistics

Q Religion

R Philosophy

S Psychology

T Education

U Geography

V History

W Polytical sc.

X Economics

Y Sociology

Z Law

## 2.4 Dewey Decimal Classification

The Dewey Decimal Classification (DDC, also called the Dewey Decimal System) is a proprietary system of library classification developed by Melvil Dewey in 1876.

It has been greatly modified and expanded through 23 major revisions, the most recent in 2011. This system organizes books on library shelves in a specific and repeatable order that makes it easy to find any book and return it to its proper place. The system is used in 200,000 libraries in at least 135 countries.

Notes



*Notes* A designation such as Dewey 16 refers to the 16th edition of the DDC.

### 2.4.1 Design

The DDC attempts to organize all knowledge into ten main classes. The ten main classes are each further subdivided into ten divisions, and each division into ten sections, giving ten main classes, 100 divisions and 1000 sections. DDC's advantage in using decimals for its categories allows it to be both purely numerical and infinitely hierarchical. It also uses some aspects of a faceted classification scheme, combining elements from different parts of the structure to construct a number representing the subject content (often combining two subject elements with linking numbers and geographical and temporal elements) and form of an item rather than drawing upon a list containing each class and its meaning.

The DDC has a number for all books, including fiction: American fiction is classified in 813. Most libraries create a separate fiction section to allow shelving in a more generalized fashion than Dewey provides for, or to avoid the space that would be taken up in the 800s, or simply to allow readers to find preferred authors by alphabetical order of surname.

Some parts of the classification offer options to accommodate different kinds of libraries. An important feature of the scheme is the ability to assign multiple class numbers to a bibliographical item and only use one of them for shelving. The added numbers appear in the classified subject catalogue (though this is not the usual practice in North America). For the full benefit of the scheme the relative index and the tables that form part of every edition must be understood and consulted when required.

### 2.4.2 Classes Listed

The system is made up of seven tables and ten main classes, each of which is divided into ten secondary classes or subcategories, each of which contain ten subdivisions.

**The tables are:**

- Standard subdivision
- Areas
- Sub-division of individual literatures
- Sub-divisions of individual languages
- Racial, ethnic, national groups
- Languages
- Persons
- The classes are:
  - 000 – Computer science, information and general works
  - 100 – Philosophy and psychology
  - 200 – Religion
  - 300 – Social sciences
  - 400 – Language
  - 500 – Science (including mathematics)

- 600 – Technology and applied Science
- 700 – Arts and recreation
- 800 – Literature
- 900 – History, geography and biography

### 2.4.3 Current Use

The Dewey Decimal Classification (DDC) system is a general knowledge organization tool that is continuously revised to keep pace with knowledge. The system was conceived by Melvil Dewey in 1873 and first published in 1876. The DDC is published by OCLC Online Computer Library Center, Inc. OCLC owns all copyright rights in the Dewey Decimal Classification, and licenses the system for a variety of uses.

The DDC is the most widely used classification system in the world. Libraries in more than 135 countries use the DDC to organize and provide access to their collections, and DDC numbers are featured in the national bibliographies of more than 60 countries. Libraries of every type apply Dewey numbers on a daily basis and share these numbers through a variety of means (including WorldCat, the OCLC Online Union Catalog). Dewey is also used for other purposes, e.g., as a browsing mechanism for resources on the web.

### 2.4.4 Development

One of Dewey's great strengths is that the system is developed and maintained in a national bibliographic agency, the Library of Congress. The Dewey editorial office is located in the Decimal Classification Division of the Library of Congress, where classification specialists annually assign over 110,000 DDC numbers to records for works cataloged by the Library. Having the editorial office within the Decimal Classification Division enables the editors to detect trends in the literature that must be incorporated into the Classification. The editors prepare proposed schedule revisions and expansions, and forward the proposals to the Decimal Classification Editorial Policy Committee (EPC) for review and recommended action.

EPC is a ten-member international board whose main function is to advise the editors and OCLC on matters relating to changes, innovations, and the general development of the Classification. EPC represents the interests of DDC users; its members come from national, public, special, and academic libraries, and from library schools.

### 2.4.5 Editions

The DDC is published in full and abridged editions in print and electronic versions. The abridged edition is a logical truncation of the notational and structural hierarchy of the corresponding full edition on which it is based, and is intended for general collections of 20,000 titles or less. WebDewey and Abridged WebDewey, the electronic versions of the full and abridged editions, respectively, are updated frequently and contain additional index entries and mapped vocabulary. The electronic versions and supplemental web postings are the chief sources of ongoing updates to the DDC. On the Dewey web site, selected new numbers and changes to the DDC are posted monthly, and mappings between selected new Library of Congress Subject Headings (LCSH) and Dewey numbers are posted biweekly.

### 2.4.6 Structure and Notation

The DDC is built on sound principles that make it ideal as a general knowledge organization tool: meaningful notation in universally recognized Arabic numerals, well-defined categories, well-



**Notes**

developed hierarchies, and a rich network of relationships among topics. In the DDC, basic classes are organized by disciplines or fields of study. At the broadest level, the DDC is divided into ten main classes, which together cover the entire world of knowledge.

Each main class is further divided into ten divisions, and each division into ten sections (not all the numbers for the divisions and sections have been used). The main structure of the DDC is presented in the DDC Summaries following this introduction. The headings associated with the numbers in the summaries have been edited for browsing purposes, and do not necessarily match the complete headings found in the schedules.

The first summary contains the ten main classes. The first digit in each three-digit number represents the main class. For example, 600 represent technology.

The second summary contains the hundred divisions. The second digit in each three-digit number indicates the division. For example, 600 is used for general works on technology, 610 for medicine and health, 620 for engineering, 630 for agriculture.

The third summary contains the thousand sections. The third digit in each three-digit number indicates the section. Thus, 610 is used for general works on medicine and health, 611 for human anatomy, 612 for human physiology, 613 for personal health and safety.

**Hierarchy**

Hierarchy in the DDC is expressed through structure and notation. Structural hierarchy means that all topics (aside from the ten main classes) are part of all the broader topics above them. Any note regarding the nature of a class holds true for all the subordinate classes, including logically subordinate topics classed at coordinate numbers.

Notational hierarchy is expressed by length of notation. Numbers at any given level are usually subordinate to a class whose notation is one digit shorter; coordinate with a class whose notation has the same number of significant digits; and super ordinate to a class with numbers one or more digits longer. The underlined digits in the following example demonstrate this notational hierarchy:

- 600      Technology
- 630      Agriculture and related technologies
- 636      Animal husbandry
- 636.7    Dogs
- 636.8    Cats

“Dogs” and “Cats” are more specific than (i.e., are subordinate to) “Animal husbandry”; they are equally specific as (i.e., are coordinate with) each other; and “Animal husbandry” is less specific than (i.e., is super ordinate to) “Dogs” and “Cats.” Sometimes, other devices must be used to express the hierarchy when it is not possible or desirable to do so through the notation. Special headings, notes, and entries indicate relationships among topics that violate notational hierarchy.

**2.4.7 Arrangement of the DDC**

The print version of the latest full edition of the DDC, Edition 22, is composed of the following major parts in four volumes:

**Volume 1**

- (A) *New Features in Edition 22*: A brief explanation of the special features and changes in DDC 22.
- (B) *Introduction*: A description of the DDC and how to use it.
- (C) *Glossary*: Short definitions of terms used in the DDC.

- (D) Index to the Introduction and Glossary.
- (E) *Manual*: A guide to the use of the DDC that is made up primarily of extended discussions of problem areas in the application of the DDC. Information in the Manual is arranged by the numbers in the tables and schedules.
- (F) *Tables*: Six numbered tables of notation that can be added to class numbers to provide greater specificity.
- (G) *Lists that compare Editions 21 and 22*: Relocations and Discontinuations; Reused Numbers.

**Volume 2**

- (H) *DDC Summaries*: The top three levels of the DDC.
- (I) *Schedules*: The organization of knowledge from 000–599.

**Volume 3**

- (J) *Schedules*: The organization of knowledge from 600–999.

**Volume 4**

- (K) *Relative Index*: An alphabetical list of subjects with the disciplines in which they are treated sub arranged alphabetically under each entry.

**Entries**

Entries in the schedules and tables are composed of a DDC number in the number column (the column at the left margin), a heading describing the class that the number represents, and often one or more notes. All entries (numbers, headings, and notes) should be read in the context of the hierarchy.

In the print version of the DDC, the first three digits of schedule numbers (main classes, divisions, sections) appear only once in the number column, when first used. They are repeated at the top of each page where their subdivisions continue. Subordinate numbers appear in the number column, beginning with a decimal point, with the initial three digits understood.

Some numbers in the schedules and tables are enclosed in parentheses or square brackets. Numbers and notes in parentheses provide options to standard practice. Numbers in square brackets represent topics that have been relocated or discontinued, or are unassigned. Square brackets are also used for standard subdivision concepts that are represented in another location. Numbers in square brackets are never used.

**Number Building**

Only a fraction of potential DDC numbers are included in the schedules. It is often necessary to build or synthesize a number that is not specifically listed in the schedules. Such built numbers allow for greater depth of content analysis. There are four sources of notation for building numbers: (A) Table 2.1 Standard Sub-divisions; (B) Tables 2.2–2.4; (C) other parts of the schedules; and (D) add tables in the schedules.

Number building is initiated only upon instructions in the schedules (except for the addition of standard subdivisions, which may take place anywhere unless there is an instruction to the contrary). Number building begins with a base number (always stated in the instruction note) to which another number is added.

**2.5 Universal Decimal Classification**

The Universal Decimal Classification is a system of library classification developed by the Belgian bibliographers Paul Otlet and Henri La Fontaine at the end of the 19th century. It is based on the

**Notes**

Dewey Decimal Classification, but uses auxiliary signs to indicate various special aspects of a subject and relationships between subjects. It thus contains a significant faceted or analytic-synthetic element, and is used especially in specialist libraries. UDC has been modified and extended through the years to cope with the increasing output in all disciplines of human knowledge, and is still under continuous review to take account of new developments.



*Notes* The documents classified by UDC may be in any form. They will often be literature, *i.e.* written documents, but may also be in other media such as films, video and sound recordings, illustrations, maps, and realia such as museum pieces.

UDC classifications use Arabic numerals and are based on the decimal system. Every number is thought of as a decimal fraction with the initial decimal point omitted, which determines filing order. For ease of reading, a UDC identifier is usually punctuated after every third digit. Thus, after 61 "Medical sciences" come the subdivisions 611 to 619; under 611 "Anatomy" come its subdivisions 611.1 to 611.9; under 611.1 come all of its subdivisions before 611.2 occurs, and so on; after 619 comes 620. An advantage of this system is that it is infinitely extensible, and when new subdivisions are introduced, they need not disturb the existing allocation of numbers.

**The Main Categories**

0. generalities
1. philosophy, psychology
2. religion, theology
3. social sciences
4. Philosophy, Linguistics, Languages
5. natural sciences
6. technology
7. the arts
8. language, linguistics, literature
9. geography, biography, history

A document may be classified under a combination of different categories through the use of additional symbols. For example:

| <i>Symbol</i> | <i>Symbol name</i> | <i>Meaning</i>        | <i>Example</i>   |
|---------------|--------------------|-----------------------|--|
| +             | plus               | addition              | <i>e.g.</i> , 59 + 636 zoology and animal breeding   |
| /             | stroke             | extension             | <i>e.g.</i> , 592/599 Systematic zoology (everything from 592 to 599 inclusive)  |
| :             | colon              | relation              | <i>e.g.</i> , 17:7 Relation of ethics to art   |
| [ ]           | square brackets    | algebraic subgrouping | <i>e.g.</i> , 311:[622 + 669](485) statistics of mining and metallurgy in Swedn (the auxiliary qualifies 622 + 669 considered as a unit) |
| =             | equals             | language              | <i>e.g.</i> , = 111 in English; 59 = 111 Zoology, in English   |

The design of UDC lends itself to machine readability, and the system has been used both with early automatic mechanical sorting devices, and modern library OPACs. A core version of UDC, with 65,000 subdivisions, is now available in database format, and is called the Master Reference File (MRF). The current full version of the UDC has 220,000 subdivisions.

## 2.6 Library of Congress

The Library of Congress is the research library of the United States Congress, de facto national library of the United States, and the oldest federal cultural institution in the United States. Located in three buildings in Washington, D.C., it is the largest library in the world by shelf space and number of books. The head of the Library is the Librarian of Congress, currently James H. Billington.

The Library of Congress was built by Congress in 1800, and was housed in the United States Capitol for most of the 19th century. After much of the original collection had been destroyed during the War of 1812.



*Did u know?* Thomas Jefferson sold 6,487 books, his entire personal collection, to the library in 1815.

After a period of decline during the mid-19th century the Library of Congress began to grow rapidly in both size and importance after the American Civil War, culminating in the construction of a separate library building and the transference of all copyright deposit holdings to the Library. During the rapid expansion of the 20th century the Library of Congress assumed a preeminent public role, becoming a “library of last resort” and expanding its mission for the benefit of scholars and the American people.

The Library’s primary mission is researching inquiries made by members of Congress through the Congressional Research Service. Although it is open to the public, only Members of Congress, Supreme Court justices and other high-ranking government officials may check out books. As the de facto national library, the Library of Congress promotes literacy and American literature through projects such as the American Folklife Center, American Memory, Center for the Book and Poet Laureate.

### 2.6.1 Origins and Jefferson’s Contribution (1800–1851)

The Library of Congress was established on April 24, 1800, when President John Adams signed an Act of Congress providing for the transfer of the seat of government from Philadelphia to the new capital city of Washington. Part of the legislation appropriated \$5,000 “for the purchase of such books as may be necessary for the use of Congress ... and for fitting up a suitable apartment for containing them....” Books were ordered from London and the collection, consisting of 740 books and 3 maps, was housed in the new Capitol. The collection covered a variety of topics but the bulk of the materials were legal in nature, reflecting Congress’ role as a maker of laws.

Thomas Jefferson played an important role in the Library’s early formation, signing into law on January 26, 1802, the first law establishing the structure of the Library of Congress. The law established the presidentially appointed post of Librarian of Congress and a Joint Committee on the Library to regulate and oversee the Library, as well as giving the president and vice president the ability to borrow books. The Library of Congress was destroyed in August 1814, when invading British troops set fire to the Capitol building and the small library of 3,000 volumes within.

Within a month, former President Jefferson offered his personal library as a replacement. Jefferson had spent 50 years accumulating a wide variety of books, including ones in foreign languages and

**Notes**

volumes of philosophy, science, literature, and other topics not normally viewed as part of a legislative library, such as cookbooks, writing that, "I do not know that it contains any branch of science which Congress would wish to exclude from their collection; there is, in fact, no subject to which a Member of Congress may not have occasion to refer."



*Notes*

In January 1815, Congress accepted Jefferson's offer, appropriating \$23,950 for his 6,487 books.

**Weakening (1851–1865)**

The antebellum period was difficult for the Library. During the 1850s the Smithsonian Institution's librarian Charles Coffin Jewett aggressively tried to move that organization towards becoming the United States' national library. His efforts were blocked by the Smithsonian's Secretary Joseph Henry, who advocated a focus on scientific research and publication and favoured the Library of Congress' development into the national library. Henry's dismissal of Jewett in July 1854 ended the Smithsonian's attempts to become the national library, and in 1866 Henry transferred the Smithsonian's forty thousand-volume library to the Library of Congress.

On December 24, 1851 the largest fire in the Library's history destroyed 35,000 books, about two-thirds of the Library's 55,000 book collection, including two-thirds of Jefferson's original transfer. Congress in 1852 quickly appropriated \$168,700 to replace the lost books, but not for the acquisition of new materials. This marked the start of a conservative period in the Library's administration under Librarian John Silva Meehan and Joint Committee Chairman James A. Pearce, who worked to restrict the Library's activities.

In 1857, Congress transferred the Library's public document distribution activities to the Department of the Interior and its international book exchange program to the Department of State. Abraham Lincoln's political appointment of John G. Stephenson as Librarian of Congress in 1861 further weakened the Library; Stephenson's focus was on non-library affairs, including service as a volunteer aide-de-camp at the battles of Chancellorsville and Gettysburg during the American Civil War. By the conclusion of the war, the Library of Congress had a staff of seven for a collection of 80,000 volumes. The centralization of copyright offices into the United States Patent Office in 1859 ended the Library's thirteen year role as a depository of all copyrighted books and pamphlets.

**Spofford's Expansion (1865–1897)**

The Library of Congress reasserted itself during the latter half of the 19th century under Librarian Ainsworth Rand Spofford, who directed the Library from 1865 to 1897. Aided by an overall expansion of the federal government and a favorable political climate, Spofford built broad bipartisan support for the Library as a national library and a legislative resource, began comprehensively collecting Americana and American literature, and led the construction of a new building to house the Library, and transformed the Librarian of Congress position into one of strength and independence.

Between 1865 and 1870, Congress appropriated funds for the construction of the Thomas Jefferson Building, placed all copyright registration and deposit activities under the Library's control, and restored the Library's international book exchange. The Library also acquired the vast libraries of both the Smithsonian and historian Peter Force, strengthening its scientific and Americana collections significantly. By 1876, the Library of Congress had 300,000 volumes and was tied with Boston Public Library as the nation's largest library.



*Notes*

The Library moved from the Capitol building to its new headquarters in 1897, it had over 840,000 volumes, 40% of which had been acquired through copyright deposit.

A year before the Library's move to its new location, the Joint Library Committee held a session of hearings to assess the condition of the Library and plan for its future growth and possible reorganization. Spofford and six experts sent by the American Library Association, including future Librarian of Congress Herbert Putnam and Melvil Dewey of the New York State Library, testified before the committee that the Library should continue its expansion towards becoming a true national library.

Based on the hearings and with the assistance of Senators Justin Morrill of Vermont and Daniel Voorhees of Indiana, Congress more than doubled the Library's staff from 42 to 108 and established new administrative units for all aspects of the Library's collection. Congress also strengthened the office of Librarian of Congress to govern the Library and make staff appointments, as well as requiring Senate approval for presidential appointees to the position.

### Post-reorganization (1897–1939)

The Library of Congress, spurred by the 1897 reorganization, began to grow and develop more rapidly. Spofford's successor John Russell Young, though only in office for two years, overhauled the Library's bureaucracy, used his connections as a former diplomat to acquire more materials from around the world, and established the Library's first assistance programmes for the blind and physically disabled.

Young's successor Herbert Putnam held the office for forty years from 1899 to 1939, entering into the position two years before the Library became the first in the United States to hold one million volumes. Putnam focused his efforts on making the Library more accessible and useful for the public and for other libraries. He instituted the interlibrary loan service, transforming the Library of Congress into what he referred to as a "library of last resort". Putnam also expanded Library access to "scientific investigators and duly qualified individuals" and began publishing primary sources for the benefit of scholars.

Putnam's tenure also saw increasing diversity in the Library's acquisitions. In 1903 he persuaded President Theodore Roosevelt to transfer by executive order the papers of the Founding Fathers from the State Department to the Library of Congress. Putnam expanded foreign acquisitions as well, including the 1904 purchase of a four-thousand volume library of Indica, the 1906 purchase of G. V. Yudin's eighty-thousand volume Russian library, the 1908 Schatz collection of early opera librettos, and the early 1930s purchase of the Russian Imperial Collection, consisting of 2,600 volumes from the library of the Romanov family on a variety of topics.



*Did u know?*

In 1903, Herbert Putnam persuaded President Theodore Roosevelt to transfer the executive order by the papers of the founding fathers from the state Department to the library of congress.

## 2.7 Bliss Bibliographic Classification

The Bliss bibliographic classification (BC) is a library classification system that was created by Henry E. Bliss (1870–1955), published in four volumes between 1940 and 1953. Although originally devised in the United States, it was more commonly adopted by British libraries than by American ones. A second edition of the system (BC2) has been developed in Britain since 1977.

The Bibliographic Classification (BC2 or Bliss) is the leading example of a fully faceted classification scheme. It provides a detailed classification for use in libraries and information services of all kinds, having a broad and detailed structure and order.

**Notes**

The vocabulary in each class is comprehensive and complemented by an exceptionally brief faceted notation considering the detail available, providing indexing to any depth the classifier wishes.

The structure of the subject within each class is clearly and simply laid out with rules provided for the quick and consistent placing of any item. A thorough A-Z index is provided in each volume. Users can access a subject catalogue record via any part of the whole, depending upon the primary interest of the user.

**BC1**

The Classification (known as BC) was originally devised by Henry Evelyn Bliss and was first published in four volumes in the USA 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 the logical principles of classification and consistent with the systems of science and education, may be available to services in libraries, “to aid revision ... of long established ... classifications” 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 of the classification is now known as BC1.

BC1 was first applied in broad outline at the College of the City of New York (where Bliss was librarian) in 1902. The full scheme followed the publication of two massive theoretical works on the organization of knowledge. Its main feature was the carefully designed main class order, reflecting the Comptean principle of gradation in specialty. Work on a radical revision of BC1, incorporating the great advances in logical facet analysis initiated by Ranganathan and developed by the Classification Research Group in Britain, began in the early 1970s.

**Origins of the System**

Bliss was born in New York in 1870 and in 1891 began work in the library of the College of the City of New York (now City College of the City University of New York).

Bliss had a lifelong interest in the organization, structure and philosophy of knowledge and was very critical of the library classification systems that were available to him. He believed that because the popular Library of Congress system had been designed for a specific library (the Library of Congress) it had no use as a standard system outside that library. He also greatly disliked the Dewey Decimal system.

Bliss wanted a classification system that would provide distinct rules yet still be adaptable to whatever kind of collection a library might have, as different libraries have different needs. His solution was the concept of “alternative location,” in which a particular subject could be put in more than one place, as long as the library made a specific choice and used it consistently.

In 1908 Bliss reclassified 60,000 of his library’s books, and in 1910 he published an article with a rough scheme of his general ideas. But as he continued to develop his system he realized that it was going to be a much larger project than he had anticipated. The first of his four volumes appeared in 1940 (the year he retired) and the last in 1953, two years before his death.

**Some of the underlying policies of the BC system were:**

Alternative location

Brief, concise notation

Organizing knowledge according to academic expertise

Subjects moving gradually from topic to topic as they naturally related to one another.



*Example:* Bliss deliberately avoided the use of the decimal point because of his objection to

Dewey's system. Instead he used capital and lower-case letters, numerals, and every typographical symbol available on his extensive and somewhat eccentric typewriter.

In the revised edition (BC2), only capital letters are used, with numerals occasionally used for special purposes. Here is an extract:

HJ Preventive medicine

HL Curative medicine

HLK Primary care; general practice

HLY Secondary care, aftercare

### Adoption and Change

BC was not used by many North American libraries. The system was not without its flaws (vague) (the result of being largely a one-person project), and the layout of Bliss's text was difficult to read. A few library schools sometimes taught the BC system to their students, but only in a minor way. The failure of the system to catch on in North America was partly because of its internal deficiencies but also because the Dewey Decimal and Library of Congress systems were already well established.

The City College library continued to use Bliss's system until 1967, when it reluctantly switched to the Library of Congress system. It had become too expensive to train new staff members to use BC, and too expensive to maintain in general. Much of the Bliss stacks remain, however, as no-one has re-catalogued the books.

The case was different, however, in Britain. BC proved more popular there and also spread to other English-speaking countries. Part of the reason for its success was that libraries in teachers' colleges liked the way Bliss had organized the subject areas on teaching and education. By the mid-1950s the system was being used in at least sixty British libraries and in a hundred by the 1970s.

In 1967 the Bliss Classification Association was formed. Its first publication was the Abridged Bliss Classification (ABC), intended for school libraries. In 1977 it began to publish and maintain a much-improved, revised version of Bliss's system, the Bliss Bibliographic Classification (Second Edition) or BC2. This retains only the broad outlines of Bliss's scheme, replacing most of the detailed notation with a new scheme based on the principles of faceted classification. 15 of approximately 28 volumes of schedules have so far been published.

The top level organization is:

2/9 – Generalia, Phenomena, Knowledge, Information science & technology

A/AL – Philosophy & Logic

AM/AX – Mathematics, Probability, Statistics

AY/B – General science, Physics

C – Chemistry

D – Astronomy and earth sciences

DG/DY – Earth sciences

E/GQ – Biological sciences

GR/GZ – Applied biological sciences: agriculture and ecology

H – Physical Anthropology, Human biology, Health sciences



**Notes**

- I – Psychology & Psychiatry
- J – Education
- K – Society (includes Social sciences, sociology & social anthropology)
- L/O – History (including area studies, travel and topography, and biography)
- LA – Archaeology
- P – Religion, Occult, Morals and ethics
- Q – Social welfare & Criminology
- R – Politics & Public administration
- S – Law
- T – Economics & Management of economic enterprises
- U/V – Technology and useful arts (including household management and services)
- W – The Arts
- WV/WX – Music
- X/Y – Language and literature
- ZA/ZW – Museology

**Self Assessment**

Multiple Choice Questions:

1. Second edition of the system (BC2) has been developed in Britain since:  
(a) 1967 (b) 1975  
(c) 1977 (d) 1976
2. Bliss was born in New York in .....  
(a) 1860 (b) 1870  
(c) 1880 (d) 1891
3. In 1908 Bliss reclassified ..... of his library's book.  
(a) 40,000 (b) 50,000  
(c) 60,000 (d) 70,000
4. In ..... the Bliss classification association was formed.  
(a) 1964 (b) 1965  
(c) 1966 (d) 1967

**2.8 Summary**

- Library classification, system of arrangement adopted by a library to enable patrons to find its materials quickly and easily.
- The Dewey Decimal Classification (DDC, also called the Dewey Decimal System).
- The DDC attempts to organize all knowledge into ten main classes.
- DDC's advantage in using decimals for its categories.
- The Universal Decimal Classification developed by the Belgian bibliographers Paul Otlet and Henri La Fontaine at the end of the 19th century.
- The Library of Congress is the research library of the United States Congress.

- The Library of Congress was established on April 24, 1800.
- The Bliss bibliographic classification was created by Henry E. Bliss.
- Bliss deliberately avoided the use of the decimal point.

Notes

## 2.9 Keywords

|   |  |
|---|--|
| <i>Library Classification</i>             | : Is a form of bibliographic classification  |
| <i>Colon Classification</i>               | : Uses 42 main classes that are combined with other letters, numbers and marks in a manner resembling the Library of Congress Classification to a sort a publication |
| <i>Bliss Bibliographic Classification</i> | : Is a library classification system that was created by Henry. E. Bliss (1870–1955).  |

## 2.10 Review Questions

1. What is Library classification system?
2. Abbreviate DDC.
3. Write the function and advantage of DDC.
4. Who developed Universal Decimal Classification (UDC)?
5. Mention the primary mission of Library of Congress.
6. Write the full form of BC and its origin.

## Answers: Self Assessment

1. (c)
2. (b)
3. (c)
4. (d)

## 2.11 Further Readings



Books

Aitchinson, J and Gilchrist, A: *Thesaurus consturction*. 2nd ed. London: Aslib, 1987.

Deegan. M. and Simon Tanner. *Digital futures*. London. LA, 2002.

Maltby, A., ed. *Sayer's manual of classification for libraries*. 5th. Ed. London: Andre Deutsch, 1975.



Online links

<http://skuastkashmir.ac.in/>

<http://xa.yimg.com/kq/groups/1392795/1453698045/name/>

<http://www.britannica.com/EBchecked/topic/126159/Colon-Classification>

<http://skuastkashmir.ac.in/>

## Unit 3: Organization in Classification Research

### CONTENTS

Objectives

Introduction

- 3.1 Documentation Research and Training Centre
- 3.2 International Society for Knowledge Organization
- 3.3 Classification Research Group
- 3.4 Summary
- 3.5 Keywords
- 3.6 Review Questions
- 3.7 Further Readings

### Objectives

After studying this unit, you will be able to:

- Know about the fundamental of classification
- Know about the research institutes and their functions
- Know about the international society for knowledge organization.

### Introduction

The Classification Research Group (CRG) was a significant contributor to classification research and theory in the field of library and information science in the latter half of the 20th century. It was formed in England in 1952 and was active until 1968. Among its members were Derek Austin, Eric Coates (classification researcher), Jason Farradane, Robert Fairthorne, Douglas Foskett, Barbara Kyle, Derek Lang ridge, Jack Mills (classification researcher), Bernard Palmer, Jack Wells and Brian Campbell Vickery.

The group formed important principles on faceted classification and also worked on the theory of Integrative levels. An integrative level, or level of organization, is a set of phenomena emerging on pre-existing phenomena of lower level. Typical examples include life emerging on non-living substances, and consciousness emerging on nervous systems. Integrative levels, or the disciplines focusing on them, form the main classes of several knowledge organization systems, including Roget's Thesaurus, the Bliss bibliographic classification, the Colon classification, and the Information coding classification.

## Characteristics of a Classification System

Notes

- Inclusive as well as comprehensive.
- Systematic.
- Flexible and expansive.
- Employ terminology that is clear and descriptive.

## The Nature of Book Classification

*Collocating objective:* Bringing like things together on library shelves:

*Subject criterion:* What about books on multiple topics?

*Author criterion:* What about books by multiple authors?

*Subject/author criteria:* What about books by the same author, but on different topics?

Solving the need for a system of unique identification in open stack libraries through notational systems and call numbers.

### 3.1 Documentation Research and Training Centre

Documentation Research and Training Centre (DRTC) is a research centre (and a department) for library and information science and allied disciplines at the Indian Statistical Institute, Bangalore. DRTC runs a graduate program leading to the award of a 'Master of Science in Library and Information Science' (MS-LIS) from the Indian Statistical Institute as well as serving as an academic and research center for Research Fellows registered for a PhD in Information Science.



*Did u know?* DRTC was established in January 1962 as a division of the Indian Statistical Institute.

DRTC developed as a result of social forces. Soon after independence, the Government of India created the Indian Standards Institution in 1947. In the same year, its Documentation (Sectional) Committee was formed with Prof. S.R. Ranganathan as chairman.

A proposal was made to Union Ministry of Education for the establishment of a National Documentation Centre. The proposal was referred to a committee of professors which included Prof. S.R. Ranganathan. In 1949, the files were taken over by Dr. Shanti Swarup Bhatnagar. There was a keenly felt need for document services to support the work done in the national laboratories that were just being established. In 1950, Dr. K.S. Krishnan, the Director of the National Physics Laboratory and Prof. S.R. Ranganathan were authorised to negotiate with UNESCO for aid in setting up a National Documentation Centre.

The result was the establishment of Indian National Scientific Documentation Centre (INSDOC) in September 1951. By about 1955, some industries had been established. The research activities in the national laboratories had also begun to accelerate to a higher pitch. Specialist Libraries to support research activities were being established in some of these institutions. Thus the time appeared to be ripe for the formation of a special library association to support specialist library activity and documentation.

**Notes**

Thus was born IASLIC. Prof. P.C. Mahalanobis, member of the Planning Commission and then Director of Indian Statistical Institute, had all along been engaged in perspective planning. He sensed the dependence of productivity of industries and of research in the country on prompt and pinpointed documentation services. As early as 1956, he requested Prof. Ranganathan, who was then in Zurich to come back to India and to establish a training school. But Prof. Ranganathan felt that the time was not ripe enough.

DRTC is widely considered to be the best research centre in India in the fields of library science and information science. It also has a very strong research program and a Ph.D collaboration with the University of Trento, Italy.

**Self Assessment**

Fill in the blanks:

1. In 1947, its documentation (sectional) committee was formed with prof. .... as chairman.
2. A proposal was made to union ministry of Education for the establishment of a .....
3. The result was the establishment of Indian National Scientific Documentation Centre (INSDOC) in .....
4. DRTC is widely considered to be the best research centre in India in the field of ..... and .....

**3.2 International Society for Knowledge Organization**

The International Society for Knowledge Organization (also referred to by its abbreviation ISKO) is the principal professional association for scholars of knowledge organization, knowledge structures, classification studies, and information organization and structure.



*Notes* Founded in 1989, ISKO's mission is "to advance conceptual work in knowledge organization in all kinds of forms, and for all kinds of purposes, such as databases, libraries, dictionaries and the Internet."

An interdisciplinary association, ISKO's worldwide membership draws from fields such as information science, philosophy, linguistics, library science, archive studies, science studies, and computer science.

ISKO "promotes research, development and applications of knowledge organization systems that advance the philosophical, psychological and semantic approaches for ordering knowledge; provides the means of communication and networking on knowledge organization for its members; and functions as a connecting link between all institutions and national societies, working with problems related to the conceptual organization and processing of knowledge."

The Society publishes the quarterly academic journal Knowledge Organization, and it holds an international conference every two years. It officially recognizes national chapters in Brazil, Canada, China, France, Germany, India, Italy, Poland, Spain, the United Kingdom, and the United States. ISKO cooperates with international and national organizations such as UNESCO, the European Commission, the International Organization for Standardization, the International Federation of Library Associations and Institutions, the American Society for Information Science and Technology, the Networked Knowledge Organization Systems/Services, and the International Information Centre for Terminology.

## Knowledge Organization (Journal)

Notes

Founded in 1973, Knowledge Organization (sometimes abbreviated as KO) is the official quarterly double-blind peer-reviewed academic journal of ISKO. It was formerly known as International Classification until 1993, when the title changed to its current form. Published in English, the Society describes the journal's scope this way:

In each issue, experts from many countries comment on questions of the adequate structuring and construction of ordering systems and on the problems of their use in providing access to the information contents of new literature, of data collections and survey, of tabular works and of other objects of scientific interest. Contributions: (1) clarify theoretical foundations (general ordering theory, philosophical foundations of knowledge and its artifacts, theoretical bases of classification, data analysis and reduction); (2) describe practical operations associated with indexing and classification, as well as applications of classification systems and thesauri, manual and machine indexing; (3) trace the history of knowledge organization; (4) discuss questions of education and training in classification; and (5) problems of terminology in general and with respect to special fields.



*Task* Find and list some international and national knowledge organizations.

## Self Assessment

Multiple Choice Questions:

5. The society publishes the quarterly academic journal knowledge organization, and it holds an international conference every:
 

|                 |                |
|-----------------|----------------|
| (a) one year    | (b) two years  |
| (c) three years | (d) four years |
6. Knowledge organization was formerly known as international classification until:
 

|          |          |
|----------|----------|
| (a) 1963 | (b) 1973 |
| (c) 1983 | (d) 1993 |

## 3.3 Classification Research Group

- It can reasonably be said that Brian Vickery was responsible for the creation of the CRG
- It finds its origins in the Royal Society Conference on Scientific Information of 1948, which Vickery attended
- Concern for the management of scientific information, its dissemination and retrieval, was one of the major themes of the conference
- This led to the setting up of a 'classification committee' under the leadership of Bernal
- After an unproductive period Bernal invited Vickery and Wells to convene a specialist group.

### Constitution of the CRG:

- the group was composed of librarians and information scientists – some of the leading names of that and the subsequent period
- there was a mix of academics, researchers, and practitioners
- figures such as Austin, Coates, Farradane, Fairthorne, Foskett, Kyle, Langridge, Mills, Palmer, and Wells, as well as Vickery, contributed to its work.

Notes

**Publications of the CRG:**

- there are a number of bibliographic and bibliometric studies of the CRG
- joint publications of the Group are relatively few
- regular (although not frequent) *Bulletins* were published in the *Journal of Documentation*
- three of these contained bibliographies of members' publications
- Vickery is by far the most prolific author, as he continued to be throughout his life.

**The new general classification scheme:**

- in the late 1950s and 1960s the main focus of CRG work was the proposed new general scheme of classification
- several papers were written, and a conference held, on this topic
- a grant from NATO subsidised the original work which never produced a classification, but did result in the PRECIS indexing system
- throughout the 1970s and later the objective was pursued through the revision of Bliss's *Bibliographic Classification*.

**Divergence of classification and IR:**

- during the 1960s 'classification' and 'information retrieval' begin to develop as distinct and separate fields
- this happens in both the US and UK
- it's at this time that Vickery ceases to contribute to the CRG's activities and his name disappears from the record.

**Factors in the 'split':**

- there are distinct 'library' and 'information science' groupings within the CRG
- bibliometric analysis of the CRG publications show quite clear associations of scholars
- at some stage the ideas of 'classification' and 'information retrieval' were uncoupled
- the main group continue with work on a classification scheme and on classification for organization
- Vickery's agenda is somewhat different and he turns in other directions.

**Faceted classification today:**

- over the last twenty years facet analysis has become increasingly important as a methodological approach for all kinds of organizational and retrieval tools
- it features in classification, subject heading lists, thesauri, search interfaces, in taxonomies, ontologies, and semantic web applications
- a number of researchers are attempting to model faceted structures using representation languages and mathematical logic
- it looks as if 'classification' and 'information retrieval' have been reconciled and re-united.

**An evaluation of Vickery's contribution:**

- a driving force in uniting and stimulating the study and understanding of classification
- two intellectual achievements:

- the development of Ranganathan's ideas into a practical tool for scientific libraries
- clarification of the role of classification within the wider activity of information retrieval.

### Cutter Expansive Classification

The Cutter Expansive Classification system is a library classification system devised by Charles Ammi Cutter. It uses all letters to designate the top categories of books. This is in contrast to the Dewey Decimal Classification, which uses only numbers, and the Library of Congress classification, which uses a mixture of letters and numbers. The system was the basis for the top categories of the Library of Congress classification.

"No one, perhaps, can remember it all; it cannot be learned, even in part, very quickly; but those who use the library much will find that they become familiar in time unconsciously with all that they have much occasion to use." from *How to Get Books* by C. A. Cutter, 1882.

### Cutter Numbers

One of the features adopted by other systems, including Library of Congress, is the Cutter number. It is an alphanumeric device to code text so that it can be arranged in alphabetical order using the fewest characters. It contains one or two initial letters and Arabic numbers, treated as a decimal. To construct a Cutter number, a cataloguer consults a Cutter table as required by the classification rules. Although Cutter numbers are mostly used for coding the names of authors, the system can be used for titles, subjects, geographic areas, and more.

### Nippon Decimal Classification

The Nippon Decimal Classification (NDC, also called the Nippon Decimal System) is a system of library classification developed for mainly Chinese and Japanese language books maintained by the Japan Library Association since 1956. It is based on the Dewey Decimal System. The system is based upon using each successive digit to divide into nine divisions with the digit zero used for those not belonging to any of the divisions.

#### Main classes

The system is made up of ten categories:

000 General

100 Philosophy

200 History

300 Social sciences

400 Natural sciences

500 Technology and engineering

600 Industry and commerce

700 Arts

800 Language

900 Literature

#### Description of the classes

000 General

010 Libraries, Library & information science



|              |  |
|--------------|--|
| <b>Notes</b> | 020 Books, Bibliography                                |
|              | 030 Encyclopaedias                                     |
|              | 040 General collected essays                           |
|              | 050 General serial publications                        |
|              | 060 Organizations                                      |
|              | 070 Journalism, Newspapers                             |
|              | 080 General collections                                |
|              | 090 Rare books, Local collections, Special collections |
|              | 100 Philosophy   |
|              | 110 Special treatises on philosophy                    |
|              | 120 Oriental philosophy                                |
|              | 130 Western philosophy                                 |
|              | 140 Psychology   |
|              | 150 Ethics & morals                                    |
|              | 160 Religion   |
|              | 170 Shintoism  |
|              | 180 Buddhism   |
|              | 190 Christianity                                       |
|              | 200 History  |
|              | 210 History of Japan                                   |
|              | 220 History of Asia and the Orient                     |
|              | 230 History of Europe and the West                     |
|              | 240 History of Africa                                  |
|              | 250 History of North America                           |
|              | 260 History of South America                           |
|              | 270 History of Oceania & Polar regions                 |
|              | 280 Biography  |
|              | 290 Geography, Topography, Travel                      |
|              | 300 Social Sciences                                    |
|              | 310 Politics   |
|              | 320 Law  |
|              | 330 Economics  |
|              | 340 Finance  |
|              | 350 Statistics   |
|              | 360 Sociology  |
|              | 370 Education  |
|              | 380 Customs, Folklore, Ethnology                       |
|              | 390 National defence, Military science                 |

- 400 Natural Sciences
- 410 Mathematics
- 420 Physics
- 430 Chemistry
- 440 Astronomy, Space science
- 450 Earth science
- 460 Biology
- 470 Botany
- 480 Zoology
- 490 Medicine, Pharmacology
  
- 500 Technology & Engineering
- 510 Construction, Civil engineering
- 520 Architecture
- 530 Mechanical engineering, Nuclear engineering
- 540 Electrical & Electronic engineering
- 550 Maritime & Naval engineering
- 560 Metal & Mining engineering
- 570 Chemical technology
- 580 Manufacturing
- 590 Domestic arts and sciences
  
- 600 Industry and Commerce
- 610 Agriculture
- 620 Horticulture
- 630 Silk industry
- 640 Animal husbandry
- 650 Forestry
- 660 Fishing
- 670 Commerce
- 680 Transportation & Traffic
- 690 Communications
  
- 700 Arts
- 710 Plastic arts (sculpture)
- 720 Painting & Calligraphy
- 730 Engraving
- 740 Photography & Printing
- 750 Craft
- 760 Music & Dance
- 770 Theatre, Motion Pictures

- Notes**
- 780 Sports, Physical Education
  - 790 Recreation, Amusements
  
  - 800 Language
  - 810 Japanese
  - 820 Chinese, other oriental languages
  - 830 English
  - 840 German
  - 850 French
  - 860 Spanish
  - 870 Italian
  - 880 Russian
  - 890 Other languages
  
  - 900 Literature
  - 910 Japanese literature
  - 920 Chinese literature, Other Oriental literature
  - 930 English & American literature
  - 940 German literature
  - 950 French literature
  - 960 Spanish literature
  - 970 Italian literature
  - 980 Russian & Soviet literature
  - 990 Other language literature

#### **Chinese Library Classification**

The Chinese Library Classification also known as Classification for Chinese Libraries (CCL), is effectively the national library classification scheme in China. It is used in almost all primary and secondary schools, universities, academic institutions, as well as public libraries. It is also used by publishers to classify all books published in China.

The Book Classification of Chinese Libraries (BCCL) was first published in 1975, under the auspices of China's Administrative Bureau of Cultural Affairs. Its fourth edition (1999) was renamed CLC. In September 2010, the fifth edition was published by National Library of China Publishing House. CLC has twenty-two top-level categories, and inherits a Marxist orientation from its earlier editions. (For instance, category A is Marxism, Leninism, Maoism & Deng Xiaoping Theory.) It contains a total of 43600 categories, many of which are recent additions, meeting the needs of a rapidly changing nation.

#### **The CLC System**

The 22 top categories and selected sub-categories of CLC (5th Edition) are as follows:

- A. Marxism, Leninism, Maoism & Deng Xiaoping Theory
  - A1 The Works of Karl Marx and Friedrich Engels
  - A2 The Works of Vladimir Lenin
  - A3 The Works of Joseph Stalin

- A4 The Works of Mao Zedong
- A49 The works of Deng Xiaoping
- A5 The Symposium/Collection of Marx, Engels, Lenin, Stalin, Mao and Deng Xiaoping
- A7 The biobibliography and biography of Marx, Engels, Lenin, Stalin, Mao and Deng Xiaoping
- A8 Study and Research of Marxism, Leninism, Maoism & Deng Xiaoping Theory
- B. Philosophy and Religion
- B0 Philosophical schools
- B1 Philosophy (Worldwide)
- B2 Philosophy in China
- B22 Pre-Qin Dynasty Philosophy (~before 220 BC)
- B222 The Confucian School
- B222.2 Confucius (Kong Qi, 551-479 BC)
- B3 Philosophy in Asia
- B4 Philosophy in Africa
- B5 Philosophy in Europe
- B6 Philosophy in Australasia
- B7 Philosophy in America
- B8 Cognitive science
- B9 Religion
- B91 Sociology of Religion, Religion and Science
- B92 Philosophy and History of Religion
- B93 Mythology and Primitive religion
- B94 Buddhism
- B95 Taoism
- B96 Islam
- B97 Christianity
- B971 Bible
- B971.1 Old Testament
- B971.2 New Testament
- B972 Doctrine, Theology
- B975 Evangelism, Sermon
- B976 Christian Denomination
- B976.1 Roman Catholic Church
- B976.2 Orthodox Christianity (Eastern Orthodoxy, Oriental Orthodoxy)
- B976.3 Protestantism (Protestant Reformation)
- B977 Ecclesiastical polity
- B978 Research on Christianity
- B979 History of Christianity
- B979.9 Biography

- Notes**
- B98 Other Religions
  - B99 Augury, Superstition
  - C. Social Sciences
    - C0 Social Scientific Theory and Methodology
    - C1 Present and Future of Social Sciences
    - C2 Organisations, Groups, Conferences
    - C3 Method of Research in Social Sciences
    - C4 Education and Popularization of Social Sciences
    - C5 Serials, Anthologies, Periodicals in Social Sciences
    - C6 Reference Materials in Social Sciences
    - C7 (no longer used)
    - C8 Statistics in Social Sciences
    - C9 Sociology
  - D. Politics and Law
    - D0 Political theory
    - D1 International Campaign of Communism
    - D2 Communist Party of China
    - D3 Communist Parties of other Countries
    - D4 Labor, Peasant, Youth, Female Organizations and Movements
    - D5 Politics (worldwide)
    - D6 Politics in China
    - D7 Politics in individual Countries
    - D8 Diplomacy, International relations
    - D9 Law
  - E. Military Science
    - E0 Military Theory
    - E1 Military (worldwide)
    - E2 Military in China
    - E3 Military in Asia
    - E4 Military in Africa
    - E5 Military in Europe
    - E6 Military in Australasia
    - E7 Military in America
    - E8 Strategies, Tactics, and Battles
    - E9 Military Technology
  - F. Economics
    - F0 Economics
    - F1 Economics, Economic history and Economic geography of individual countries
    - F2 Economic Planning and Management

F3 Agricultural Economics  
F4 Industrial Economics  
F5 Economics of Transport  
F6 Economics of Postal and Cable Services  
F7 Economics of Commerce  
F8 Finance, Banking  
G. Culture, Science, Education and Sports  
G0 Philosophy of Culture  
G1 Culture  
G2 Knowledge transmission  
G3 Science, Scientific Research  
G4 Education  
G5 Education in individual Countries  
G6 Education (Primary, Secondary, Tertiary)  
G7 Education (specialized)  
G8 Sports  
H. Languages and Linguistics  
H0 Linguistics  
H1 Chinese language  
H2 Languages of China's ethnic minorities  
H3 Commonly Used Foreign Languages  
H31 English language  
H32 French language  
H33 German language  
H34 Spanish language  
H35 Russian language  
H36 Japanese language  
H37 Arabic language  
H4 Family of Sino-Tibetan languages (China, Tibet and Burma)  
H5 Family of Altaic languages (Turkic, Mongolian and Tungusic)  
H6 Language families in other areas of the World  
H61 Austro-Asiatic languages and Tai languages (Mainland Southeast Asia))  
H62 Dravidian languages (South India)  
H63 Austronesian languages (Malayo-Polynesian)  
H64 Paleosiberian languages (Siberia)  
H65 Ibero-Caucasian languages (Caucasus Mountains)  
H66 Uralic languages  
H67 Afroasiatic languages (Southwest Asia, Arabian Peninsula, North Africa)  
H7 Indo-European languages

- Notes**
- H8 Language families on other Continents
  - H81 African languages
  - H83 American languages
  - H84 Papuan languages
  - H9 International Auxiliary Languages (Interlingua, Ido, Esperanto, etc.)
  - I. Literature
  - I0 Literary Theory
  - I1 Literature (worldwide)
  - I2 Literature in China
  - I3 Literature in Asia
  - I4 Literature in Africa
  - I5 Literature in Europe
  - I6 Literature in Australasia
  - I7 Literature in America
  - J. Art
  - J0 Theory of Fine Art
  - J1 Fine Art of the World
  - J2 Painting
  - J3 Sculpture
  - J4 Photography
  - J5 Applied arts
  - J6 Music
  - J7 Dance
  - J8 Drama
  - J9 Cinematography, Television
  - K. History and Geography
  - K0 Historical Theory
  - K1 History of the World
  - K2 History of China
  - K3 History of Asia
  - K4 History of Africa
  - K5 History of Europe
  - K6 History of Australasia
  - K7 History of America
  - K8 Biography, Archaeology
  - K9 Geography
  - N. Natural Science
  - N0 Theory and Methodology
  - N1 Present state

N2 Organisations, Groups, Conferences  
N3 Research Methodology  
N4 Education and Popularization  
N5 Serials, Anthologies, Periodicals  
N6 Reference Materials  
N8 Field Surveys  
N9 Minor Sciences  
O. Mathematics, Physics and Chemistry  
O1 Mathematics  
O2 Applied Mathematics  
O3 Mechanics  
O4 Physics  
O6 Chemistry  
O7 Crystallography  
P. Astronomy and Geoscience  
P1 Astronomy  
P2 Geodesy  
P3 Geophysics  
P4 Meteorology  
P5 Geology  
P6 Mineralogy  
P7 Oceanography  
P9 Physiography  
Q. Life Sciences  
Q1 General Biology  
Q2 Cytology  
Q3 Genetics  
Q4 Physiology  
Q5 Biochemistry  
Q6 Biophysics  
Q7 Molecular Biology  
Q8 Bioengineering  
Q9 Zoology and Botany  
R. Medicine and Health Sciences  
R1 Preventive Medicine, Public health  
R2 Traditional Chinese Medicine  
R3 Human anatomy, Physiology, Pathology, Microbiology, Parasitology  
R4 Clinical Medicine  
R5 Internal medicine  
R6 Surgery



Notes

R7 Medical Specialties  
R71 Obstetrics, Gynecology  
R72 Pediatrics  
R73 Oncology  
R74 Neurology, Psychiatry  
R75 Dermatology, Venereology  
R76 Otolaryngology  
R77 Ophthalmology  
R78 Dentistry  
R79 Non-Chinese Traditional Medicine  
R8 Radiology, Sport medicine, Diving medicine, Aerospace medicine  
R9 Pharmacology, Pharmacy  
S. Agricultural Science  
S1 Fundamental Agricultural Science  
S2 Agricultural Engineering  
S3 Agronomy  
S4 Phytopathology  
S5 Individual Crops  
S6 Horticulture  
S7 Forestry  
S8 Animal Husbandry, Veterinary medicine, Hunting, Sericulture, Apiculture  
S9 Aquaculture, Fishery  
T. Industrial Technology  
TB General Industrial Technology  
TD Mining Engineering  
TE Petroleum, Natural Gas  
TF Extractive metallurgy, Smelting  
TG Metallurgy, Metalworking  
TH Machinery, Instrumentation  
TJ Military Technology  
TK Power Plant  
TL Nuclear technology  
TM Electrical Engineering  
TN Electronic Engineering, Telecommunication Engineering  
TP Automation, Computer Engineering  
TQ Chemical Engineering  
TS Light Industry, Handicraft  
TU Construction Engineering  
TV Water Resources, Hydraulic Engineering  
U. Transportation

- U1 General Transport
- U2 Railway Transport
- U4 Highway Transport
- U6 Marine Transport
- V. Aviation and Aerospace
- V1 Research and Exploration of Aviation and Aerospace Technology
- V2 Aviation
- V4 Aerospace (Spaceflight)
- X. Environmental Science
- X1 Fundamental Environmental Science
- X2 Environmental Research
- X3 Environmental Protection and Management
- X4 Disaster Protection
- X5 Pollution Control
- X7 Waste Management and Recycling
- X8 Environmental Quality Monitoring
- X9 Occupational safety and health
- Z. General, Miscellaneous, Auxiliary and Others
- Z1 Series
- Z2 Encyclopedia
- Z3 Dictionary
- Z4 Symposium, Anthologies, Selected Works, Essay
- Z5 Almanac
- Z6 Serial, Periodicals
- Z8 Catalogue, Abstract, Index

**Korean Decimal Classification**

The Korean Decimal Classification (KDC) is a system of library classification used in Korea. The main classes are the same as in the Dewey Decimal Classification but these are in a different order: Natural sciences 400; Technology and engineering 500; Arts 600; Language 700.

**Main classes**

- 000 General
- 100 Philosophy
- 200 Religion
- 300 Social sciences
- 400 Natural sciences
- 500 Technology and engineering
- 600 Arts
- 700 Language
- 800 Literature
- 900 History

Notes

### British Classification Society

The British Classification Society exists to encourage the co-operation and exchange of views and information among those interested in principles and practice of classification in any discipline where they are used. Its membership includes anthropologists, archaeologists, astronomers, biologists, chemists, computer scientists, forensic scientists, geologists, information specialists, librarians, psychologists, soil scientists and statisticians. The Society organises meetings, some by itself, but often jointly with societies representing application areas for classification.

### 3.4 Summary

- Documentation Research and Training Centre (DRTC) is a research centre (and a department) for library and information science and allied disciplines at the Indian Statistical Institute, Bangalore.
- DRTC was established in January 1962 .
- DRTC is widely considered to be the best research centre in India in the fields of library science and information science.
- The International Society for Knowledge Organization (also referred to by its abbreviation ISKO) is the principal professional association for scholars.
- ISKO “promotes research, development and applications of knowledge organization systems that advance the philosophical, psychological and semantic approaches for enhancing knowledge.
- Brian Vickery was responsible for the creation of the CRG.
- CRG group was composed of librarians and information scientists – some of the leading names of that and the subsequent period.
- In the late 1950s and 1960s the main focus of CRG work was the proposed new general scheme of classification.
- It features in classification, subject heading lists, thesauri, search interfaces, in taxonomies, ontologies, and semantic web application.

### 3.5 Keywords

|   |   |
|---|---|
| <i>DRTC</i>                                   | : Documentation Research and Training Centre—was established in January 1962. |
| <i>Cutter Expansive Classification System</i> | : Is a library classification devised by Charles Ammi Cutter.                 |
| <i>Nippon Decimal Classification</i>          | : Is based on Dewey Decimal system.   |
| <i>Chinese Library Classification</i>         | : Is the library classification used in China.                                |
| <i>Korean Decimal Classification</i>          | : Is the library classification used in Korea.                                |

### 3.6 Review Questions

1. What is the full form of DRTC and when it was established?
2. Write the function of DRTC.

3. Abbreviate ISKO.
4. What promotes ISKO?
5. What does CRG mean?
6. Who created CRG?
7. Who constituted CRG?

### Answers: Self Assessment

1. S.R. Ranganathan
2. National documentation centre
3. September 1951
4. Library science, information science
5. (b)                      6. (d)

### 3.7 Further Readings



#### Books

Deegan, M. and Simon Tanner. *Digital futures*. London. LA, 2002.

Maltby, A., ed. *Sayer's manual of classification for libraries*. 5th. Ed. London: Andre Deutsch, 1975.

Oddy, P. *Future libraries, future catalogs*. London: LA, 1996.



#### Online links

<http://hurights.pbworks.com/w/page/11947504/India%20Centers>

<http://www.ib.hu-berlin.de/~wumsta/iskobroc.html#1>

## Unit 4: Cataloguing–Development and Trends

### CONTENTS

Objectives

Introduction

4.1 International Standard Bibliographic Description

4.2 Summary

4.3 Keywords

4.4 Review Questions

4.5 Further Readings

### Objectives

After studying this unit, you will be able to:

- Describe international standard bibliographic description
- Define structure of an ISBD record.

### Introduction

Cataloguing is the process of listing or includes something in a Catalogue. In library science is the producing of bibliographical descriptions of books or other kinds of documents. Today the study of Cataloguing has broaden and merged with the study of metadata (“data about data contents”) and is sometimes termed resource description and access.

The International Standard Bibliographic Description (ISBD) is intended to serve as a principal standard to promote universal bibliographic control, that is, to make universally and promptly available, in a form that is internationally acceptable, basic bibliographic data for all published resources in all countries.



*Task* Explain the common cataloguing methods used in national level in India.

### 4.1 International Standard Bibliographic Description

The main goal of the ISBD has been since the beginning, to provide consistency when sharing bibliographic information. The ISBD is the standard that determines the data elements to be recorded or transcribed in a specific sequence as the basis of the description of the resource being catalogued. In addition, it employs prescribed punctuation as a means of recognizing and displaying data elements and making them understandable independently of the language of the description.

A new Statement of International Cataloguing Principles was published by IFLA in 2009. In these principles, which replace and broaden the Paris Principles of 1961, the fifth section is devoted to bibliographic description where it is stated that “Descriptive data should be based on an internationally agreed standard.” A footnote identifies the ISBD as the standard for the library community, as the statement of principles is intended not only for libraries but also for archives, museums, and other communities.

Although the development of this standard was originally motivated by the automation of bibliographic control as well as by the economic necessity of sharing cataloguing, the ISBD continues to be useful for and applicable to bibliographic descriptions of all kinds of resources in any type of catalogue, whether online or in a form less technologically advanced. Those agencies using national and multinational cataloguing codes could apply this internationally agreed-upon standard conveniently in their catalogues.

**Work on the ISBD has been guided by the following objectives and principles:**

- The ISBD provides consistent stipulations for description of all types of published resources, to the extent that uniformity is possible, and specific stipulations for specific types of resources as required to describe those resources.
- The ISBD provides the stipulations for compatible descriptive cataloguing worldwide in order to aid the international exchange of bibliographic records between national bibliographic agencies and throughout the international library and information community (including producers and publishers).
- The ISBD accommodates different levels of description, including those needed by national bibliographic agencies, national bibliographies, universities and other research collections.
- The descriptive elements needed to identify and select a resource must be specified.
- The set of elements of information rather than the display or use of those elements in a specific automated system provides the focus.
- Cost-effective practices must be considered in developing the stipulations.

The organization of provisions in the present text is to give first the general stipulations that apply to all types of resources, then the specific stipulations that add information required for that specific type of resource or are exceptions to a general rule.

In general, the ISBD is applied to describe manifestations, by means of description of the item in hand as an exemplar of the entire manifestation, in the terminology of Functional Requirements for Bibliographic Records (FRBR).



*Notes* The ISBD applies the Statement of International Cataloguing Principles, which establishes that “A bibliographic description typically should be based on the item as representative of the manifestation”.

In the ISBD, national bibliographic agencies are called upon to prepare definitive descriptions that “contain all the mandatory elements set out in the ISBD insofar as the information is applicable to the resource. This practice is also recommended for application by libraries that share bibliographic data with each other. Inclusion of a data element is considered “mandatory” in all cases for certain elements, and in other cases is considered “mandatory” when necessary for identification of the resource being described or otherwise considered important to users of a bibliography or a catalogue.

Notes



*Did u know?* One of the original purposes of the ISBD was to provide a standard form of bibliographic description that could be used to exchange records internationally. This would support IFLA's program of universal bibliographic control.

### Structure of an ISBD Record

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

1. title and statement of responsibility area, with the contents of
  - (a) Title proper
  - (b) General material designation
  - (c) Parallel title
  - (d) Other title information
  - (e) Statements of responsibility
2. edition area
3. material or type of resource specific area (for example, the scale of a map or the numbering of a periodical)
4. publication, production, distribution, etc., area
5. physical description area (for example: number of pages in a book or number of CDs issued as a unit)
6. series area
7. notes area
8. resource identifier (*e.g.* ISBN, ISSN) and terms of availability area

ISBD(A) is governing the antiquarian bibliographic publications, which could apply to the ones in archaeology, museum, antique auction or canonical texts etc.

### Self Assessment

Multiple Choice Questions:

1. A new statement of international cataloguing principles was published by IFLA in:
  - (a) 2019
  - (b) 2009
  - (c) 2029
  - (d) 2008
2. The ISBD prescribes ..... areas of description.
  - (a) Five
  - (b) Six
  - (c) Seven
  - (d) Eight

Fill in the blanks:

3. The main goal of the ISBD has been since the beginning, to provide consistency when sharing .....
4. .... and ..... that do not apply to a particular resource are omitted from the description.

## 4.2 Summary

Notes

- The International Standard Bibliographic Description (ISBD) is intended to serve as a principal standard to promote universal bibliographic control.
- The ISBD is the standard that determines the data elements to be recorded or transcribed in a specific sequence as the basis of the description of the resource being catalogued.
- The ISBD provides consistent stipulations for description of all types of published resources, to the extent that uniformity is possible, and specific stipulations for specific types of resources as required to describe those resources.

## 4.3 Keywords

*Volume(s)* : Some works comprise several bound books.

*Page(s)* : Number of arabic-numeral-numbered pages that make up the main body of the book.

## 4.4 Review Questions

1. What is the objective of the ISBD?
2. Mention the key function of ISBD.
3. Describe the structure of an ISBD record.

## Answers: Self Assessment

1. (b)
2. (d)
3. Bibliographic information
4. Elements, areas.

## 4.5 Further Readings



Books

Best, DP, Ed. *The fourth resource: information and its management*. Aldershot: Aslib, 1996.

Cooke, A : *A guide to finding quality Information on the Internet*. 2nd Edition. London: Facet Publishing, 2001.

Deegan, M. and Simon Tanner. *Digital futures*. London. LA, 2002.



Online links

<http://bcppwww.csu.edu.au/faculty/educat/sis/CIS/3365/>  
[www.wikipedia.com](http://www.wikipedia.com).



## Unit 5: MACHine-Readable Cataloguing and Online

### CONTENTS

Objectives

Introduction

- 5.1 MACHine-Readable Cataloguing
- 5.2 Common Communication Format
- 5.3 History of Online Public Access Catalogue
- 5.4 Summary
- 5.5 Keywords
- 5.6 Review Questions
- 5.7 Further Readings

### Objectives

After studying this unit, you will be able to:

- Describe machine-readable cataloguing
- Define common communication format
- Discuss about the history of online public access catalogue.

### Introduction

MARC is an acronym, used in the field of library science that stands for MACHine-Readable Cataloguing. MACHine-readable means that one particular type of machine, a computer, can read and interpret the data in the cataloguing record and thus, provide online a means of discovering information. Many words have been written on the subject of the MACHine-Readable Cataloguing (MARC) Program: the events that led to the pilot project, the development of the format, the operational Distribution Service, the influence of MARC on standardization, and the impetus it gave to library automation projects and to the creation of networks here and abroad.

An Online Public Access Catalogue (often abbreviated as OPAC or simply Library Catalogue) is an online database of materials held by a library or group of libraries. Users search a library catalogue principally to locate books and other material physically located at a library.

### 5.1 MACHine-Readable Cataloguing

The MARC standards consist of the MARC formats, which are standards for the representation and communication of bibliographic and related information in machine-readable form, and related documentation. It defines a bibliographic data format that was developed by Henriette Avram at

the Library of Congress beginning in the 1960s. It provides the protocol by which computers exchange, use, and interpret bibliographic information. Its data elements make up the foundation of most library catalogues used today.

Notes



*Did u know?* The record structure of MARC is an implementation of ISO 2709, also known as ANSI/NISO Z39.2.

MARC records are composed of three elements: the record structure, the content designation, and the data content of the record. The record structure implements national and international standards (e.g., Z39.2, ISO2709). The content designation is “the codes and conventions established to identify explicitly and characterize ...data elements within a record” and support their manipulation. The content of data elements in MARC records is defined by standards outside the formats such as AACR2, L.C. Subject Headings, and MeSH.



*Notes* The future of the MARC formats is a matter of some debate in the worldwide library science community.

On the one hand, the storage formats are quite complex and are based on outdated technology. On the other, there is no alternative bibliographic format with an equivalent degree of granularity. The huge user base, billions of records in tens of thousands of individual libraries (including over 50,000,000 belonging to the OCLC consortium alone), also creates inertia.

## MARC Formats

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

## MARC 21

MARC 21 is a result of the combination of the United States and Canadian MARC formats (USMARC and CAN/MARC). MARC 21 is based on the ANSI standard Z39.2, which allows users of different software products to communicate with each other and to exchange data.


**Notes**

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

MARC 21 allows the use of two character sets, either MARC-8 or Unicode encoded as UTF-8. MARC-8 is based on ISO 2022 and allows the use of Hebrew, Cyrillic, Arabic, Greek, and East Asian scripts. MARC 21 in UTF-8 format allows all the languages supported by Unicode.

**MARC XML**

MARC XML is an XML schema based on the fairly common MARC 21 standards.



*Did u know?* MARC XML was developed by the US Library of Congress and adopted by it and others as a means of easy sharing of, and networked access to, bibliographic information.

Being easy to parse by various systems allows it to be used as an aggregation format, as it is in software packages such as MetaLib, though that package merges it into a wider DTD specification.

The MARC XML primary design goals included:

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

**Self Assessment**

Fill in the blanks:

1. MARC is an acronym, used in the field of library science that stands for ..... .
2. MARC defines a bibliographic data format that was developed by ..... at the library of congress beginning in the 1960's.
3. MARC records are composed of three elements: ..... , ..... and ..... .
4. MARC 21 has formats for the following five types of data: Bibliographic Format, Authority Format, Holdings Format, ..... and ..... .
5. MARC 21 in ..... format allows all the languages supported by unicode.

**5.2 Common Communication Format**

The Unesco Common Communication Format (CCF) is described in the context of other exchange formats. A definition is given of 'exchange format', and the CCF is compared against this definition. The history of its development is outlined and its major technical features are summarized. Examples

are given of the ways in which it is being used and is likely to be used in the future, and a number of implementation manuals are mentioned which have been developed to assist in its use.

### **5.3 History of Online Public Access Catalogue**

#### **Early online catalogues**

Although a handful of experimental systems existed as early as the 1960s, the first large-scale online catalogues were developed at Ohio State University in 1975 and the Dallas Public Library in 1978.

These and other early online catalogue systems tended to closely reflect the card catalogues that they were intended to replace. Using a dedicated terminal or telnet client, users could search a handful of pre-coordinate indexes and browse the resulting display in much the same way they had previously navigated the card catalogue.

Throughout the 1980s, the number and sophistication of online catalogues grew. The first commercial systems appeared, and would by the end of the decade largely replace systems built by libraries themselves. Library catalogues began providing improved search mechanisms, including Boolean and keyword searching, as well as ancillary functions, such as the ability to place holds on items that had been checked-out.

At the same time, libraries began to develop applications to automate the purchase, cataloguing, and circulation of books and other library materials. These applications, collectively known as an integrated library system (ILS) or library management system, included an online catalogue as the public interface to the system's inventory. Most library catalogues are closely tied to their underlying ILS system.

#### **Stagnation and dissatisfaction**

The 1990s saw a relative stagnation in the development of online catalogues. Although the earlier character-based interfaces were replaced with ones for the web, both the design and the underlying search technology of most systems did not advance much beyond that developed in the late 1980s.

Prior to the widespread use of the Internet, the online catalogue was often the first information retrieval system library users ever encountered. Now accustomed to web search engines, newer generations of library users have grown increasingly dissatisfied with the complex (and often arcane) search mechanisms of older online catalogue systems.

This has, in turn, led to vocal criticisms of these systems within the library community itself, and in recent years to the development of newer (often termed 'next-generation') catalogues.

#### **Next-generation catalogues**

The newest generation of library catalogue systems are distinguished from earlier OPACs by their use of more sophisticated search technologies, including relevancy ranking and faceted search, as well as features aimed at greater user interaction and participation with the system, including tagging and reviews.

These newer systems are almost always independent of the library's integrated library system, instead providing drivers that allow for the synchronization of data between the two systems. While older online catalogue systems were almost exclusively built by ILS vendors, libraries are increasingly turning to next generation catalogue systems built by enterprise search companies and open source projects, often led by libraries themselves. The costs associated with these new systems, however, have slowed their adoption, particularly at smaller institutions.

Notes

**Union catalogues**

Although library catalogues typically reflect the holdings of a single library, they can also contain the holdings of a group or consortium of libraries. These systems, known as union catalogues, are usually designed to aid the borrowing of books and other materials among the member institutions via interlibrary loan. The largest such union catalogue is WorldCat, which includes the holdings of over 70,000 libraries worldwide.

**Related systems**

There are a number of systems that share much in common with library catalogues, but have traditionally been distinguished from them. Libraries utilize these systems to search for items not traditionally covered by a library catalogue.

These include bibliographic databases—such as Medline, ERIC, PsycINFO, and many others—which index journal articles and other research data. There are also a number of applications aimed at managing documents, photographs, and other digitized or born-digital items. Particularly in academic libraries, these systems (often known as digital library systems or institutional repository systems) assist with efforts to preserve documents created by faculty and students.



*Task* Make a report on early online catalogue.

**Self Assessment**

Multiple Choice Questions:

6. The first largescale online catalogues were developed at ohio state university in ..... .  
(a) 1974 (b) 1975  
(c) 1978 (d) 1985
7. The ..... saw a relative stagnation in the development of online catalogues.  
(a) 1979s (b) 1980s  
(c) 1990s (d) 1995s
8. The largest such union catalogue is worldcat, which includes the holdings of over ..... libraries worldwide.  
(a) 40,000 (b) 50,000  
(c) 60,000 (d) 70,000

**5.4 Summary**

- MARC is an acronym, used in the field of library science that stands for MACHine-Readable Cataloguing.
- The Unesco Common Communication Format (CCF) is described in the context of other exchange formats.
- An Online Public Access Catalogue (often abbreviated as OPAC or simply Library Catalogue) is an online database of materials held by a library or group of libraries.
- The newest generation of library catalogue systems are distinguished from earlier OPACs by their use.
- MARC is an acronym, used in the field of library science that stands for MACHine-Readable Cataloguing.

## 5.5 Keywords

**MARC** : MACHINE—Readable Cataloguing and Online—is an acronym—used in the field of library science.

**CCF** : Common Communication Format, is described in the context of other exchange formats.

## 5.6 Review Questions

1. Write the function of MARC.
2. What do you mean by CCF?
3. Give a brief history of online public access catalogue.

## Answers: Self Assessment

1. MACHINE-readable cataloguing
2. Henriette Avram
3. The record structure, the content designation, the data content of the record.
4. Community format, classification data format.
5. UTF-8.
6. (b)                      7. (c)                      8. (d)

## 5.7 Further Readings



### Books

Aitchinson, J and Gilchrist, A: *Thesaurus construction*. 2<sup>nd</sup> ed. London: Aslib, 1987.

Best, DP, Ed. *The fourth resource: information and its management*. Aldershot: Aslib, 1996.

Chowdhary, GG: *Introduction to Modern Information Retrieval*. London: LA, 1999.



### Online links

<http://unesdoc.unesco.org/images/0008/000806/080626eb.pdf>

<http://searchsqlserver.techtarget.com/definition/OPAC>

<http://lib.kedah.uitm.edu.my/pslibrary/notes/is110/layout/notes/>

## Unit 6: Cataloguing

### CONTENTS

Objectives

Introduction

6.1 Cataloguing

6.2 History

6.3 Summary

6.4 Keywords

6.5 Review Questions

6.6 Further Readings

### Objectives

After studying this unit, you will be able to:

- Explain cataloguing
- Explain brief history of cataloguing.

### Introduction

The process of creating entire for a catalogue. In libraries, this usually includes bibliographic description, subject analysis, assignment of classification notation, and activities involved in physically preparing the item for the shelf, tasks usually performed under the supervision of a librarian trained as a cataloguer. A library catalogue consisting of a collection of bibliographic records in machine-readable format, maintained on a dedicated computer that provides uninterrupted interactive access via terminals or workstations in direct, continuous communication with the central computer. Although the software used in online catalogues is proprietary and not standardized, most online catalogues are searchable by author, title, subject heading, and keywords and most public and academic libraries in the United States provide free public access, usually through a Web-based graphical user interface.

Library catalogues have a very extensive history, and can be traced back to the libraries of Antiquity. In the 7th century B.C., important libraries in Mesopotamia had author and title catalogues that were posted on walls for user convenience. Callimachus, scholar and chief librarian of the Alexandrian Library in the 3rd century B.C. compiled a huge catalogue of the works contained there, called the Pinakes. This work later became the foundation for the analytical analysis of Greek Literature.

Catalogues have changed dramatically over the centuries, having appeared in many forms, from clay tablets, papyrus scrolls, printed books and cards, microform, to the online versions that are prevalent today.

## 6.1 Cataloguing

Notes

A library catalogue (or library catalogue) is a register of all bibliographic items found in a library or group of libraries, such as a network of libraries at several locations. A bibliographic item can be any information entity (*e.g.*, books, computer files, graphics, realia, cartographic materials, etc.) that is considered library material (*e.g.*, a single novel in an anthology), or a group of library materials (*e.g.*, a trilogy), or linked from the catalogue (*e.g.*, a webpage) as far as it is relevant to the catalogue and to the users (patrons) of the library.



*Notes* The card catalogue was a familiar sight to library users for generations, but it has been effectively replaced by the online public access catalogue (OPAC). Some still refer to the online catalogue as a “card catalogue”.

Some libraries with OPAC access still have card catalogues on site, but these are now strictly a secondary resource and are seldom updated. Many of the libraries that have retained their physical card catalogue post a sign advising the last year that the card catalogue was updated. Some libraries have eliminated their card catalogue in favour of the OPAC for the purpose of saving space for other use, such as additional shelving.



*Task* What is cataloguing and its goal? Explain.

### Goal

Charles Ammi Cutter made the first explicit statement regarding the objectives of a bibliographic system in his *Rules for a Printed Dictionary Catalogue* in 1876. According to Cutter, those objectives were

1. to enable a person to find a book of which either (Identifying objective)
  - the author
  - the title
  - the subject
  - the category
  - is known.
2. to show what the library has (Collocating objective)
  - by a given author
  - on a given subject
  - in a given kind of literature.
3. to assist in the choice of a book (Evaluating objective)
  - as to its edition (bibliographically)
  - as to its character (literary or topical).

These objectives can still be recognized in more modern definitions formulated throughout the 20th century. 1960/61 Cutter's objectives were revised by Lubetzky and the Conference on Cataloguing



**Notes**

Principles (CCP) in Paris. The latest attempt to describe a library catalogue's goals and functions was made in 1998 with Functional Requirements for Bibliographic Records (FRBR) which defines four user tasks: find, identify, select, and obtain.

**Catalogue Card**

Main Entry *e.g.*,

Arif, Abdul Majid.

Political structure in a changing Pakistani

villages/by Abdul Majid and Basharat Hafeez

Andaleeb.—2nd ed.—Lahore : ABC Press, 1985.

xvi, 367p. : ill. ; 22 cm.

Includes index.

ISBN 969-8612-02-8 (hbk.)

**Types**

Traditionally, there are the following types of catalogue:

Author card: a formal catalogue, sorted alphabetically according to the authors' or editors' names of the entries.

Title catalogue: a formal catalogue, sorted alphabetically according to the title of the entries.

Dictionary catalogue: a catalogue in which all entries (author, title, subject, series) are interfiled in a single alphabetical order. This was the primary form of card catalogue in North American libraries just prior to the introduction of the computer-based catalogue.

Keyword catalogue: a subject catalogue, sorted alphabetically according to some system of keywords.

Mixed alphabetic catalogue forms: sometimes, one finds a mixed author/title, or an author/title/keyword catalogue.

Systematic catalogue: a subject catalogue, sorted according to some systematic subdivision of subjects. Also called a Classified catalogue.

Shelf list catalogue: a formal catalogue with entries sorted in the same order as bibliographic items are shelved. This catalogue may also serve as the primary inventory for the library.

**Self Assessment**

State whether the following statements are true or false:

1. 1960/61 cutter's objectives were revised by Lubetzky and the Conference on Cataloguing Principles (CCP) in Paris.
2. Author Card: a formal catalogue, sorted alphabetically according to the title of the entries.
3. Keyword catalogue: a subject catalogue, sorted alphabetically according to some system of keywords.
4. Shelf list catalogue is also called a classified catalogue.
5. A library catalogue is a register of all bibliographic items found in a library and group of libraries.

## 6.2 History

### Notes

Library catalogues originated as manuscript lists, arranged by format (folio, quarto, etc.) or in a rough alphabetical arrangement by author. Printed catalogues, sometimes called dictionary catalogues enabled scholars outside a library to gain an idea of its contents. These would sometimes be interleaved with blank leaves on which additions could be recorded, or bound as guardbooks in which slips of paper were bound in for new entries. Slips could also be kept loose in cardboard or tin boxes, stored on shelves. The first card catalogues appeared in the nineteenth century, enabling much more flexibility, and towards the end of the twentieth century the OPAC was developed.

245 BC : Callimachus is considered the first bibliographer and is the one that organized the library by authors and subjects. The Pinakes was the first ever library catalogue. Variations on this system were used in libraries until the late 1800s when Melvil Dewey developed the Dewey Decimal Classification in 1876, which is still in use today.

800 : Library catalogues are introduced in the House of Wisdom and other medieval Islamic libraries where books are organized into specific genres and categories.

1595 : Nomenclator of Leiden University Library appears, the first printed catalogue of an institutional library.

1674 : Thomas Hyde's catalogue for the Bodleian Library.

More about the early history of library catalogues has been collected in 1956 by Strout.

## Cataloguing Rules

Cataloguing rules have been defined to allow for consistent cataloguing of various library materials across several persons of a cataloguing team and across time. Users can use them to clarify how to find an entry and how to interpret the data in an entry.

Cataloguing rules prescribe -> which information from a bibliographic item is included in the entry; -> how this information is presented on a catalogue card or in a cataloguing record; -> how the entries should be sorted in the catalogue. The larger a collection, the more elaborate cataloguing rules are needed. Users cannot and do not want to examine hundreds of catalogue entries or even dozens of library items to find the one item they need.

Currently, most cataloguing rules are similar to, or even based on, the International Standard Bibliographic Description (ISBD), a set of rules produced by the International Federation of Library Associations and Institutions (IFLA) to describe a wide range of library materials.

These rules organize the bibliographic description of an item in the following areas: title and statement of responsibility (author or editor), edition, material specific details (for example, the scale of a map), publication and distribution, physical description (for example, number of pages), series, notes, and standard number (ISBN). The most commonly used set of cataloguing rules in the English speaking world are the Anglo-American Cataloguing Rules, 2nd Edition, or AACR2 for short. In the German-speaking world there exists the Regeln für die alphabetische Katalogisierung, abbreviated RAK. AACR2 has been translated into many languages, however, for use around the world.



*Did u know?* AACR2 provides rules for descriptive cataloguing only and does not touch upon subject cataloguing.

Library items that are written in a foreign script are, in some cases, transliterated to the script of the catalogue.

Notes

## Cataloguing Terms

Main entry – generally refers to the first author named on the item. Additional authors are added as “added entries.” In cases where no clear author is named, the title of the work is considered the main entry.

## Self Assessment

Multiple Choice Questions:

6. In ..... the pinakes was the first ever library catalogue.  
(a) 225 BC (b) 245 BC  
(c) 265 BC (d) 285 BC
7. In ..... Thomas Hyde’s catalogue for the Bodleian library.  
(a) 1574 (b) 1584  
(c) 1664 (d) 1674
8. Main entry—generally refers to the ..... author named on the item.  
(a) First (b) Second  
(c) Third (d) Fourth

## 6.3 Summary

- A library catalogue (or library catalogue) is a register of all bibliographic items found in a library or group of libraries, such as a network of libraries at several locations.
- Charles Ammi Cutter made the first explicit statement regarding the objectives of a bibliographic system in his Rules for a Printed Dictionary Catalogue in 1876.
- Printed catalogues, sometimes called dictionary catalogues enabled scholars outside a library to gain an idea of its contents.
- Cataloguing rules have been defined to allow for consistent cataloguing of various library materials across several persons of a cataloguing team and across time.

## 6.4 Keywords

*Systematic Catalogue* : A subject catalogue, sorted according to some systematic subdivision of subjects. Also called a Classified catalogue.

*Cataloguing Rules* : Users can use them to clarify how to find an entry and how to interpret the data in the entry.

## 6.5 Review Questions

1. Define library catalogue.
2. Write the cataloguing rules.
3. Explain history of library catalogue.
4. Name the various types of catalogue.

**Answers: Self Assessment**

**Notes**

- |         |          |         |          |
|---------|----------|---------|----------|
| 1. True | 2. False | 3. True | 4. False |
| 5. True | 6. (b)   | 7. (d)  | 8. (a)   |

**6.6 Further Readings**



*Books*

Best, DP, Ed. *The fourth resource: information and its management*. Aldershot: Aslib, 1996.

Aitchinson, J and Gilchrist, A: *Thesaurus construction*. 2<sup>nd</sup> ed. London: Aslib, 1987.

Deegan, M. and Simon Tanner. *Digital futures*. London. LA, 2002.



*Online links*

<http://lib.kedah.uitm.edu.my/psblibrary/notes/is110/layout/notes/>

[http://www.ehow.com/about\\_5449148\\_types-card-catalogs.html](http://www.ehow.com/about_5449148_types-card-catalogs.html)

[http://www.ehow.com/about\\_5098670\\_library-catalogue.html](http://www.ehow.com/about_5098670_library-catalogue.html)

## Unit 7: Sorting and Indexing

### CONTENTS

Objectives

Introduction

- 7.1 Sorting
- 7.2 Online Catalogues
- 7.3 Online Research
- 7.4 Concept Indexing
- 7.5 Summary
- 7.6 Keywords
- 7.7 Review Questions
- 7.8 Further Readings

### Objectives

After studying this unit, you will be able to:

- Define sorting
- Describe online catalogues and online research
- Explain concept indexing.

### Introduction

Indexing and sorting are two approaches for establishing the order of data in a table. You use them to answer different needs in an application. In general, you index a table to establish a specific order of the rows, to help you locate and process information quickly. Indexing makes applications run more efficiently. Use sorting only when you want to create another table with a different natural order of rows.

Indexing orders rows in a specific sequence, usually in ascending or descending order on one field. Indexing creates a list of rows arranged in a logical order, such as by date or by name, and stores this list in a separate file called an *index file*. A dbase index (.MDX) file can have up to 47 indexes, but only one controls the order of rows at any time. The index that is controlling the order is the current master index.

Sorting an operation that segregates items into groups according to specified criterion.

A = {3 1 6 2 1 3 4 5 9 0}

A = {0 1 1 2 3 3 4 5 6 9}

Indexing is not a new activity: it has been practiced since libraries came into existence. Modern information services are in essence the same as traditional libraries. Specific problems and emphases may change, but the fundamental principles of indexing remain the same.

## 7.1 Sorting

In a title catalogue, one can distinguish two sort orders:

In the grammatical sort order (used mainly in older catalogues), the most important word of the title is the first sort term. The importance of a word is measured by grammatical rules; for example, the first noun may be defined to be the most important word.

In the mechanical sort order, the first word of the title is the first sort term. Most new catalogues use this scheme, but still include a trace of the grammatical sort order: they neglect an article (The, A, etc.) at the beginning of the title.

The grammatical sort order has the advantage that often, the most important word of the title is also a good keyword (question 3), and it is the word most users remember first when their memory is incomplete. However, it has the disadvantage that many elaborate grammatical rules are needed, so that only expert users may be able to search the catalogue without help from a librarian.

In some catalogues, person's names are standardized, *i.e.*, the name of the person is always (catalogued and) sorted in a standard form, even if it appears differently in the library material. This standardization is achieved by a process called authority control. An advantage of the authority control is that it is easier to answer question 2 (which works of some author does the library have?). On the other hand, it may be more difficult to answer question 1 (does the library have some specific material?) if the material spells the author in a peculiar variant. For the cataloguer, it may incur (too) much work to check whether Smith, J. is Smith, John or Smith, Jack.

For some works, even the title can be standardized. The technical term for this is uniform title. For example, translations and re-editions are sometimes sorted under their original title. In many catalogues, parts of the Bible are sorted under the standard name of the book(s) they contain.



*Did u know?* The plays of William Shakespeare are another frequently cited example of the role played by a uniform title in the library catalogue.

Many complications about alphabetic sorting of entries arise. Some examples:

Some languages know sorting conventions that differ from the language of the catalogue. For example, some Dutch catalogues sort IJ as Y. Should an English catalogue follow this suit? And should a Dutch catalogue sort non-Dutch words the same way?

## 7.2 Online Catalogues

Online cataloguing has greatly enhanced the usability of catalogues, thanks to the rise of Machine Readable Cataloguing = MARC standards in the 1960s. Rules governing the creation of catalogue MARC records include not only formal cataloguing rules like AACR2 but also special rules specific to MARC, available from the Library of Congress and also OCLC. MARC was originally used to automate the creation of physical catalogue cards; Now the MARC computer files are accessed directly in the search process. OPACs have enhanced usability over traditional card formats because:

The online catalogue does not need to be sorted statically; the user can choose author, title, keyword, or systematic order dynamically.

**Notes**

Most online catalogues offer a search facility for any word of the title; the goal of the grammatical word order (provide an entry on the word that most users would look for) is reached even better.

Many online catalogues allow links between several variants of an author name. So, authors can be found both under the original and the standardised name (if entered properly by the cataloguer).

The elimination of paper cards has made the information more accessible to many people with disabilities, such as the visually impaired, wheelchair users, and those who suffer from mold allergies.

### **Current and Emerging Trends in Cataloguing**

We live in a fast paced digital age. The growing popularity of the Web influences all aspects of our life, has changed the way we live, work, study and even think. As a result the role of library and information professionals is subject to radical changes. Catalogue is the core of every library, a basic tool of retrieval in any document collection. A library without a catalogue cannot fulfill its functions efficiently. The practice of collecting written knowledge in some sort of repository in a certain order is old as civilization itself. It does not lose its importance nowadays when we strive to retrieve some relevant information in the chaos of the net.

Attention to the profession of a cataloguer, which I love, and which is more often, taking into consideration my colleagues, attract people demonstrating such qualities as dedication, creativity, persistency, and enthusiasm. Cataloguers played a key role in organization of recorded knowledge of the human civilization thus making records searchable and retrievable. I am telling these well known facts in view of a modern trend to acquire a wrong attitude to a role of cataloguing profession.

One can come across with reports of the so-called "restructuring" and closure of cataloguing departments. There is a tendency to reduce and eliminate the professionals who catalogue which results in databases full of errors, low quality records, duplicating of records and inconsistencies, and eventually to the "de-professionalisation" of cataloguing. It is important that collections are being developed and maintained by professionals who understand the structure of the information.

Another trend arises, the tendency is that the present generation of cataloguers is retiring or is close to join those "young at heart", which means the loss of professional memory. We need specialists with broad understanding of the principles of cataloguing and bibliographic control. The library users depend of the dedicated and quality work of cataloguers which can save time and frustration while searching for the desired information.

The growth of information technology and computerization add to the need for that quality. In this situation we turn to the library schools, and to our greatest disappointment, find out that cataloguing is not even a core subject in many of them. The concern over the cataloguing training is international. I have come across over a very interesting online survey conducted by Cynthia Boeke (the assistant curator for The American Society for Cell Biology's Image & Video Library).



*Task* "What do you think is the most important issue facing catalogue profession right now, and why it is so important?"

Many participants showed their concern that not enough students are attracted to the subject of cataloguing, as well as the decreasing amount and quality of cataloguing training in library schools nowadays.

### **Career in Cataloguing**

"The lack of professional librarians who want to pursuer a career in cataloguing is the most important issue right now. Why? Without cataloguers, access to the bibliographic resources owned by libraries won't happen." Cynthia Whitecre, Manager, Metadata Quality Dept., OCLC.

During the past ten years, many library schools ceased making 'organization of information' a required course. As result, many library school students haven't had the opportunity to discover cataloguing as possible profession for them. Cataloguing courses that do exist are often inadequate, inconsistent, or too difficult" Billie Hackney, Head of Monograph Cataloguing, Getty Research Institute.

"Cataloguing is less and less represented in courses (at least in France), whilst it should be more and more developed. A student recently said to me: 'The catalogue stands at the core of all library services, why does it not stand at the core of a library curator's formal training?'" Patrick Le Boeuf, curator of the National Library of France.

Recent literature recognizes the challenges faced by library schools today and prompt educators to implement effective teaching strategies and methodologies.

### 7.3 Online Research

Having said this, mention a highly interesting and innovative approach to teaching descriptive cataloguing at the University of Queensland by utilizing various innovative methods and media for teaching activities, where students were encouraged to think critically about broader issues. The results of this pilot project are "beyond expectations" as students clearly indicated that learning about descriptive cataloguing, a good knowledge of AACR and MARC is highly relevant.



*Did u know?* Gillian Hallam "Beyond our expectations: a review of an Independent learning module at the Queensland University of Technology"

The project at the University of Queensland is an evidence that there are ways to attract the students to this challenging and, to my mind, satisfying profession. The profession that has a long and honorable history behind it. Nowadays the ever-expanding growth of information and information technology, increasing volumes and multiple formats of information, changing user expectations and behaviors brought about even higher levels of challenges for cataloguers.

To pursue professional ethics in creating timely and high quality records cataloguers are to develop a new mindset to deal with the increased complexity in cataloguing. New technologies require new skills. The modern cataloguer has to be multi-skilled, computer literate, able to operate different in-house library systems, able to use the online packages, such as MARC 21 standard online, WEB Dewey, LCSH Authorities, LA Search interface, Sanborn Cutter, national and international union online catalogues.

The cultural changes in the society brought about additional requirement. The modern cataloguer has to be multilingual, able to catalogue materials in different languages, including those in non-Roman script, able to insert vernacular statements into the bibliographic record.

Cataloguers have to keep pace with the changing environment, managing materials in new formats, manipulate different metadata schemes, catalogue for diverse user environments and audiences. Print materials do not go away.



*Notes* Materials in traditional format are to be used in combination with new digital and electronic formats.

At the moment we catalogue a variety of electronic resources, among them are CDs, DVDs, CD-ROMs, VCDs, etc. It is important for cataloguers to maintain the quality of cataloguing for effective



Notes

discovery of these materials. Remote access electronic resources are the next challenge to cataloguing which require talented individuals orientated in the electronic area, able to identify the most valuable resources for the on-line catalogue and to create collections of well organized information available in digital form.

Development of new formats require modifications of the classification, bibliographic rules and subjects headings. In the next few years we will have the new addition of the Dewey Decimal classification. We can expect the expansion of class numbers for computer science, philology and literature of languages not represented at present. LC Subject Headings are being frequently reassessed and updated.

Of high importance are the news from the Joint Steering Committee which is preparing a new addition of cataloguing rules for publication in early months of 2009. Recently the committee decided that the new cataloguing code will be called "Resource Description and Access" or RDA.

The most notable changes between RDA and AACR2 will be a statement of cataloguing principles, revised rules of the chapter on the electronic resources, addition of sections of bibliographic relationships, and authority control, incorporating FRBR terminology and concepts. (FRBR = Functional Requirements for Bibliographic Records formulated by the International Federation of Library Associations-IFLA, in 1998).

RDA will provide a set of guidelines and instructions on formulating descriptive data and access point control data to support resource discovery. Being developed as a web based product RDA is especially designed for description and access for digital resources. Where AACR2 is an arrangement of rules based on the format of the item described, the focus of RDA is to be a standard for describing content rather than a display standard. The terminology of AACR2 is being revised, but many of the concepts are being retained. For example, instead of "heading," RDA might use "access point." The concept of "main entry," becomes "primary access point." "Uniform title" will be retained as "preferred title."

To pursue global sharing of information resources RDA is being developed in line with a set of objectives and principles which are based on the IME ICC (IFLA Meeting of Experts on an International Cataloguing Code) draft Statement of International cataloguing principles.

There are three parts in the Resource Description and Access cataloguing rules. The first part will relate to descriptive cataloguing, and will outline general rules for description of an item. This will be followed by the supplementary rules for specific formats.

An important feature of RDA will be its independence from the presentation of data. It will provide guidance on the recording of data, the content, and not on how it might be organized on the screen. That means that RDA-based cataloguing can be readily accommodated in many other than MARC encoding standards and metadata schema, thus intended to be independent of any cataloguing code. As a result, the more user-friendly presentations of bibliographic data can be introduced.

Another difference is in the format of general material designations (GMD).



*Did u know?* The RDA will allow, as for a proposal, for a two-part GMD which might be called "type and form of resource".

The first part will describe the content and the second part will describe the carrier or the physical format. For example, a map or atlas on CD-ROM will be assigned the GMD of Music CD might have GMD [music recording: CD audio], and videocassette –Part II will cover description of bibliographic relationships, which will allow bibliographic records to express the relationships described in Functional Requirements for Bibliographic Records.

FRBR supports user tasks showing what tasks the user of a catalogue is to be able to accomplish: Find information that is similar to the user's search criteria Identify information user wants and eliminate information or entities user does not want Select a particular entity appropriate to user's needs Obtain it through loan or remotely.

FRBR comprises 3 groups of entities that key objects of interest to the users of bibliographic information :

Group 1 represents intellectual or artistic products: Work, Expression, Manifestation, and Item.

Group 2 entities are responsible for the intellectual or artistic content: Person and Corporate body. (responsibility relationships)

Group 3 entities are subjects of Group 1 or Group 2's intellectual endeavor, and include Concepts, Objects, Events, and Places.

## Self Assessment

Fill in the blanks:

1. The online catalogue does not need to be stored .....
2. The growth of ..... and computerization add to the need for that quality.
3. The catalogue stands at the core of all ..... services.
4. At the moment we catalogue a variety of electronic resources, among them are ..... etc.
5. Recently the committee decided that the new cataloguing code will be called ..... or RDA.

## 7.4 Concept Indexing

### Concept Indexing with WordNet Synsets

The popularity of the bag of words model is justified by the fact that words and its stems carry an important part of the meaning of a text, specially regarding subject-based classification. However, this representation faces two main problems: the synonymy and the polysemy of words. These problems are addressed by a concept indexing model using WordNet synsets.

The basic idea of concept indexing with WordNet synsets is recognizing the synsets to which words in texts refer, and using them as terms for representation of documents in a Vector Space Model. Synset weights in documents can be computed using the same formulas for word stem terms in the bag of words representation. This concept based representation can improve IR, as commented by Gonzalo et al. "(...) using WordNet synsets as indexing space instead of word forms (...) combines two benefits for retrieval: one, that terms are fully disambiguated (this should improve precision); and two, that equivalent terms can be identified (this should improve recall)."

It is important to note that available information in SemCor allows both sense and concept indexing. As sense indexing, we understand using word senses as indexing units. For instance, we could use the pair (car, sense 1) or "car s1" as indexing unit. Concept indexing involves a word-independent normalization that allows recognizing "car s1" and "auto-mobile s1" as occurrences of the same concept, the noun of code 02573998 in WordNet (thus addressing synonymy and polysemy simultaneously).



Task Explain the term Concept Indexing.

## Text Representation Approaches

We have tested two kinds of text representation approaches for TC: a bag of words model, and a concept indexing model. Given that the TC problem is more oriented to genre than to subject, we have considered four possibilities: using a stoplist and stemming, using a stoplist, using stemming, and finally using words (without stoplist and stemming).

These approaches are coded below as BSS, BSN, BNS, and BNN respectively. The motivation of considering these four approaches is that words occurring in a stoplist (*e.g.* prepositions, etc.) and original word forms (*e.g.* past suffixes as “-ed”) can be good indicators of different text genres. The three concept indexing approaches considered in our experiments are those regarding the level of disambiguation, which are: using the correct word sense (CD), using the first sense given the part of speech (CF), and using all the senses for the given part of speech (CA). We have weighted terms and synsets in documents using the popular TF.IDF formula from the Vector Space Model.

## Conclusion

In general, we have not been able to prove that concept indexing is better than the bag of words model for TC. However, our results must be examined with care, because of at least two reasons: (1) the lack of enough training data: stronger evidence is got when the number of documents increases; also, the behavior of some algorithms is surprising (Naive Bayes, kNN); and (2) the genre identification problem is such that meaning of used words is not more relevant than other text features, including structure, capitalization, tense, punctuation, etc. In other words, this study needs to be extended to a more populated, subject-oriented TC test collection (*e.g.* Reuters-21578). The work by Petridis et al. adds evidence of concept indexing outperforming the bag of words model on the SemCor collection, specially with SVM, and Fukumoto and Suzuki work with them on Reuters-21578 allow to say that such an study is well motivated and promising.

## Concept Indexing for Production Databases

To explore the feasibility of using the National Library of Medicine’s Unified Medical Language System (UMLS) Metathesaurus as the basis for a computational strategy to identify concepts in medical narrative text preparatory to indexing. To quantitatively evaluate this strategy in terms of true positives, false positives (spuriously identified concepts) and false negatives (concepts missed by the identification process).

## Methods

Using the 1999 UMLS Metathesaurus, the authors processed a training set of 100 documents (50 discharge summaries, 50 surgical notes) with a concept-identification programme, whose output was manually analyzed. They flagged concepts that were erroneously identified and added new concepts that were not identified by the program, recording the reason for failure in such cases. After several refinements to both their algorithm and the UMLS subset on which it operated, they deployed the program on a test set of 24 documents (12 of each kind).

## Results

Of 8,745 matches in the training set, 7,227 (82.6 percent ) were true positives, whereas of 1,701 matches in the test set, 1,298 (76.3 percent) were true positives. Matches other than true positive indicated potential problems in production-mode concept indexing. Examples of causes of problems were redundant concepts in the UMLS, homonyms, acronyms, abbreviations and elisions, concepts that were missing from the UMLS, proper names, and spelling errors.

## Conclusions

The error rate was too high for concept indexing to be the only production-mode means of preprocessing medical narrative. Considerable curation needs to be performed to define a UMLS subset that is suitable for concept matching.

**In a nutshell such are the emerging trends in cataloguing:**

Notes

- Importance of workforce planning taking into consideration
  - o Retirement and difficulty to find trained staff
- Increasing complexity of the cataloguing process as a result of
  - o The increasing number of online tools for cataloguers
  - o Rapid emergence of new formats
  - o Emphasis to indexing and metadata
  - o Cataloguing for diverse user environments and audiences.
- Growing need to provide multilingual cataloguing
- Increased rate of updates to cataloguing rules, subject headings and MARC 21.
- New type of presentation in OPACs prompted by
  - o the changes in user environments - cataloguers will need to include additional feature into the bibliographic record, such as book cover art, reader reviews, book summaries, etc.
  - o application of new cataloguing rules in OPACs – cataloguers will need to create links between the related records in the database, the expressions and manifestations of the work.

**Self Assessment**

Multiple Choice Questions:

6. Using the 1999 UMLS metathesaurus, the authors processed a training set of ..... documents with a concept-identification programme, whose output was manually analyzed.
- |         |         |
|---------|---------|
| (a) 100 | (b) 200 |
| (c) 300 | (d) 400 |
7. Of 8,745 matches in the training set, 7,227 were true positives, whereas of 1,701 matches in the test set .....
- |           |           |
|-----------|-----------|
| (a) 1,278 | (b) 1,288 |
| (c) 1,298 | (d) 1,308 |

**7.5 Summary**

- RDA will provide a set of guidelines and instructions on formulating descriptive data and access point control data to support resource discovery.
- In the grammatical sort order (used mainly in older catalogues), the most important word of the title is the first sort term.
- “Cataloguing is less and less represented in courses (at least in France), whilst it should be more and more developed.
- The project at the University of Queensland is an evidence that there are ways to attract the students to this challenging and, to my mind, satisfying profession.
- Development of new formats require modifications of the classification, bibliographic rules and subjects headings.

Notes

## 7.6 Keywords

**Indexing** : Sorting an operation that segregates items into groups according to specified criterion.

**RDA** : Resource Description and Access—provides a set of guidelines and instructions on formulating descriptive data and access point control data to support resource discovery.

## 7.7 Review Questions

1. Distinguish between grammatical short order and mechanical short order.
2. Write about current and emerging trends in cataloguing.
3. Give an idea about career in cataloguing.
4. Write a short note on concept indexing.

## **Answers: Self Assessment**

1. Statically
2. Information technology
3. Library
4. CDs, DVDs, CD-ROMs, VCDs
5. Resource description and Access
6. (a)            7. (c)

## 7.8 Further Readings



**Books**

Chowdhary, GG: *Introduction to Modern Information Retrieval*. London: LA, 1999.

Maltby, A., ed. *Sayer's manual of classification for libraries*. 5th. Ed. London: Andre Deutsch, 1975.

Oddy, P. *Future libraries, future catalogs*. London: LA, 1996.



**Online links**

<http://www-old.pgcc.edu/library/tutorial/catalog.htm>  
[www.wikipedia.com](http://www.wikipedia.com).

## Unit 8: Indexing

### CONTENTS

Objectives

Introduction

8.1 Indexing Development

8.1.1 Indexing Process

8.2 Index Development and Trends

8.3 Summary

8.4 Keywords

8.5 Review Questions

8.6 Further Readings

### Objectives

After studying this unit, you will be able to:

- Define indexing development
- Describe index development and trends
- Explain design phase and development phase.

### Introduction

Indexing is depending both on the document to be indexed and on the indexer performing the process under specific conditions in a specific environment. Different documents are of course indexed differently by the same indexer. If they were not the index would be non-discriminative and total useless. Any theory of indexing has to deal with this fact and thus with how document attributes or properties should influence its representation.

The same document may be indexed differently by different indexers or by the same indexer at different times or by different indexing systems or in different libraries, for different target groups or for different ideal purposes.

The indexing is close to the document if it is constructed by a set of terms selected mechanically from the document (*e.g.* from titles, references or full-text). This is the objective pole because the document is the object of the indexing process. Also the rhetorical view of indexing (Andersen 2004) is close to the objective pole emphasizing what the author of the document is arguing.

The subjective pole of indexing theory emphasizes that the same document may be seen differently by different people or systems and that the indexing should not aim at a purely objective representation but should also consider, for example, the collection to which the document belongs or the tasks for which the indexing is made. Automatic indexing usually represent the terms of a document relative to the terms frequency in a collection of documents. In this way is the representation not just a function of the document itself, but also a function of a collection.

Notes

Another example is that the same book may be indexed differently for library for gender studies compared to a library of historical studies. Still, the indexing has to be loyal to the document being indexed, but different aspects of the document may be emphasized and the subject may be expressed in different controlled vocabularies constructed to support either collection.

The importance of indexing documents specific to a specific discipline, task or point of view may be illustrated by an example from the Royal Library in Copenhagen. First, the practice in this library is that a given book is circulated to different subject bibliographers. Each subject bibliographer then make a decision whether the book is relevant to his or her discipline or not. If it is relevant it is then indexed within that discipline. In this way a given document may be indexed from multiple points of view in the same catalogue.



*Did u know?* Nynne Koch, began about 1972 to collect printed catalogue cards which she regarded important to a new field, which she defined and termed “feminology”.

This initiative later developed and became an important independent library and research center “KVINFO”. The important point in relation to indexing theory is that this new library was not started by a special collection of books, but by a new way of indexing books belonging to other disciplines. This example demonstrates the importance of the subjectivity of indexing: to regard the indexing in relation to the aim of the indexing system.

Indexing should not, of course, aim at an idiosyncratic understanding of the individual indexer. It is not his or her special interests or points of view, which should be emphasized. An indexer work in order to accomplish a goal which is implicit or explicit in a given library or information system. It is this goal, not the individual indexers goal which should form the basis for the indexing. This insight has led to an ideal of inter-indexer consistency. However, as pointed out by Cooper (1969), indexing may be consistently wrong, why studies of inter-indexer consistency may not necessarily provide a basis for indexing quality.

Discussions on abstracting cover such concepts as the different types of abstracts, purpose of an abstract, structured versus narrative abstracts, informative versus indicative abstracts, subject slanting, modular abstracts, and writing and evaluating an abstract.

Various styles of indexing used in printed publications such as Index Medicus, the Engineering Index, and Chemical Abstracts are illustrated in the text; although the author is quick to note that printed tools are used much less today in favor of their online counterparts. In the online world, indexing has even greater importance in the effort to retrieve relevant data efficiently. Related concepts such as weighted indexing, linking of terms, and relational indicators are discussed as aids to precision.

The idiosyncrasies of indexing special formats such as images and sounds and the Internet, as well as the use of computer-generated or automated indexing and abstracting, are also reviewed. The author admits that the Web has become so large and complex that it is beyond the scope of any single book to explain all of its components. He suggests the use of Web-based services such as The Extreme Web Searcher’s Internet Handbook News and Updates to keep current with new developments.

Lancaster quotes several authors who see indexing of the Web becoming more impossible with time, but, at the same time, see that the need for automatic abstracts or summarizations continuing to grow in importance. With automation, the need for human intervention at the local level, be it Website design or local resources management, will also increase.



*Notes* Lancaster work is primarily a teaching textbook that gives a good overview of the historic theory and principles behind indexing and abstracting and then discusses various applications, practices, and issues related to content analysis.

Adequate representation of the material being described is the core challenge with indexing and abstracting. Another work that addresses this core issue is *Explorations in Indexing and Abstracting: Pointing, Virtue, and Power* by Brian C. O'Connor. O'Connor defines "pointing" as the fundamental definition of indexing; "virtue," the essence of a work, as equal to abstracting; and the two tools together as giving a person "power" to make meaningful use of the information.

Another related title, *Introduction to Indexing and Abstracting* by Donald B. Cleveland and Ana D. Cleveland, is more practical than the other two titles in that it provides many examples of what is being discussed and includes a section on "Ninety-nine Web Resources for Indexers and Abstractors," which leads to useful tools such as indexing services, standards, indexing organizations, and search services. One of the listed search services is a company working to answer Lancaster's challenge about the daunting task of indexing the Web. "The perfect search engine would understand exactly what you mean and give back exactly what you want." Many of the principles discussed in Lancaster's work—precision, specificity, and depth of indexing—are just as applicable and essential in today's online world as companies like Google seek to develop the "perfect search engine".



*Task* Find the latest trends in indexing in India.

## Self Assessment

Multiple Choice Questions:

- The ..... document may be indexed differently by different indexers.
  - Different
  - Same
  - None of these.
- The indexing is ..... to the document if it is constructed by a set of terms selected mechanically from the document.
  - Close
  - Open
  - None of these.
- ....., the practice in this library is that a given book is circulated to different subject bibliographers.
  - First
  - Second
  - Third
  - Fourth.

## 8.1 Indexing Development

An **index** is a list of words or phrases ('headings') and associated pointers ('locators') to where useful material relating to that heading can be found in a document. In a traditional back-of-the-book index the headings will include names of people, places and events, and concepts selected by a person as being relevant and of interest to a possible reader of the book. The pointers are typically page numbers, paragraph numbers or section numbers. In a library catalogue the words are authors, titles, subject headings, etc., and the pointers are call numbers.

### 8.1.1 Indexing Process

#### Conventional Indexing

The indexer reads through the text, identifying indexable concepts (those for which the text provides useful information and which will be of relevance for the text's readership). The indexer creates index headings, to represent those concepts, which are phrased such that they can be found when in alphabetical order (so 'indexing process' rather than 'how to create an index'). These headings and



**Notes**

their associated locators (indicators to position in the text) are entered into specialist indexing software which handles the formatting of the index and facilitates the editing phase. The index is then edited to impose consistency throughout the index.

Indexers must analyze the text to enable presentation of concepts and ideas in the index that may not be named within the text. The index is intended to help the reader, researcher, or information professional, rather than the author, find information, so the professional indexer must act as a liaison between the text and the its ultimate user. Indexing is often done by freelancers hired by authors, publishers or book packagers. Some publishers and database companies employ indexers. There are several dedicated, indexing software programs available to assist with the special sorting and copying needs involved in index preparation. The most widely known include Cindex, macrex, PDF Index Generator, SkyIndex and TExtract.

**Embedded Indexing**

Embedded indexing involves including the index headings in the midst of the text itself, but surrounded by codes so that they are not normally displayed. A usable index is then generated automatically from the embedded text using the position of the embedded headings to determine the locators. Thus, when the pagination is changed the index can be regenerated with the new locators.

LaTeX documents support embedded indexes primarily through the makeIndex package. Several widely-used XML DTDs, including DocBook and TEI, have elements that allow index creation directly in the XML files. An embedded index requires essentially the same amount of work to create as a conventional static index; however, this work differs slightly in character as the original source files are being edited, which may slow the process or prove distracting. An embedded index saves considerable work if the material will be updated even infrequently.



*Task* What is an Index? Explain Indexing process.

**8.2 Index Development and Trends**

The notion of building indexes and forgetting about them should not be used as philosophy when thinking about database indexes. Indexes need to be well thought out and tweaked over time. You need to develop an indexing development lifecycle to build and manage your indexes appropriately over time. It will give you some ideas that you can use to establish an indexing development lifecycle for your environment.



*Did u know?* In most IT shops, there are at least three different environments: Production, Quality Assurance/Test, and Development.

Your T-SQL code migrates through these different environments as your code progresses from one development phase to another, and so should your indexes. Therefore, why not have the following lifecycle phases for developing and maintaining indexes: Design, Development, Acceptance Testing, Production, and Maintenance. Let go through each one of these phases and discuss the kinds of tasks you should consider performing in each phase.

**Design Phase**

The design phase for indexes is just like the design phase for developing code. In this phase, you should look at the data model of your new database and consider the processing requirements your

application will need to go through to meet the business rules you have defined. Your programs will need to read data a particular way to build reports. Alternatively, you will have an online screen that will allow users to enter some search criteria, so different screens can be displayed. Pay attention to these different data access requirements of the application. If a report needs data to come back in sorted order based on a column value then this column would be a good candidate for being in an index. If customers need to enter an ID and date range to return some customer records on a screen, then the ID and date column associated with the date range would be additional candidate columns for indexes.

In the design phase, you need to get a sense for which columns are being used, and what order those columns will be returned to the application. You can then use this information to design some best guesses at what indexes your application will need. By doing this data analysis you will have the information you need to start identifying some indexes that will most likely be useful for your application.

## Notes

### Development Phase

In the development phase, you will review how well those best guess indexes are meeting the needs of your application. Just like any other development phase, this is the phase where you will be tweaking those indexes in your design when you realize they are not meeting your application data access requirements.



*Notes*

As you find your code is performing poorly you will add more indexes. Keep in mind indexes don't come for free. The more indexes you have the more costly an INSERT, UPDATE and DELETE statement will be if it has to update a bunch of indexes.

Therefore, create indexes for those queries in your application that are going to be run frequently. If you also have monthly or yearly queries that take a long time, possibly you can do without an index. Possibly you can create these indexes once a month, or year for these monthly/yearly processes. These are the kinds of decisions you need to make when you are developing indexes. You need to have well balanced indexes so your application performs acceptable most of the time for the frequently run queries. You need to make sure you do not have too many indexes that cause the INSERT, UPDATE, and DELETE statements to take a long time.

Once you are done with developing your application code, but prior to moving into user acceptance testing, you need to go back and review the indexes you have. Make sure all the indexes you created during your development phase are the ones that really made your queries run faster and they are not ones you created that provided little value. If you removed those useless indexes as you went along then you probably can omit this step. Another thing to do is verify that you don't have any duplicate indexes. If you have duplicate indexes you are just wasting resources, both disk space and the cost of maintaining them over time. Here is a script, written by Paul Neilson, that can be used to identify those duplicates:

```
- From Paul Neilson: http://sqlblog.com/blogs/paul\_nielsen/archive/2008/06/25/find-duplicate-indexes.aspx
- exact duplicates
  with indexcols as
(
select object_id as id, index_id as indid, name,
(select case keyno when 0 then NULL else colid end as [data()])
```

**Notes**

```
from sys.sysindexkeys as k
where k.id = i.object_id
and k.indid = i.index_id
order by keyno, colid
for xml path('') as cols,
(select case keyno when 0 then colid else NULL end as [data()])
from sys.sysindexkeys as k
where k.id = i.object_id
and k.indid = i.index_id
order by colid
for xml path('') as inc
from sys.indexes as i
)
select
object_schema_name(c1.id) + '.' + object_name(c1.id) as 'table',
c1.name as 'index',
c2.name as 'exactduplicate'
from indexcols as c1
join indexcols as c2
on c1.id = c2.id
and c1.indid < c2.indid
and c1.cols = c2.cols
and c1.inc = c2.inc;
```

### Acceptance Testing Phase

Now that you have promoted your database and application code to the next level in the development lifecycle, you need to start reviewing those indexes you created. Hopefully, in this phase of your application development cycle you will have users actually trying out your application. They will be running your application code through its paces by testing those reports, and search capability you've built into your application. Having your testing staff use your application gives you an opportunity to now see which indexes are being used by real application usage.

One of the things you should consider doing in this phase is to monitor your index usage. Look at how often each index is being used. Pay particular attention to which indexes are being updated heavily. As part of this analysis, you should determine those indexes that are constantly being updated, and never used. For those indexes that are not being used but are being update, you might want to consider whether you need those indexes.

In addition, for indexes that are being used and updated frequently you might want to review the FILLFACTOR to make sure it is appropriate, so that you don't get too many page splits. You can use the Dynamic Management Views (DMV's) that became available with SQL Server 2005 to find index usage information. Keep in mind this dynamic information is only collected since the last time SQL Server started, so it might not contain all of the statistics associated with indexes, especially those indexes that are infrequently used. Here is a sample script that shows you the index usage information for the AdventureWorks database.

```
USE AdventureWorks;
GO
-- Display Index Usage Information
-- Written By Gregory A. Larsen
SELECT o.name Object_Name,
       SCHEMA_NAME(o.schema_id) Schema_name,
       i.name Index_name,
       i.Type_Desc,
       CASE WHEN (s.user_seeks > 0
                 OR s.user_scans > 0
                 OR s.user_lookups > 0)
            AND s.user_updates > 0
       THEN 'USED AND UPDATED'
       WHEN (s.user_seeks > 0
            OR s.user_scans > 0
            OR s.user_lookups > 0)
            AND s.user_updates = 0
       THEN 'USED AND NOT UPDATED'
       WHEN s.user_seeks IS NULL
            AND s.user_scans IS NULL
            AND s.user_lookups IS NULL
            AND s.user_updates IS NULL
       THEN 'NOT USED AND NOT UPDATED'
       WHEN (s.user_seeks = 0
            AND s.user_scans = 0
            AND s.user_lookups = 0)
            AND s.user_updates > 0
       THEN 'NOT USED AND UPDATED'
       ELSE 'NONE OF THE ABOVE'
       END AS Usage_Info,
       COALESCE(s.user_seeks,0) AS user_seeks,
       COALESCE(s.user_scans,0) AS user_scans,
       COALESCE(s.user_lookups,0) AS user_lookups,
       COALESCE(s.user_updates,0) AS user_updates
FROM sys.objects AS o
     JOIN sys.indexes AS i
ON o.object_id = i.object_id
   LEFT OUTER JOIN
```

**Notes**

```

sys.dm_db_index_usage_stats AS s
ON i.object_id = s.object_id
    AND i.index_id = s.index_id
WHERE o.type = 'U'
    - Clustered and Non-Clustered indexes
AND i.type IN (1, 2)
AND (DB_NAME(s.database_id) = 'SmarTPH' or s.database_id IS NULL);

```

Another thing to look at is missing index statistics. As the database engine is processing queries, it determines whether a query would perform better if an additional index were added to the database. This information is called missing indexes. The missing index information can be exposed using DMV's. You can use this missing index information to determine which new indexes you might need. Keep in mind there might be lots of missing indexes identified.

You should only consider adding those missing indexes that are identified to have a large number user seeks and scans. These are the missing indexes that will be used quite frequently. Once again, you need to keep a good balance of indexes, so don't add every missing index. Here is a query that identifies the missing indexes for the AdventureWorks DB, and orders them by user scans and seeks:

**Build Create Index Statements From Missing Indexes**

```

d.statement AS [ObjectName],
gs.unique_compiles,
gs.user_seeks,
gs.user_scans,
gs.avg_total_user_cost,
gs.avg_user_impact,
'CREATE INDEX MissingIndex_' + rtrim(cast(d.index_handle AS char(100)))
    + ' ON ' + d.statement + ' (' +
CASE WHEN equality_columns IS NOT NULL THEN equality_columns ELSE ''
    ND +
CASE WHEN equality_columns IS NOT NULL AND
    inequality_columns IS NOT NULL THEN ', ' ELSE '' END +
CASE WHEN inequality_columns IS NOT NULL THEN inequality_columns ELSE
    '' END + ') ' +
CASE WHEN included_columns IS NOT NULL THEN 'INCLUDE (' + included_columns
    + ')' ELSE '' END AS MissingIndex
FROM sys.dm_db_missing_index_groups g
    join sys.dm_db_missing_index_group_stats gs ON gs.group_handle
        = g.index_group_handle
    join sys.dm_db_missing_index_details d ON g.index_handle
        = d.index_handle

```

After you have reviewed your index usage statistic and your missing indexes, you need to determine what index modification you need. Make any index modification you need first in your development environment, and then promote them up to your acceptance testing environment. Every time you

add new indexes to your environment, don't forget to review all of your indexes to make sure you have not included any new duplicate indexes.

#### *Production Phase:*

Once you have promoted your code and your database to the production environment, are you done developing indexes? No! There is always index work to consider once you are in the production phase.

Now that your application code is finalized and in production, you will be able to examine real life usage of your application to determine how those indexes are really being used and updated. Therefore, you should start gathering index usage statistics. You need to consider keeping the index usage statistics over time, as well as the missing index information. Keeping your index statistics and missing index information over time will allow you to be able to determine how useful those indexes really are, and how often they have been updated.

After you have tracked your index usage information for a week, a month, and/or a year you will really be able to tell which indexes really have not been used. After you have gathered a sufficient amount of index usage and missing index information, you can use this information to manage your indexes. This collected index usage information will help identify those indexes that are being updated frequently or rarely, if ever used, as well as those missing indexes that are frequently missing. You can then use this information to tweak your indexes. Lastly don't forget to look for duplicate indexes once and while, especially after you have modified your indexes and/or added new indexes.

#### *Maintenance Phase:*

The last phase of the index development lifecycle is maintenance. Of course, you have been doing index maintenance all along the way by added, dropping and modifying indexes.

As your indexes are updated, they will become fragmented. The more fragmented your indexes the more pages they take up, the more I/O it takes to traverse the index, and the slower they perform. Periodically you need to review your index fragmentation to determine how fragmented your indexes are.

By reviewing the fragmentation information, you will be able to determine if you should identify a new FILL FACTOR for an index. Based on how fragmented your indexes are you might want to perform an index rebuild versus an index organization operation. The Microsoft recommendation is to rebuild an index if the index fragmentation is greater than 30% and reorganize it if the index fragment is between 5% and 30%. Here is some code that will help you identify the index fragment information for your indexes:

```
SELECT DB_NAME(ps.[database_id]) AS [database_name],
       OBJECT_NAME(ps.[object_id], DB_ID()) AS [object_name],
       si.[name] AS [index_name], ps.partition_number,
       ps.index_type_desc, ps.alloc_unit_type_desc,
       ps.index_level,
       ps.[avg_fragmentation_in_percent],
       ps.[page_count]
FROM sys.dm_db_index_physical_stats(DB_ID(), NULL, NULL, NULL, 'LIMITED')
     ps
JOIN sys.sysindexes si
ON ps.OBJECT_ID = si.id
AND ps.index_id = si.indid
WHERE index_type_desc <> 'HEAP'
```

## Notes

One last thing to talk about is index statistics. When you create an index, SQL Server generates index statistics. SQL Server will also keep these statistics up to date as your index gets updates, provided your database setting for "AUTO\_CREATE\_STATISTICS" is on (which is the default). The database engine keeps these statistics up to date based the percentage of pages that have been updated. When around 20% or more of the data rows have been updated, SQL Server will automatically update your index statistics. Keep in mind when SQL Server automatically creates statistics it does s by only sampling that data rows. If your tables are quite large, in the millions or billions of rows, it might take quite a while to update 20% of the rows for the statistics to get automatically updated.

Therefore, because of the sampling method and the 20% rule you might want to consider routinely updating your index statistics manually. Doing this will give the database engine updated statistics and this can help drastically in improving the performance of your application. Up to date statistics allow the database engine to make the appropriate choices when selecting an execution plan. Here is a script that will tell you when the statistics were last updated:

### Display Index Statistic Update Date

```
SELECT s.name, o.name, i.name, STATS_DATE(i.object_id, i.index_id)
StatisticsLastUpdated, i.type_desc
FROM sys.indexes I
JOIN sys.objects o
ON i.object_id = o.object_id
JOIN sys.schemas s
ON o.schema_id = s.schema_id
WHERE o.name NOT like 'sys%'
AND STATS_DATE(i.object_id, i.index_id) IS NOT NULL
ORDER BY STATS_DATE(i.object_id, i.index_id)
```

### Good Indexes are Not a Mistake

Creating good indexes for your application does not happen by mistake. You need to have a plan for how you will develop your indexes. Think of developing indexes using an indexing development lifecycle approach. Using this method gives you the best shot at hitting the mark when it comes to developing good indexes for your application.

### Self Assessment

Fill in the blanks:

4. The notion of building indexes and forgetting about them should not be used as philosophy when thinking about .....
5. The ..... for indexes is just like the design phase for developing code.
6. The missing index information can be exposed using .....
7. You should only consider adding those missing indexes that are identified to have a large number user ..... and .....
8. The last phase of the index development lifecycle is .....

### 8.3 Summary

- Indexing is depending both on the document to be indexed and on the indexer performing the process.
- The importance of indexing documents specific to a specific discipline, task or point of view may be illustrated by an example from the Royal Library in Copenhagen.
- Various styles of indexing used in printed publications.
- In most IT shops, there are at least three different environments: Production, Quality Assurance/Test, and Development.
- The design phase for indexes is just like the design phase for developing code.

### 8.4 Keywords

**Indexing** : Depends upon both on the document to be indexed and on the indexer performing the process under specific conditions in a specific environment.

**Clustering** : Aims to bring together group of closely related documents.

### 8.5 Review Questions

1. Write brief note on indexing development.
2. Explain the design phase for indexes.
3. Good indexes are not a mistake. Comment.

### Answers: Self Assessment

1. (b), 2. (b), 3. (a)
4. Database indexes
5. Design phase
6. DMV's
7. Seeks and scans.
8. Maintenance

### 8.6 Further Readings



Books

Cooke, A : *A guide to finding quality Information on the Internet*. 2<sup>nd</sup> Edition. London: Facet Publishing, 2001.

Oddy, P. *Future libraries, future catalogs*. London: LA, 1996.

Chowdhary, GG: *Introduction to Modern Infomation Retrieval*. London: LA, 1999.



Online links

[http://everything.explained.at/index\\_%28publishing%29/www.wikipedia.com](http://everything.explained.at/index_%28publishing%29/www.wikipedia.com).



## Unit 9: Trends in Indexing

### CONTENTS

Objectives

Introduction

9.1 Derived Indexing

9.2 Assigned Indexing

9.3 Alphabetical Indexing

9.4 Keyword Indexing

9.5 Pre-coordinate Indexing

9.6 Post-coordinated Indexing

9.7 Pre or Post-coordinate Indexing

9.8 Pre-coordinate Indexing/Post-coordinate Indexing System

9.8.1 Pre-coordinate Indexing System

9.8.2 Post-coordinate Indexing System

9.9 Citation Indexing

9.10 Summary

9.11 Keywords

9.12 Review Questions

9.13 Further Readings

### Objectives

After studying this unit, you will be able to:

- Describe derived indexing and assigned indexing.
- Explain alphabetical indexing and keyword indexing
- Describe pre-coordinate indexing and post-coordinate indexing
- Explain citation indexing.

## Introduction

Notes

Key trends in indexing include Islamic indices, frontier markets and alpha-producing indices.

There is growing demand for better representation of Asia, especially the Hong Kong-China-Taiwan relationship.

Stock exchanges on a global basis are becoming increasingly interested in the index business and are looking to use indices for derivative purposes.

Any further slicing and dicing of indices is probably redundant for retail investors.



*Did u know?* According to Andrew Clark, some of the hot topics in indexing include Islamic indices, frontier markets (including Africa), and alpha-producing indices.

On the ETF side, he said there appears to be interest in bringing out indices which produce alpha. From what happened during the financial crisis, the hope is that the risk controls are in place, so people are willing to take on more risk, he explained.

At the same time, there seems to be a need for a better representation of Asia, said Clark, especially the Hong Kong-China-Taiwan relationship.

Market participants generally feel that the very few indices out there which represent all three markets are not very good, he said.

There is also a focus from stock exchanges on using indices for derivative purposes – either from indices which already exist, or by creating custom ones, said Clark.

Exchanges on a global basis are becoming increasingly interested in the index business, he said.

## Working with Stock Exchanges

When it comes to index providers and stock exchanges working together, an exchange might, for example, have an existing index which it wants to turn into a derivative, explained Clark.

However, if there are 300 securities in the index, the exchange will need to get a good representation that gives the same basic return but with only one-tenth of the number of securities.

Index providers can reduce such an index to 30 stocks, with the same basic return and fairly good correlation, said Clark. The exchange can then take that smaller index and turn it into a derivative.

## Trends in Global Indexing

New indices and index strategies

- Broader benchmarks
- Narrower benchmarks
- Customized benchmarks
- Benchmarks for alternative asset classes
- “Active” benchmarks

Notes

**Self Assessment**

Multiple Choice Questions:

1. Key trends in indexing include ..... indices.  
(a) Chinese (b) European  
(c) Islamic (d) African
2. Clark said, market participants generally feel that the very few indices out there which represent all ..... market are not very good.  
(a) Two (b) Three  
(c) Four (d) One
3. Index providers can reduce such an index to ..... stocks, with the same basic return and Fairly good correlation, said Clark.  
(a) 30 (b) 40  
(c) 50 (d) 60

**9.1 Derived Indexing**

Derived indexing terms are terms occurring in the text to be indexed. Assigned terms are terms not occurring in the text.

“There are essentially two approaches to the creation and maintenance of this document or knowledge representation. One is to create a knowledge system in advance and assign the documents to it afterward: assigned indexing. The other is to derive the terms of the index language from the documents themselves: derived indexing.

The manual library systems, in which books were classified according to an existing classification system, for example, Dewey or UDC, are assigned indexing systems; computerized IR-systems which extract keywords from the documents according to a weighting scheme are typically derived indexing systems. We will extend the definition of assigned indexing systems to contain all systems that use terms in their docreps (document representations) that are not taken from the documents themselves, because such external terms belong to a knowledge representation outside the document.

The derived indexing systems became very popular when the computer made it easy to create an inverted list of all the words occurring in a document base. In the 1970s and '80s much effort was put into the development of techniques to identify such words (phrases, sentences) in the inverted lists as were most efficient in retrieving particular documents.

Assigned terms may come from external semantic resources (*e.g.*, authority files, classification systems or thesauri) or other kinds of external information.

Derived indexing systems are generally more primitive compared to assigned systems (or, of course, combinations). It is easy mechanically to mark a text for words to appear in an index, and to construe an index on this basis. However, users searching for a concept using a synonymous term, a broader term or a narrower term, will miss the information. This is the rationale behind controlled vocabularies.



*Task* Explain Derived Indexing.

It is common to classify documents according to an organization of disciplines. Documents may or may not describe their disciplinary memberships. Even if they do, the authors' organization of disciplines may be different from those chosen to be assigned by a library or an information system.

Assigning terms, which is not simple substitutions of synonyms, but which represents independent conceptualizations of document contents may turn out to be the most important area in which human indexing is better than automatic indexing.

Notes

As traditional classification is a time-consuming and expensive process, it is obvious that investigations into the use of automated solutions are worthwhile. At the same time, classification is an activity where a significant level of human expertise, abstract thinking and understanding is needed and this is not easy to replace by artificial intelligence or expert systems. There are no known examples of traditional library classification being undertaken completely by computer software. Knowledge structuring on the Internet has to cope with far larger numbers of resources, exponential growth rates and a high risk of changes occurring in documents which already exist.

This is the background to a growing number of research projects and experimental systems which are trying to support knowledge-structuring activities on the Internet with automatic methods. Most of these projects use methods of derived indexing, *i.e.* they extract information from the documents and then use it for structuring tasks.

Automated classification will probably not replace intellectual classification as far as quality subject services are concerned, but will rather support and complement selection and subject indexing efforts. Intellectual classification is always needed to validate and improve the automatic methods. However, robot-generated databases, as an add-on to quality services in a subject area, will be automatically classified. One practical goal in DESIRE II is to explore simple applications of automated classification methods on a robot-generated subject index to the Web.



Notes

Many different tests will be carried out on the 'All' Engineering (AE) robot-generated database of engineering documents from the Internet.

The effort required will be studied and the resulting outcomes evaluated. A pilot service of the 'All' Engineering Web index will offer a full classification and browsing structure with the most suitable solution found during the project. In addition, a comprehensive state-of-the-art report on projects, methods, alternatives and problems concerning automatic classification will also be presented.

## 9.2 Assigned Indexing

Assigned terms may, on the one hand simply substitute terms represented in the document with other terms, *e.g.* from a controlled vocabulary. On the other hand, an assigned term may represent a conceptualization of the document, which is not expressed in the document with any terms. A romantic poem, for example, does not describe itself as such, but may be assigned the term "romantic poem". It is common to classify documents according to an organization of disciplines.

Documents may or may not describe their disciplinary memberships. Even if they do, the authors organization of disciplines may be different from those chosen to be assigned by a library or an information system. Assigning terms, which is not a simple substitutions of synonyms, but which represents independent conceptualizations of document contents may turn out to be the most important area in which human indexing is better than automatic indexing.

From the preceding discussion, it is clear that if the terms are selected from the title or the text of a document and used without any alteration as index terms, then this is referred to as natural language indexing or derived indexing. If however, the selected terms are translated or encoded into authorized terms by the help of a prescribed list, then the indexing language becomes controlled or artificial. This process is called Assigned Indexing.

**Notes**

Derived Indexing solely relies on information which is manifest in the document, without attempting to add to this from indexer's own knowledge or other sources. We looked at ways in which printed indexes could be derived from information manifest in a document. We can also consider some of the ways in which files may be searched online, again using the information manifest in the document, e.g. titles, abstracts or full text. By doing so we have to face the problems of natural language. A discussion to these problems leads to the idea of assigned indexing.

If we are to use a list of words to help us in our searching, we would increase the chances of achieving successful matches if we used the same list of words to encode the appropriate words to the documents ourselves rather than rely on authors' choice. In other words, we devise an indexing language and used this for both encoding operations: input and question. Such systems are referred to as assigned indexing systems. Assigned indexing involves intellectual process. Subject heading schemes, thesaurus and classification schemes are the popular forms of assigned indexing. Assigned indexing is also known as concept indexing.

### **9.3 Alphabetical Indexing**

Every business must develop and maintain an organized way to store written communication, such as reports, letters, memorandums, order forms, invoices, and other such information so that it is available for efficient retrieval or reference. This method of storing records is called filing. While there are a number of different methods for storing or filing information—alphabetic, subject, numeric, and geographic - the most common method is the alphabetic filing system.

Procedures for storing records alphabetically will vary among organizations and even among departments within an organization. Therefore, the filing procedures to be used in any one office needs to be determined, recorded, approved, and followed, without exception. Without written rules for storing records, procedures will vary with time, changes in personnel, etc. These changes could cause difficulty in future retrieval of records or even in the loss of records.

The Association of Records Managers and Administrators, Inc. (ARMA) is an organization designed to help professionals in records management perform their jobs easier and better.



*Did u know?*

ARMA has published a list of Alphabetic Filing Rules, containing standard rules for storing records alphabetically.

### **Basic Filing Terms**

Before learning the 12 filing rules, an understanding of filing terms is necessary.

**Unit** – Each part of a name is a unit. Names are alphabetized unit by unit. If there are two parts in a name, the name has two units.

**Indexing** – Indexing is determining the order and format of the units in a name. Is a person's record filed by first or last name? Is a business record filed under T if the name begins with The? Is punctuation considered with alphabetizing a name? Indexing is deciding which name to file a record under and then arranging the units in that order.

**Alphabetizing** – When you arrange names in alphabetical order, you are alphabetizing them. There are 3 basic categories for alphabetizing names: Personal Names, Business or Company Names, and Government Names.

**Alphabetizing Unit by Unit.** The first step in alphabetizing is to alphabetize Unit by Unit. If the names in Unit 1 are exactly the same, then continue to alphabetize by Unit 2. If the first and second units are the same, the next step is to alphabetize Unit 3, and so on.

Notes

| NAME                | UNIT 1 | UNIT 2  | UNIT 3 |
|---------------------|--------|---------|--------|
| Jessica Marie Adams | ADAMS  | JESSICA | MARIE  |
| Susan K. Adams      | ADAMS  | SUSAN   | K      |
| Susan P. Adams      | ADAMS  | SUSAN   | P      |

- **Nothing Comes Before Something.**
- **In alphabetizing, it is important to remember that nothing comes before something.**

| NAME              | UNIT 1    | UNIT 2 | UNIT 3 |
|-------------------|-----------|--------|--------|
| Ann B. Shoemaker  | SHOEMAKER | ANN    | B      |
| Anne B. Shoemaker | SHOEMAKER | ANNE   | B      |
| J. Tilden         | TILDEN    | J      |        |
| John Tilden       | TILDEN    | JOHN   |        |

- **Case.** The **case** of a letter refers to whether the letter is written as a capital letter (A), called uppercase, or written as a small letter (a), called lowercase.
- **In alphabetizing, uppercase and lowercase letters are considered the same.**
- **For example, McAdams and Mcadams are considered to be exactly the same when alphabetizing**

| NAME              | UNIT 1     | UNIT 2  | UNIT 3 |
|-------------------|------------|---------|--------|
| Ashley Mcadams    | MCADAMS    | ASHLEY  |        |
| Ashley K. McAdams | MCADAMS    | ASHLEY  | K      |
| Patrick McDonald  | MCDONALD   | PATRICK |        |
| Phillip Mcdonald  | MCDONALD   | PHILLIP |        |
| A.B. Stillworth   | STILLWORTH | AB      |        |
| Abe Stillworth    | STILLWORTH | ABE     |        |

## 9.4 Keyword Indexing

The success of this type of index depends entirely on the keywords you choose. The choice of words needs to be carefully thought through. If words are not chosen carefully, the resulting index can be overly long and unusable. Do not use simple words or words that repeat constantly throughout the text. For example, if you are writing a cookbook, do not index the word egg. Every time the word egg is mentioned, the computer will index that page. We presume that egg would show up on many recipe pages.

### Keyword Indexing by Simple Software

Simple Index is a powerful yet affordable solution for Keyword Indexing. Simple Index is designed to allow a desktop user with a high-speed scanner to quickly scan and file many documents with a minimal amount of manual data entry.

**Notes**

Simple Index is also much easier to use because it is designed for single-user scanning environments instead of large scanning service bureaus. Its unique 1-click interface lets you scan, process barcodes, OCR, rename files and upload them to a server all in one step or even as an unattended process.

A computer-generated keyword index lists a page number for a key term each time it occurs in the book. You simply provide Trafford with a list of words that you want to appear in the index. Trafford will create an index by tagging the selected words during the production process. The resulting alphabetical list of keywords gives the page numbers for each occurrence of the word.

### **9.5 Pre-Coordinate Indexing**

Since the indexing is coordinating (combining, pulling together concepts) then they engage in an act of synthesis to build one long index entry.

This can happen in three ways:

- Represent a single subject flowers or flowers and shrubs
- An aspect of a single subject fertilization of flowers, arrangement of flowers
- Two or more subjects treated in relation to one another flowers in art, flowers in religion folklore, etc.

### **9.6 Post-coordinated Indexing**

Post-coordinated indexing is the opposite of pre-coordinated indexing. There are pros and cons to each method of indexing.

Terms from the index are combined during the searching, rather than before, to create an index based on the individual search result. Facets can be combined ad finitum following the standpoint of the user. The index is created after the information is added to the database. Post-coordinated indexing will general increase recall but will usually decrease precision.



*Notes*

Post-coordinated indexing is almost always associated with computers, and usually uses some type of boolean logic and descriptors. Post-coordinated indexing allows searchers the freedom to freely combine many terms that are relevant to the search.

### **9.7 Pre or Post-coordinate Indexing**

Most people think about what they want to search for and are willing to combine their concepts at the time of search. They put in the combination of terms they are thinking of; they are doing the coordination of terms. It is up to the search software to do the intersection of the terms for them and figure out the post-coordination.

In the current online environment, very seldom do we put together terms in a pre-coordinate fashion. That is one of the challenges in taking older classified lists – back-of-the-book indexes—and making them into a post-coordinate system.

Post-coordination of terms is typical of traditional classification systems, and not of most modern taxonomies and thesauri. Classification systems often concatenate separate concepts into a string of terms. Natural language is not used.

Most people search by typing in words as they think of them, so we need to support natural language in our systems. This is part of why Access Innovations uses post-coordination in the taxonomies and thesauri it creates.

In a post-coordinate system, a single concept is represented by a single term. We're not combining two concepts, we are keeping them separate. That way we are able to do a large amount of automated indexing. If we try to mine the text at the beginning, it is very system-intensive. So, we have taken another path, by and large, for that.



*Task* Differentiate between Pre or Post-coordinate Indexing.

## **9.8 Pre-coordinate Indexing/Post-coordinate Indexing System**

Purpose of all kinds of indexing is the retrieval of information. There are basically two types of retrieval systems.

### **9.8.1 Pre-coordinate Indexing System**

The kind of system in which coordination is done at the time of indexing is called pre-coordinate indexing system. In this system documents or searched under the same terms which the indexer originally assigned to them without any further manipulation of terms at the time of searching. It means that whatever compound terms are used they are created at the time of indexing. Rather than at the time of searching.



*Did u know?* Co-relations are made during the indexing process and prior to use at the index, it is also called pre-coordinate or pre-correlative indexing.

The subjects represented in pre-coordinate indexes are shown with the entire component concepts coordinated. Thus, the entries in an index based upon pre-coordination are as complex as is necessary to describe the subject. But complex or composite subjects demand a series of entries and terms in order that they are described adequately.

An example:

- Chain indexing by S.R.Ranganathan
- PRESIS–Preserved context index system by derrick Austin
- POPSI–Postulate based Permuted Subject Indexing by G.Bhattacharya
- SLIC–Selective Listing In Combination by J.P.Sharp

### **Advantages:**

1. Pre-coordinate indexes eliminate the need for sophisticated search logic. The use at the index just looks under the terms that are expected to find the subject described. This is a direct method of search with which users are well acquainted.
2. It requires no special features in their physical format. Almost all printed indexes reflecting pre-coordinate indexing principles, are hard copy.
3. Its principles are applicable to a limited extent in on-line or off-line searched computer based information retrieval systems.



**Notes**

4. These also have found some application in subject indexes to library catalogues and the shelf arrangement of book-stock. These are to be found in abstracting and indexing journals, national bibliographies and indexes to journals.
5. In this single or multiple entry, present certain advantages at the search stage. It is possible for a number of searches to be conducted simultaneously.

**Limitations:**

1. In pre-coordinate systems, the multidimensional character at the subject matter is forced into a one-dimensional representations, which then necessitates to repeat the index entry in some way for example by rotation of the terms.
2. In this system relationships among topics are built once and for all into the system vocabulary or index entries formed from its components by the indexes. There are non manipulative.
3. A multiple access approach is possible, if we enter the document several times in the index by duplicating the citation.
4. These are also criticized on the ground that even the extensive duplication of entries does not provide the true multidimensional retrieval capability to multidimensional subject matter.
5. Efficient approaches to information retrieval demand such systems that permit the free "combination" of classes and the terms representing them.
6. A number of ways have been suggested to provide multiple approach to retrieval in pre-coordinate indexes without complete permutation of index terms.

### 9.8.2 Post-coordinate Indexing System

As the coordination of index terms is done after the index files has been compiled, this indexing system is called post-coordinate indexing system.

Examples for post-coordinate indexing system:

- Uniterm system of Taube dates about 1951
- Peek- aboo by batter in England and cordonnier in France by 1940
- Edge- notched card system by calerin mooers.

#### Common Features

1. None of the entries in the system are specific. There are relatively large number of documents under each heading and if the searches approaches the index as a conventional index, be in liable to become involved in extensive scanning of entries in order to discriminate between relevant and less relevant documents.
2. There are usually a larger number of entries in a post-coordinate indexing system than in an index based upon pre-coordinate indexing principles.
3. The number of different heading is the index is relevant small, because, as in classification a system scheme needless categories or heading than an equivalent enumerative scheme.

#### Conclusion

Thus in indexing it has pre and post-coordinated indexing system. There have some similarity and dissimilarities. It can be summed up as follows:

## Similarities

## Notes

- The subject content has to be analyzed and then, the standardized term has to be identified.
- In both types, the terms have to be co-ordinated.
- Both the systems involve the arrangement of the indexed cards in some logical order.

## Differences

- In input preparation
- Differences in access point
- Differences in arrangement
- Differences in search time
- Differences in browse ability.

## 9.9 Citation Indexing

A citation index is an index of citations between publications, allowing the user to easily establish which later documents cite which earlier documents. The first citation indices were legal citators such as Shepard's Citations (1873). In 1960, Eugene Garfield's Institute for Scientific Information (ISI) introduced the first citation index for papers published in academic journals, starting with the Science Citation Index (SCI), and later expanding to produce the Social Sciences Citation Index (SSCI) and the Arts and Humanities Citation Index (AHCI). The first automated citation indexing was done by CiteSeer in 1997.

Citation indexing is a way to look forward in the literature from the starting point of a particular paper or group of papers. This is a different and complementary approach to ordinary word-based literature searching, which looks backward in the literature from the present time.

For example, if you have an excellent paper on a particular topic that was published in 1992, you can use Science Citation Index (via Web of Science) to find papers published after 1992 that cited that paper. Citation implies a direct subject relationship between the papers. So, by searching for later papers citing your known paper, you can find more documents on the same or similar topic without using any keywords or subject terms.

## Major Citation Indexing Services

There are two publishers of general-purpose academic citation indexes, available to libraries by subscription:



*Notes*

ISI (now part of Thomson Scientific), which publishes the ISI citation indexes in print and compact disc. They are now generally accessed through the Web under the name Web of Science, which is in turn part of the group of databases in the Web of Knowledge.

Elsevier, which publishes Scopus, available online only, which similarly combines subject searching with citation browsing and tracking in the sciences and social sciences.

Each of these offers an index of citations between publications and a mechanism to establish which documents cite which other documents. They differ widely in cost: the ISI databases and Scopus are subscription databases, the others mentioned are freely available online.

## Notes

Citation data have long been used to rank journals within particular subject areas, usually based on the ISI Impact Factor. The impact factor is simply a numerical ratio of the total number of citations a journal receives in ISI Source Journals in one year to the total number of “citable” articles it published in the previous two years. It is a useful way to see how journals perform in relation to others in the same subject area. It is not useful in comparing journals across subject areas, and the number taken out of this context is essentially meaningless.

For example, Journal A has an impact factor of 4.327, and Journal B has an impact factor of 1.045. Is Journal A “better” than Journal B? You could conceivably make that argument, if you first accept the notion that quality equates with citedness, AND if journals A and B are both in the same field. But if A is in Biochemistry, and B is in Clinical Pharmacy, no such judgment can be made, as citation behavior varies considerably from field to field.

Impact factor can also vary based on the number and types of articles a journal publishes. Review articles tend to be more heavily cited than full papers or communications, so journals and annuals that publish mostly reviews will often have high impact factors. Journals that publish only a few articles in a given year may also have disproportionately high impact factors. Similarly, one very highly cited paper can skew a journal’s impact factor significantly.

Impact factors for journals covered by ISI are published annually in an electronic compilation called Journal Citation Reports. All ISI Source Journals are ranked within one or more relevant subject categories, such as CHEMISTRY, ORGANIC or SPECTROSCOPY. You can also compile customized lists. JCR also contains data on historical trends, immediacy index, cited half-life, etc.

## Citation Analysis

While citation indexes were originally designed for information retrieval purposes, they are increasingly used for bibliometrics and other studies involving research evaluation. Citation data is also the basis of the popular journal impact factor.

Legal citation analysis is a citation analysis technique for analyzing legal documents to facilitate the understanding of the inter-related regulatory compliance documents by the exploration the citations that connect provisions to other provisions within the same document or between different documents. Legal citation analysis uses a citation graph extracted from a regulatory document.

## Self Assessment

State whether the following statements are true or false:

4. Derived indexing terms are terms occurring in the text to be indexed.
5. Assigned indexing is a powerful yet affordable solution for keyword indexing.
6. The first citation indices were legal citators such as shepard’s citations (1873).
7. The first automated citation indexing was done by Citeseer in 1987.
8. Impact factors for journals covered by ISI are published annually in an electronic compilation called Journal Citation Reports.

## 9.10 Summary

- Derived Indexing solely relies on information which is manifest in the document, without attempting to add to this from indexer’s own knowledge or other sources.
- There are 3 basic categories for alphabetizing names: Personal Names, Business or Company Names, and Government Names.
- In alphabetizing, it is important to remember that nothing comes before something.

- In alphabetizing, uppercase and lowercase letters are considered the same.
- Simple Index is a powerful yet affordable solution for Keyword Indexing.
- Simple Index is also much easier to use because it is designed for single-user scanning environments instead of large scanning service bureaus.
- Post-coordinated indexing is the opposite of pre-coordinated indexing.
- In a post-coordinate system, a single concept is represented by a single term.
- The kind of system in which coordination is done at the time of indexing is called pre-coordinate indexing system.
- Pre-coordinate indexes eliminate the need for sophisticated search logic.
- In pre-coordinate systems, the multidimensional character at the subject matter is forced into a one-dimensional representations.
- A citation index is an index of citations between publications, allowing the user to easily establish which later documents cite which earlier documents.

### 9.11 Keywords

*Derived Indexing Terms* : Are terms occurring in the text to be indexed.

*Simple Index* : Is a powerful yet affordable solution for keyword Indexing.

### 9.12 Review Questions

1. Write about current Trends in Indexing.
2. What do you know about Assign Indexing and Alphabetical Indexing?
3. Define Pre-coordinate indexing system.
4. Explain advantages and disadvantages of pre-coordinate indexing system.
5. Write the major citation indexing services.

### **Answers: Self Assessment**

- |          |          |         |
|----------|----------|---------|
| 1. (c)   | 2. (b)   | 3. (a)  |
| 4. True  | 5. False | 6. True |
| 7. False | 8. True  |         |

### 9.13 Further Readings



Books

Best, DP, Ed. *The fourth resource: information and its management*. Aldershot: Aslib, 1996.

Chowdhary, GG: *Introduction to Modern Information Retrieval*. London: LA, 1999.

Deegan, M. and Simon Tanner. *Digital futures*. London. LA, 2002.



Online links

[http://www.geocities.ws/salman\\_mlisc/dissertation/chap3.htm](http://www.geocities.ws/salman_mlisc/dissertation/chap3.htm)

<http://www.iva.dk/bh/core%20concepts%20in%20lis/articles%20a-z/>

## Unit 10: Information Storage and Retrieval System

### CONTENTS

Objectives

Introduction

10.1 Information Retrieval System Evaluation

10.2 Precision and Recall

10.3 Precision

10.4 Recall

10.5 Relevance

10.6 Online Searching Basics; Library Databases; Web Sites

10.7 Summary

10.8 Keywords

10.9 Review Questions

10.10 Further Readings

### Objectives

After studying this unit, you will be able to:

- Define information storage and retrieval system
- Describe precision and recall
- Explain relevance
- Describe keywords searching and boolean operators.

### Introduction

An information storage and retrieval system (ISRS) is a network with a built-in user interface that facilitates the creation, searching, and modification of stored data. An ISRS is typically a peer-to-peer (P2P) network operated and maintained by private individuals or independent organizations, but accessible to the general public. Some, but not all, ISRSs can be accessed from the Internet.

Characteristics of an ISRS include lack of centralization, graceful degradation in the event of hardware failure, and the ability to rapidly adapt to changing demands and resources. The lack of centralization helps to ensure that catastrophic data loss does not occur because of hardware or programme failure, or because of the activities of malicious hackers. Graceful degradation is provided by redundancy of data and programming among multiple computers. The physical and electronic diversity of an ISRS, along with the existence of multiple operating platforms, enhances robustness, flexibility, and adaptability. (These characteristics can also result in a certain amount of chaos.) In addition to these features, some ISRSs offer anonymity, at least in theory, to contributors and users of the information. A significant difference between an ISRS and a database management system (DBMS) is the fact that an ISRS is intended for general public use, while a DBMS is likely to be proprietary, with access privileges restricted to authorized entities. In addition, an ISRS, having no centralized management, is less well-organized than a DBMS.



*Task* What do you mean by ISRS. Define?

## 10.1 Information Retrieval System Evaluation

To measure ad hoc information retrieval effectiveness in the standard way, we need a test collection consisting of three things:

- A document collection
- A test suite of information needs, expressible as queries
- A set of relevance judgments, standardly a binary assessment of either relevant or non-relevant for each query-document pair.

The standard approach to information retrieval system evaluation revolves around the notion of relevant and non-relevant documents. With respect to a user information need, a document in the test collection is given a binary classification as either relevant or non-relevant. This decision is referred to as the gold standard or ground truth judgment of relevance. The test document collection and suite of information needs have to be of a reasonable size: you need to average performance over fairly large test sets, as results are highly variable over different documents and information needs. As a rule of thumb, 50 information needs has usually been found to be a sufficient minimum.

Relevance is assessed relative to and not a query. For example, an information need might be:

Information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine.

This might be translated into a query such as: wine and red and white and heart and attack and effective.

A document is relevant if it addresses the stated information need, not because it just happens to contain all the words in the query. This distinction is often misunderstood in practice, because the information need is not overt. But, nevertheless, an information need is present. If a user types python into a web search engine, they might be wanting to know where they can purchase a pet python. Or they might be wanting information on the programming language Python.

From a one word query, it is very difficult for a system to know what the information need is. But, nevertheless, the user has one, and can judge the returned results on the basis of their relevance to it. To evaluate a system, we require an overt expression of an information need, which can be used for judging returned documents as relevant or non-relevant. At this point, we make a simplification: relevance can reasonably be thought of as a scale, with some documents highly relevant and others marginally so. But for the moment, we will use just a binary decision of relevance.

**Notes**

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

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

**Self Assessment**

Multiple Choice Questions:

1. The physical and electronic diversity of an ....., along with the existence of multiple operating platforms, enhances robustness, Flexibility, and adaptability.  
(a) IRSE (b) ISRS  
(c) DBMS (d) P2P
2. In Information retrieval a query does not uniquely identify a ..... object in the collection.  
(a) Single (b) Double  
(c) Triple (d) Fourth

**10.2 Precision and Recall**

Precision and recall are two widely used metrics for evaluating the correctness of a pattern recognition algorithm. They can be seen as extended versions of accuracy, a simple metric that computes the fraction of instances for which the correct result is returned.

When using precision and recall, the set of possible labels for a given instance is divided into two subsets, one of which is considered “relevant” for the purposes of the metric. Recall is then computed as the fraction of correct instances among all instances that actually belong to the relevant subset, while precision is the fraction of correct instances among those that the algorithm believes to belong to the relevant subset.



*Notes*

Precision can be seen as a measure of exactness or fidelity, whereas recall is a measure of completeness.

In even simpler terms, a high recall means you haven’t missed anything but you may have a lot of useless results to sift through (which would imply low precision). High precision means that everything returned was a relevant result, but you might not have found all the relevant items (which would imply low recall).

As an example, in an information retrieval scenario, the instances are documents and the task is to return a set of relevant documents given a search term; or equivalently, to assign each document to one of two categories, “relevant” and “not relevant”. In this case, the “relevant” documents are simply those that belong to the “relevant” category. Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

In a classification task, the precision for a class is the number of true positives (*i.e.* the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (*i.e.* the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (*i.e.* the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

Often, there is an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other. For example, an information retrieval system (such as a search engine) can often increase its recall by retrieving more documents, at the cost of increasing number of irrelevant documents retrieved (decreasing precision). Similarly, a classification system for deciding whether or not, say, a fruit is an orange, can achieve high precision by only classifying fruits with the exact right shape and color as oranges, but at the cost of low recall due to the number of false negatives from oranges that did not quite match the specification.

### Information Retrieval Context

In information retrieval contexts, precision and recall are defined in terms of a set of retrieved documents (*e.g.* the list of documents produced by a web search engine for a query) and a set of relevant documents (*e.g.* the list of all documents on the internet that are relevant for a certain topic).

### 10.3 Precision

In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the search:

Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at  $n$  or  $P@n$ .

For example for a text search on a set of documents precision is the number of correct results divided by the number of all returned results.

Precision is also used with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system.



Notes

The meaning and usage of “precision” in the field of Information Retrieval differs from the definition of accuracy and precision within other branches of science and technology.


### 10.4 Recall

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

For example for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned.



Notes

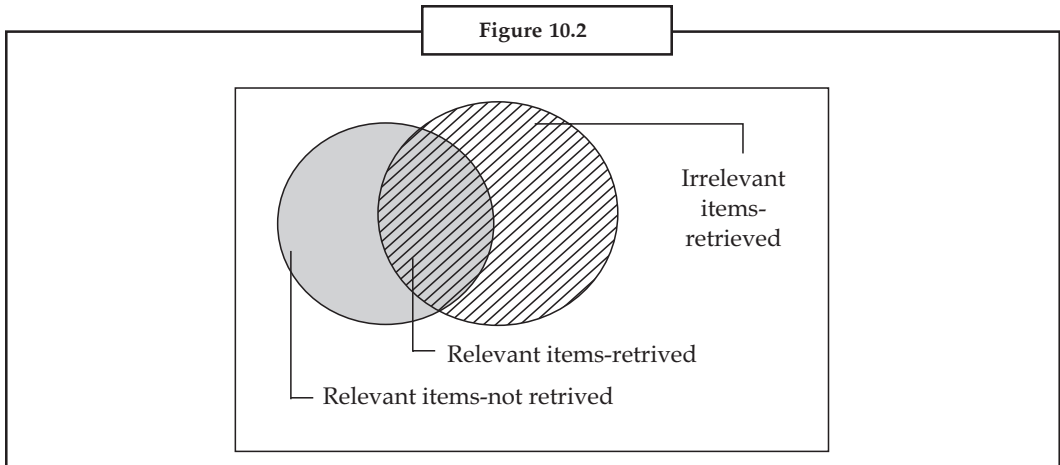
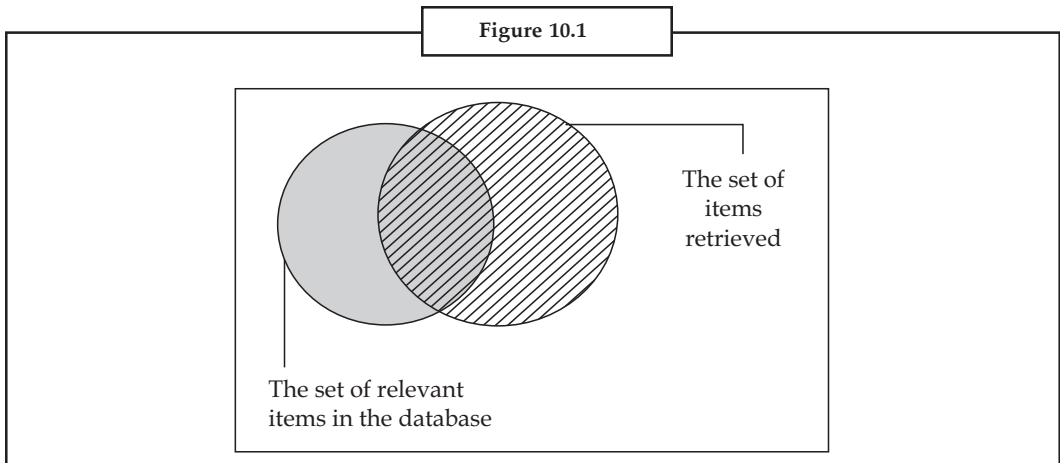
 *Did u know?* In binary classification, recall is called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query.

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

**Precision and recall are the basic measures used in evaluating search strategies**

As shown in the first two figures on the left, these measures assume:

1. There is a set of records in the database which is relevant to the search topic.
2. Records are assumed to be either relevant or irrelevant (these measures do not allow for degrees of relevancy).

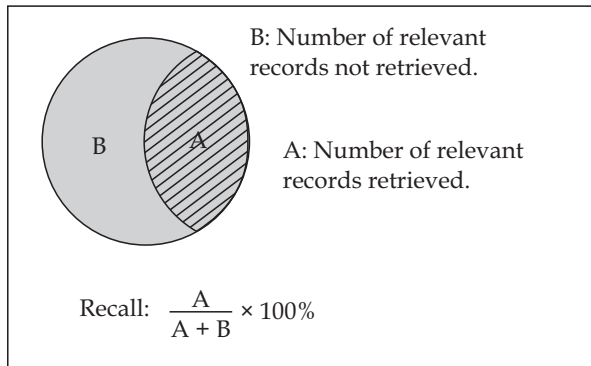


3. The actual retrieval set may not perfectly match the set of relevant records.

**RECALL** is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

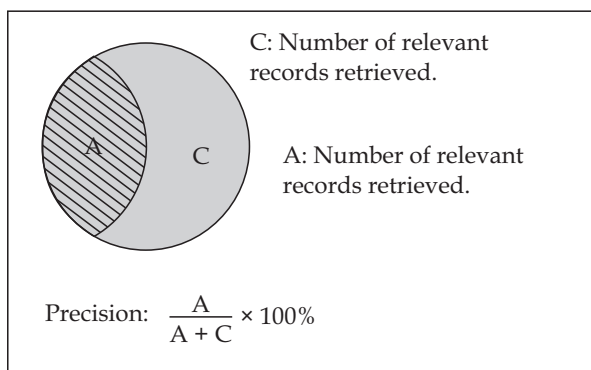
Notes

Figure 10.3



**PRECISION** is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

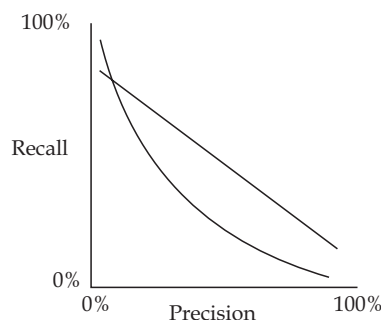
Figure 10.4



Recall and precision are inversely related:

Figure 10.5

As recall ↑ precision ↓  
 conversely:  
 As recall ↓ precision ↑



In the graph above, the two lines may represent the performance of different search systems. While the exact slope of the curve may vary between systems, the general inverse relationship between recall and precision remains.

Notes

**Other problems with Precision and Recall:**

As noted earlier, records must be considered either relevant or irrelevant when calculating precision and recall. Obviously records can exist which are marginally relevant or somewhat irrelevant. Others may be very relevant and others completely irrelevant. This problem is complicated by individual perception: what is relevant to one person may not be relevant to another. Measuring recall is difficult because it is often difficult to know how many relevant records exist in a database. Often recall is estimated by identifying a pool of relevant records and then determining what proportion of the pool the search retrieved. There are several ways of creating a pool of relevant records: one method is to use all the relevant records found from different searches, another is to manually scan several journals to identify a set of relevant papers.

**Precision and Recall are useful measures despite their limitations:**

As abstract ideas, recall and precision are invaluable to the experienced searcher. Knowing the goal of the search — to find everything on a topic, just a few relevant papers, or something in-between — determines what strategies the searcher will use. There are a variety of search techniques which may be used to effect the level recall and precision. A good searcher must be adept at using them. Many of these techniques are discussed in the section on search strategies.

**Self Assessment**

Fill in the blanks:

3. In binary classification, recall is called .....
4. .... is the ratio of the number of relevant records retrieved to the total number of relevant records in the database.
5. .... is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved.
6. Records must be considered either relevant or irrelevant when calculating ..... and .....
7. As abstract ideas, recall and precision are invaluable to the .....

**10.5 Relevance**

In information science and information retrieval, relevance denotes how well a retrieved document or set of documents meets the information need of the user.

**Types**

Relevance most commonly refers to topical relevance or aboutness, *i.e.*, to what extent the topic of a result matches the topic of the query or information need. Relevance can also be interpreted more broadly, referring to generally how “good” a retrieved result is with regard to the information need. The latter definition of relevance, sometimes referred to as user relevance, encompasses topical relevance and possibly other concerns of the user such as timeliness, authority or novelty of the result.

**History**

The concern with the problem of finding relevant information dates back at least to the first publication of scientific journals in 17th Century. The formal study of relevance began in the 20th Century with

the study of what would later be called bibliometrics. In the 1930s and 1940s, S. C. Bradford used the term “relevant” to characterize articles relevant to a subject (cf., Bradford’s law). In the 1950s, the first information retrieval systems emerged, and researchers noted the retrieval of irrelevant articles as a significant concern. In 1958, B. C. Vickery made the concept of relevance explicit in an address at the International Conference on Scientific Information. Since 1958, information scientists have explored and debated definitions of relevance.



Notes

A particular focus of the debate was the distinction between “relevance to a subject” or “topical relevance” and “user relevance”.

## Evaluation

The information retrieval community has emphasized the use of test collections and benchmark tasks to measure topical relevance, starting with the Cranfield Experiments of the early 1960s and culminating in the TREC evaluations that continue to this day as the main evaluation framework for information retrieval research. In order to evaluate how well an information retrieval system retrieved topically relevant results, the relevance of retrieved results must be quantified. In Cranfield-style evaluations, this typically involves assigning a relevance level to each retrieved result, a process known as relevance assessment. Relevance levels can be binary (indicating a result is relevant or that it is not relevant), or graded (indicating results have a varying degree of match between the topic of the result and the information need).

Once relevance levels have been assigned to the retrieved results, information retrieval performance measures can be used to assess the quality of a retrieval system’s output. In contrast to this focus solely on topical relevance, the information science community has emphasized user studies that consider user relevance. These studies often focus on aspects of human-computer interaction.

## Clustering and Relevance

The cluster hypothesis, proposed by C. J. van Rijsbergen in 1979, asserts that two documents that are similar to each other have a high likelihood of being relevant to the same information need. With respect to the embedding similarity space, the cluster hypothesis can be interpreted globally or locally. The global interpretation assumes that there exist some fixed set of underlying topics derived from inter-document similarity. These global clusters or their representatives can then be used to relate relevance of two documents (*e.g.* two documents in the same cluster should both be relevant to the same request). Methods in this spirit include.

### Cluster-based Information Retrieval

Cluster-based document expansion such as latent semantic analysis or its language modelling equivalents. It is important to ensure that clusters—either in isolation or combination – successfully model the set of possible relevant documents.


A second interpretation, most notably advanced by Ellen Voorhees, focuses on the local relationships between documents. The local interpretation avoids having to model the number or size of clusters in the collection and allow relevance at multiple scales. Methods in this spirit include, multiple cluster retrieval spreading activation and relevance propagation methods.

Notes

**Epistemological Issues**

Most research about relevance in information retrieval in recent years have implicitly assumed that the users evaluation of the output a given system should be used to increase “relevance” output. An alternative strategy would be to use journal impact factor to rank output and thus base relevance on expert evaluations.

Other strategies, such as including diversity of the search results, may be used as well. The important thing to recognize is, however, that relevance is fundamentally a question of epistemology, not psychology. (Peoples’ psychology reflects certain epistemological influences).

|   |  |
|---|--|
|  | <p><i>Task</i> Give a brief report on—Theory and Implementation of Information Storage and Retrieval system.</p> |
|---|--|

**10.6 Online Searching Basics; Library Databases; Web Sites**

Effective searching in web-based databases depends on how well your search is formulated. Usually at least two types of searches can be performed: subject and keyword. Most online searches tend to default to the keyword variety. Database users have to read onscreen directions to see how, or if, subject and keyword searches are differentiated.

When you type in a keyword search in a database, you are instructing the search mechanism to analyze every eligible document in its inventory, and return to you only the ones that include all the relevant terms you typed in. Since the computer will only retrieve articles containing all of the significant terms that you type in, good researchers must choose their terms well. Which “key” words will the authors of documents be likely to use when writing on a certain topic? What terms, if any, have database designers decided to formalize into an official list of searchable key words? It pays to take a few minutes before beginning your search to organize it. Read onscreen directions to help optimize your research strategies. Using too many terms, or too few, will too often retrieve nothing useful.

Analyze your search strategy frequently. As necessary, narrow the search by deleting extraneous terms, or broaden it by adding appropriate ones, to get the best results. Using synonyms or related terms in one or a series of keyword inquiries will help to make searches more effective.

To perform a subject search, database searchers usually have to use another interface of the database other than the keyword approach. This subject connection to database contents would allow searches by concept rather than by individual words.

**Keyword Searching and Boolean Operators**

If you were writing a paper asserting, for example, that “marijuana use in California should not be legal”, you might achieve good results with a keyword search. Keyword searching allows you to “customize” your searches by combining more terms than a less-flexible subject search could do. Use only the key words (important words; words that have substance and meaning) in the thesis (such as marijuana, California, and legal in the example above) and type them in. Discard meaningless words such as in, should, not, and be, since they are unimportant for retrieving information in the search. Such insignificant terms are called stopwords.



*Did u know?* Keyword searches that combine two or more terms imply a connection between the terms. These connections are expressed in what is called Boolean terms or operators. The three Boolean operators for keyword searching are AND, OR, and NOT (or AND NOT).

Notes

Using AND and NOT will narrow search results; OR will expand the results. The sample thesis above assumes an AND connection between terms: when you type in the search words, you're asking for all the documents with the words marijuana AND California AND legal in them. Many databases and some search engines default into an AND search.

The sample search above could be broadened by using the OR option. If you used, for instance, the words "legal OR lawful", you would add another variable, a word meaning almost the same thing as the word "legal". Since this is a word search, the computer would now pick up articles that might have the word "lawful" in place of "legal".

Similarly, the NOT option changes the parameters of a search, cutting down on the number of articles retrieved. The search "marijuana AND legal AND California NOT San Francisco" would omit articles that mentioned San Francisco when discussing, for example, legal use of marijuana in California.

Try to keep your search to only 3 to 4 relevant terms at a time, with their proper Boolean connections. If you need to use more terms, try breaking your unwieldy single search into two or more smaller searches.

## 10.7 Summary

- An information storage and retrieval system (ISRS) is a network with a built-in user interface that facilitates the creation, searching, and modification of stored data.
- An ISRS is typically a peer-to-peer ( P2P ) network operated and maintained by private individuals or independent organizations.
- A significant difference between an ISRS and a database management system (DBMS ) is the fact that an ISRS is intended for general public use, while a DBMS is likely to be proprietary, with access privileges restricted to authorized entities.
- An information retrieval process begins when a user enters a query into the system.
- Precision and recall are two widely used metrics for evaluating the correctness of a pattern recognition algorithm.
- Precision can be seen as a measure of exactness or fidelity, whereas recall is a measure of completeness.
- In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the search.
- Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.
- Precision and recall are the basic measures used in evaluating search strategies.
- In information science and information retrieval, relevance denotes how well a retrieved document or set of documents meets the information need of the user.
- Effective searching in web-based databases depends on how well your search is formulated.

Notes

## 10.8 Keywords

**ISRS** : Information storage and retrieval system, is a network with a built-in user interface.

**Precision** : Is a measure of exactness.

**Recall** : Is a measure of completeness.

**DBMS** : Database Management System.

## 10.9 Review Questions

1. Write the characteristics of an ISRS.
2. Describe information retrieval process.
3. Where precision and recall are mostly used?
4. How can you evaluate information retrieval system?
5. Explain cluster-based information retrieval.

## **Answers: Self Assessment**

1. (b), 2. (a)
3. Sensitivity
4. Recall
5. Precision
6. Precision and recall
7. Experienced searcher.

## 10.10 Further Readings



*Books*

Aitchinson, J and Gilchrist, A: *Thesaurus construction*. 2nd ed. London: Aslib, 1987.

Deegan. M. and Simon Tanner. *Digital futures*. London. LA, 2002.

Maltby, A., ed. *Sayer's manual of classification for libraries*. 5th. Ed. London: Andre Deutsch, 1975.



*Online links*

<http://en.wikipedia.org/wiki/>

<http://thenoisychannel.com/2009/03/17/precision-and-recall/>

[http://www.columbia.edu/cu/lweb/help/clio/boolean\\_operators.html](http://www.columbia.edu/cu/lweb/help/clio/boolean_operators.html)

## Unit 11: Online Searching: Library Databases

### CONTENTS

Objectives

Introduction

11.1 Search Strategies

11.2 Summary

11.3 Keywords

11.4 Review Questions

11.5 Further Readings

### Objectives

After studying this unit, you will be able to:

- Explain search strategies.

### Introduction

The fundamentals of Boolean searching, the basics of using the OPAC, and an introduction to using print indexes and the Periodicals Holdings List have been explored in previous pages. And several of the searching principles learned in prior readings can be applied to searching Ebscohost MasterFile Premier. When you are at SMC you have on-campus access to Ebscohost without having to log in with your SMC username and password. However, for off-campus access to Ebscohost, and to other Library online resources, you will need to use the username and password provided by your SMC student account.

Ebscohost is one of SMC Library's electronic periodical indexes, used in order to find appropriate articles from magazines, newspapers, or journals. Ebscohost is a web-based database to which the SMC Library subscribes. This online general index comes to us via the World Wide Web, but it is a proprietary database published by the Ebsco company. Since Ebscohost is a general index, it contains articles on many subject areas instead of just one. In addition, it offers keyword, customizable searching and the full text of many of its articles, which come from magazines, journals, newspapers, and other sources.

### 11.1 Search Strategies


#### **Search Strategies – Keyword Searching**

Let's examine some search strategies you can use to get more relevant results from your library research. Keyword searching is the default in many of the library databases. So what does this mean? It means the database will look "everywhere" for your search terms, including: titles, author names, summaries, and sometimes even the full text of an article, dissertation, or book.



Notes

One of the most beneficial things you can do to improve your keyword searching results is to plan out your search. Here's how: Write down the keywords or search terms you plan to use.



*Did u know?* Brainstorm possible synonyms for your search terms (*e.g.*, community organizing vs. grassroots movements), and be sure to incorporate those into your search.

Identify terms that are most specific to the concepts that you're researching (*e.g.*, substance abuse vs. alcoholism).

When you find relevant articles, examine them for potential (new) keywords.

Keep notes on what worked – and what didn't – so you don't repeat the same searches over and over again.

Keyword searching ignores context unless we tell it otherwise. Thus keyword searching works well if you're using very specific terminology, if you want to search full text, or if you just want to perform a broad search on a topic. It can also work well if you use it to construct complex search strings;

#### Top Search Mistakes – Database Mismatch

The Walden librarians are often asked "what are the most common mistakes people make when searching?"


Well, that's a surprisingly hard question to answer!

Library databases are complicated, there are many different types of resources, and the internet has very little structure. There are lots of places where things can go wrong. So, we're starting a series here on the blog that will look at some of the common problems searchers face, and how you can solve them.

Today we'll talk about database mismatch.

This is when the information you need exists, but you aren't in the right spot to find it. With over 60 databases in the Walden Library, this can happen easily and often. It even happens to librarians (almost daily).

So, how do you make sure you're picking the right database?



*Task* Write down the keywords or search terms.

Become familiar with the databases in your subject area. Read through the descriptions and take a few minutes to explore each database. Does it have popular resources or scholarly ones? Both? Does it have books, which provide more general information, or articles, which cover very specific topics? Is it a large database or a relatively small one? Does it have specialized content, such as reviews, videos, or SWOT analyses? If you have a basic understanding of what's available to you, you'll be better at picking a database the next time you need to use one.

Decide if your topic is cross-disciplinary or multi-disciplinary. Some topics are discussed by researchers in different fields. You may need to try databases from other fields or try a search in a multidisciplinary database such as Academic Search Premier or ProQuest Central.

Try more than one database. Many subject areas have several databases that contain similar content. It's almost impossible to know which one is better for your particular research topic until you try them. And you may find that you need articles from several different databases. You can do the same search (or similar searches) in several different databases; librarians do this all the time!



*Notes* If you think you may be suffering from database mismatch, and you aren't sure which database you need, you can always ask a librarian.

Notes

## Self Assessment

State whether the following statements are true or false:

1. Boolean searching is an introduction to using print indexes and the periodicals holdings list.
2. Ebscohost is one of SMC library database periodical indexes.
3. Keyword searching is the default in many of the SMC library.
4. Many subject areas have several databases that contains similar content.

## 11.2 Summary

- Let's examine some search strategies you can use to get more relevant results from your library research. Keyword searching is the default in many of the library databases.
- Keyword searching ignores context unless we tell it otherwise. Thus keyword searching works well if you're using very specific terminology.
- It's almost impossible to know which one is better for your particular research topic until you try them.
- One of the most beneficial things you can do to improve your keyword searching results is to plan out your search.

## 11.3 Keywords

*Ebscohost Searches* : Are not case sensitive and can be done in two major ways, by subject or by several variations on a keyword search.

*Keyword Searching* : To improve keyword searching result plan out your search.

## 11.4 Review Questions

1. Keyword searching is the default. Explain.

## Answers: Self Assessment

1. True
2. False
3. False
4. True

Notes

**11.5 Further Readings**



*Books*

Oddy, P. *Future libraries, future catalogs*. London: LA, 1996.

Cooke, A : *A guide to finding quality Information on the Internet*. 2nd Edition. London: Facet Publishing, 2001.

Maltby, A., ed. *Sayer's manual of classification for libraries*. 5th. Ed. London: Andre Deutsch, 1975.



*Online links*

<http://library.uaf.edu/goldmine-keyword-searching>  
[www.wikipedia.com](http://www.wikipedia.com).

## Unit 12: Vocabulary Control

### CONTENTS

Objectives

Introduction

12.1 Methodology

12.2 Library Science

12.3 Indexing Languages

12.4 Trends and Development

12.5 Summary

12.6 Keywords

12.7 Review Questions

12.8 Further Readings

### Objectives

After studying this unit, you will be able to:

- Define methodology and Library Science
- Explain indexing language
- Describe trends and development.

### Introduction

Vocabulary control is used to improve the effectiveness of information storage and retrieval systems, Web navigation systems, and other environments that seek to both identify and locate desired content via some sort of description using language. The primary purpose of vocabulary control is to achieve consistency in the description of content objects and to facilitate retrieval. Controlled vocabularies provide a way to organize knowledge for subsequent retrieval.

They are used in subject indexing schemes, subject headings, thesauri and taxonomies. Controlled vocabulary schemes mandate the use of predefined, authorised terms that have been preselected by the designer of the vocabulary, in contrast to natural language vocabularies, where there is no restriction on the vocabulary.


There are four important principles of vocabulary control that guide their design and development.

These are:

- eliminating ambiguity
- controlling synonyms
- establishing relationships among terms where appropriate
- testing and validation of terms.

**Notes**

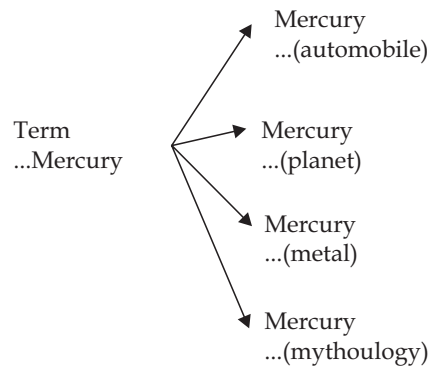
A major goal of vocabulary control is to ensure that each distinct concept refers to a unique linguistic form. These types of linguistic relationships should be controlled or regularized so that information or content that is provided to a user is not spread across the system under multiple access points, but is gathered together in one place. Eliminating ambiguity and compensating for synonymy through vocabulary control assures that each term has only one meaning and that only one term may be used to represent a given concept or entity.



*Task* Why is vocabulary control necessary in organizations?

**Ambiguity**

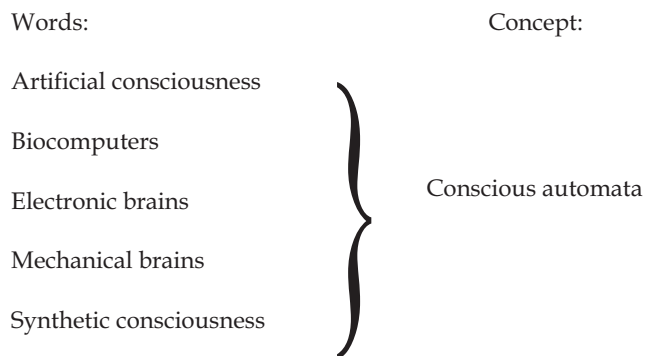
Ambiguity occurs in natural language when a word or phrase (a homograph or polyseme) has more than one meaning. Below figure provides an example and shows how a single word may be used to represent multiple, very different concepts.



A controlled vocabulary must compensate for the problems caused by ambiguity by ensuring that each term has one and only one meaning.

**Synonymy**

A different problem occurs when a concept can be represented by two or more synonymous or nearly synonymous words or phrases. This is called synonymy. This means that desired content may be scattered around an information space or database because it can be described by different but equivalent terminology. Below figure illustrates this case:



A controlled vocabulary must compensate for the problems caused by synonymy by ensuring that each concept is represented by a single preferred term. The vocabulary should list the other synonyms and variants as non-preferred terms with USE references to the preferred term.

There are other types of “equivalent” terms besides synonyms which require vocabulary control.

### Semantic Relationships

Various types of semantic relationships may be identified among the terms in a controlled vocabulary. These include equality relationships, hierarchical relationships and associative relationships which may be defined as required for a particular application.

### Using Warrant to Select Terms

The process of selecting terms for inclusion in controlled vocabularies involves consulting various sources of words and phrases as well as criteria based on:

- the natural language used to describe content objects (literary warrant),
- the language of users (user warrant), and
- the needs and priorities of the organization (organizational warrant).

### Literary Warrant

Assessing literary warrant involves consulting reference sources such as dictionaries or textbooks as well as existing vocabularies. The word or phrases chosen should match as closely as possible the prevailing descriptions for the concept in the literature.

### Organizational Warrant

Determining organization warrant requires identifying the form or forms of terms that are preferred by the organization or organizations that will use the controlled vocabulary.

### User Warrant

Creating lists of potential terms to enhance completeness of the vocabulary.

- Organizing candidate terms into broad categories to determine what categories users prefer and what they should be called.
- Placing candidate terms into a tentative set of broad categories to validate categories that have been created.
- Reviewing drafts of the vocabulary to add missing terms, delete terms that are incorrect or obsolete, create more useful term forms, and identify and correct missing and/or incorrect relationships among terms.

### Self Assessment

Fill in the blanks:

1. There are ..... principles of vocabulary control that guide their design and development.
2. A major goal of vocabulary control is to ensure that each distinct concept refers to a ..... .
3. .... in natural language when a word or phrase has more than one meaning.

Notes

4. Various types of semantic relationships may be identified among the terms in a ..... .
5. Accessing literary warrant involves consulting reference sources such as dictionaries or textbooks as well as ..... .

## 12.1 Methodology

In library and information science controlled vocabulary is a carefully selected list of words and phrases, which are used to tag units of information (document or work) so that they may be more easily retrieved by a search. Controlled vocabularies solve the problems of homographs, synonyms and polysemes by a bijection between concepts and authorized terms. In short, controlled vocabularies reduce ambiguity inherent in normal human languages where the same concept can be given different names and ensure consistency.

For example, in the Library of Congress Subject Headings (a subject heading system that uses a controlled vocabulary), authorized terms — subject headings in this case — have to be chosen to handle choices between variant spellings of the same concept (American versus British), choice among scientific and popular terms (Cockroaches versus *Periplaneta americana*), and choices between synonyms (automobile versus cars), among other difficult issues.

Choices of authorized terms are based on the principles of user warrant (what terms users are likely to use), literary warrant (what terms are generally used in the literature and documents), and structural warrant (terms chosen by considering the structure, scope of the controlled vocabulary).

Controlled vocabularies also typically handle the problem of homographs, with qualifiers. For example, the term “pool” has to be qualified to refer to either swimming pool, or the game pool to ensure that each authorized term or heading refers to only one concept.



*Did u know?* There are two main kinds of controlled vocabulary tools used in libraries: subject headings and thesauri.

While the differences between the two are diminishing, there are still some minor differences.

Historically subject headings were designed to describe books in library catalogues by cataloguers while thesauri were used by indexers to apply index terms to documents and articles. Subject headings tend to be broader in scope describing whole books, while thesauri tend to be more specialized covering very specific disciplines. Also because of the card catalogue system, subject headings tend to have terms that are in indirect order (though with the rise of automated systems this is being removed), while thesaurus terms are always in direct order.

Subject headings also tend to use more pre-coordination of terms such that the designer of the controlled vocabulary will combine various concepts together to form one authorized subject heading. (e.g., children and terrorism) while thesauri tend to use singular direct terms. Lastly thesauri list not only equivalent terms but also narrower, broader terms and related terms among various authorized and non-authorized terms, while historically most subject headings did not.



*Notes* The Library of Congress Subject Heading itself did not have much syndetic structure until 1943, and it was not until 1985 when it began to adopt the thesauri type term “Broader term” and “Narrow term”.

The terms are chosen and organized by trained professionals (including librarians and information scientists) who possess expertise in the subject area. Controlled vocabulary terms can accurately

describe what a given document is actually about, even if the terms themselves do not occur within the document's text. Well known subject heading systems include the Library of Congress system, MeSH, and Sears. Well known thesauri include the Art and Architecture Thesaurus and the ERIC Thesaurus.

Choosing authorized terms to be used is a tricky business, besides the areas already considered above, the designer has to consider the specificity of the term chosen, whether to use direct entry, inter consistency and stability of the language. Lastly the amount of pre-co-ordinate (in which case the degree of enumeration versus synthesis becomes an issue) and post co-ordinate in the system is another important issue.

Controlled vocabulary elements (terms/phrases) employed as tags, to aid in the content identification process of documents, or other information system entities (e.g. DBMS, Web Services) qualifies as metadata.

## **12.2 Library Science**

This functional analysis provides a path away from the arguments that used to characterize information retrieval in the post-World War II period. Any relatively complete functional analysis of information storage and retrieval should provide a basis for the mapping of research and development activities. We suggest that the following components that have been and remain of most interest in Library Science:

### **The Use of Human Intermediaries in Query Development**

An emphasis on incorporating external knowledge (expert descriptive cataloguing, classification, and assignment of subject headings) into the representation (catalogue record).

Vocabulary control (alias authority control) in creating representations, in syndetic structure, and in query development.

In online catalogues, minimally a two-stage approach (a Boolean operation to partition the Representations, followed by the alphabetization of the Retrieved Set) and commonly, a three stage approach (the two-stage approach preceded by a search of the Searchable Index only, for feedback).

The activities generally referred to information retrieval research have historically tended to emphasize:

In storage, the use of algorithmic alternatives to human expertise in creating representations and indexes. Good examples are automatic keyword indexes (e.g. KWIC) and the generation of vector space representations of documents' terms.



*Notes*

In retrieval, the use of highly elaborate partitioning (retrieval) and transforming algorithms leading to the strict ranking of a set of retrieved documents.

Others might prefer to nominate other techniques as being characteristic of these streams of research and development, but any realistic mapping on to a general framework of information storage and retrieval theory is likely to reveal how complementary rather than contradictory these interests are. There is a difference in emphasis: the former tending to emphasize quality of data, consistency, and expert human intervention; the latter, exploring efficient algorithmic approaches to large volumes of data. Neither approach alone can provide a complete approach to selection systems in theory or in practice.



### **12.3 Indexing Languages**

There are three main types of indexing languages.

- Controlled indexing language: Only approved terms can be used by the indexer to describe the document.
- Natural language indexing language: Any term from the document in question can be used to describe the document.
- Free indexing language: Any term (not only from the document) can be used to describe the document.

When indexing a document, the indexer also has to choose the level of indexing exhaustivity, the level of detail in which the document is described. For example using low indexing exhaustivity, minor aspects of the work will not be described with index terms. In general the higher the indexing exhaustivity, the more terms indexed for each document.

Controlled vocabularies are often claimed to improve the accuracy of free text searching, such as to reduce irrelevant items in the retrieval list. These irrelevant items (false positives) are often caused by the inherent ambiguity of natural language. Take the English word football for example. Football is the name given to a number of different team sports. Worldwide the most popular of these team sports is Association football, which also happens to be called soccer in several countries.

Compared to free text searching, the use of a controlled vocabulary can dramatically increase the performance of an information retrieval system, if performance is measured by precision (the percentage of documents in the retrieval list that are actually relevant to the search topic).

In some cases controlled vocabulary can enhance recall as well, because unlike natural language schemes, once the correct authorized term is searched, you don't need to worry about searching for other terms that might be synonyms of that term.

However, a controlled vocabulary search may also lead to unsatisfactory recall, in that it will fail to retrieve some documents that are actually relevant to the search question.

This is particularly problematic when the search question involves terms that are sufficiently tangential to the subject area such that the indexer might have decided to tag it using a different term (but the searcher might consider the same). Essentially, this can be avoided only by an experienced user of controlled vocabulary whose understanding of the vocabulary coincides with the way it is used by the indexer.

Controlled vocabularies are also quickly out-dated and in fast developing fields of knowledge, the authorized terms available might not be available if they are not updated regularly. Even in the best case scenario, controlled language is often not as specific as using the words of the text itself. Indexers trying to choose the appropriate index terms might misinterpret the author, while a free text search is in no danger of doing so, because it uses the author's own words.

The use of controlled vocabularies can be costly compared to free text searches because human experts or expensive automated systems are necessary to index each entry. Furthermore, the user has to be familiar with the controlled vocabulary scheme to make best use of the system. But as already mentioned, the control of synonyms, homographs can help increase precision.

Numerous methodologies have been developed to assist in the creation of controlled vocabularies, including faceted classification, which enables a given data record or document to be described in multiple ways.

### **Types of Controlled Vocabularies**

Currier (2005) distinguish between the following kinds of controlled vocabularies to which we added metadata schemes.

**Flat list**

A simple flat list of terms

**Glossary**

An alphabetical list of terms with some explanation

**Subject headings list**

See subject heading. A systematic list of subject headings like the ones used for library catalogues. A subject header provides one of the access points to information.

**Taxonomy**

In a wide sense almost any kind of well defined list of terms

In one narrow sense, a mono-hierarchical classification of terms, *i.e.*, a child term inherits in principle the properties of the parent term. *E.g.* controlled vocabularies are a kind of vocabularies, or XHTML is a kind of XML application which is a kind of formalism for defining a formal grammar. This is the equivalent of a kind of typology.

In another narrow sense: “controlled vocabulary in which concepts are represented by preferred terms, formally organized so that paradigmatic relationships between the concepts are made explicit, and the preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms” (Willpower Information, retrieved 15:08, 27 February 2009 (UTC)).



*Did u know?* One also could define a taxonomy with non-hierarchical relationships, but we would rather call these “thesauri”.

**12.4 Trends and Development**

Controlled vocabularies, such as the Library of Congress Subject Headings, are an essential component of bibliography, the study and classification of books. They were initially developed in library and information science. In the 1950s, government agencies began to develop controlled vocabularies for the burgeoning journal literature in specialized fields; an example is the Medical Subject Headings (MeSH) developed by the U.S. National Library of Medicine. Subsequently, for-profit firms (called Abstracting and indexing services) emerged to index the fast-growing literature in every field of knowledge.

In the 1960s, an online bibliographic database industry developed based on dialup X.25 networking. These services were seldom made available to the public because they were difficult to use; specialist librarians called search intermediaries handled the searching job.



*Did u know?* In the 1980s, the first full text databases appeared; these databases contain the full text of the index articles as well as the bibliographic information.

Online bibliographic databases have migrated to the Internet and are now publicly available; however, most are proprietary and can be expensive to use. Students enrolled in colleges and universities may be able to access some of these services without charge; some of these services may be accessible without charge at a public library.

In large organizations, controlled vocabularies may be introduced to improve technical communication. The use of controlled vocabulary ensures that everyone is using the same word to mean the same thing. This consistency of terms is one of the most important concepts in technical

Notes

## Notes

writing and knowledge management, where effort is expended to use the same word throughout a document or organization instead of slightly different ones to refer to the same thing.

Web searching could be dramatically improved by the development of a controlled vocabulary for describing Web pages; the use of such a vocabulary could culminate in a Semantic Web, in which the content of Web pages is described using a machine-readable metadata scheme. One of the first proposals for such a scheme is the Dublin Core Initiative. An example of a controlled vocabulary which is usable for indexing web pages is PSH.

It is unlikely that a single metadata scheme will ever succeed in describing the content of the entire Web. To create a Semantic Web, it may be necessary to draw from two or more metadata systems to describe a Web page's contents.



Notes

The eXchangeable Faceted Metadata Language (XFML) is designed to enable controlled vocabulary creators to publish and share metadata systems. XFML is designed on faceted classification principles.

## Self Assessment

Multiple Choice Questions:

6. .... is the name given to a number of different team sports.  
(a) Cricket (b) Football  
(c) Hockey (d) None of these.
7. Controlled vocabularies also typically handle the problem of .....  
(a) Synonyms (b) Polysemes  
(c) Homographs (d) None of these
8. In the ....., government agencies began to develop controlled vocabularies.  
(a) 1940 (b) 1960  
(c) 1950 (d) None of these.

## 12.5 Summary

- Controlled vocabularies, such as the Library of Congress Subject Headings, are an essential component of bibliography, the study and classification of books.
- In the 1960s, an online bibliographic database industry developed based on dialup X.25 networking.
- Controlled vocabularies are also quickly out-dated and in fast developing fields of knowledge, the authorized terms available might not be available if they are not updated regularly.
- Vocabulary control (alias authority control) in creating representations, in syndetic structure, and in query development.
- Subject headings also tend to use more pre-coordination of terms such that the designer of the controlled vocabulary will combine various concepts together to form one authorized subject heading.
- Controlled vocabularies also typically handle the problem of homographs, with qualifiers.

- Vocabulary control is used to improve the effectiveness of information storage and retrieval systems, Web navigation systems, and other environments that seek to both identify and locate desired content via some sort of description using language.
- In large organizations, controlled vocabularies may be introduced to improve technical communication.

## 12.6 Keywords

**Vocabulary Control** : Is used to improve the effectiveness of information storage and retrieval systems.

**Ambiguity** : Occurs when a word or phrase has more than one meaning.

## 12.7 Review Questions

1. What is the use of Vocabulary control?
2. Write the four important principles of vocabulary control.
3. In the 1960s, an online bibliographic database industry developed. Explain.

## **Answers: Self Assessment**

1. Four important
2. Unique linguistic form
3. Ambiguity occurs
4. Controlled vocabulary
5. Existing vocabularies
6. (b), 7. (c), 8. (c)

## 12.8 Further Readings



Books

Maltby, A., ed. *Sayer's manual of classification for libraries*. 5th. Ed. London: Andre Deutsch, 1975.

Aitchinson, J and Gilchrist, A: *Thesaurus construction*. 2nd ed. London: Aslib, 1987.

Deegan. M and Simon Tanner. *Digital futures*. London. LA, 2002.



Online links

[nkos.slis.kent.edu/2005workshop/z3919.ppt](http://nkos.slis.kent.edu/2005workshop/z3919.ppt)

<http://www.palgrave-journals.com/dam/journal/v6/n5/full/dam201029a.html>

## Unit 13: Subject Headings

### CONTENTS

Objectives

Introduction

- 13.1 Sears List of Subject Heading
- 13.2 Library of Congress Subject Headings
- 13.3 LCSH Policy Issues
- 13.4 Medical Subject Headings (MeSH)
- 13.5 Summary
- 13.6 Keywords
- 13.7 Review Questions
- 13.8 Further Readings

### Objectives

After studying this unit, you will be able to:

- Define sears list of subject heading
- Explain library of congress subject heading
- Describe medical subject headings.

### Introduction

Access problems such as these have, over time, spurred the use of subject headings that indicate the topics covered by materials in the library. Because consistency is an important issue when providing access to information in a library catalogue, there have been a small number of very comprehensive and regularly updated subject lists developed for use in libraries. The two most commonly used lists for public, academic and school libraries are *Sears List of Subject Headings* and *Library of Congress Subject Headings*. These lists were developed to try to cover most known topics in a consistent manner, enabling libraries to provide access to materials on similar subjects less than one consistent term.

Specialized lists have also been developed for the use of libraries that deal only with areas such as medical or agricultural information. This type of specialized list contains a more detailed breakdown of the information area, appropriate to the more detailed level of coverage in special libraries. These specialized lists have usually been developed under the auspices of a knowledgeable, respected institution, such as the *National Library of Medicine* or the *National Agricultural Library*, or by large organizations or database creators who are dealing with a specific topic.

### 13.1 Sears List of Subject Heading

Notes

The 19th edition of the Sears List of Subject Headings melds the traditional with the new. New headings and categories that address contemporary issues join the traditional streamlined approach to subject headings found in Sears. The combination continues to make Sears the choice of most school and public libraries. Sears 19 breaks new ground in helping both teacher-librarians and public librarians enhance records by adding subject headings and metadata that will enhance the cataloguing description of the item and thus provide contextual reference for both identification and content analysis.



*Did u know?* Several notable 19th edition features include over 440 completely new subject headings and two totally new categories — “Islam” and “Graphic Novels.”

Expanded coverage of headings in the science/technology, lifestyle/entertainment, politics/world affairs, and literature/arts categories are also noteworthy additions.

Once again, Joseph Miller and Barbara Bristow maintain the primary mission of Sears by using simplified vocabulary specifically geared for the school and small public library setting and by retaining an awareness of the descriptive searching and cataloguing needs from an educator’s perspective. In addition, both Miller and Bristow take time to focus in-depth on the mechanics of creating Sears subject headings. They emphasize that there exist certain “Sears-specific “ subject heading elements and developmental strategies which apply specifically to identifying and/or creating Sears headings to use in metadata describing materials from both a “child-centered” and “K-12 curriculum perspective.”

For example, Sears 19 provides broad conceptual overviews and detailed instructions for the four types of Sears subject headings — topical, form, geographic, and proper names. Both the forms of the headings (*i.e.*, single words, compound headings, phrase headings, etc.) and the rules for creating and adding subdivisions (*i.e.*, chronological, geographic, form, etc.) — as well as their order of presentation — are included in the “Principles of the Sears List” chapter.

While Sears does a fine job of providing comprehensive subject heading coverage, there are some subjects that are difficult to address descriptively using the Sears structure, notably descriptions for biographies (both individual and collective), nationality descriptions, literary forms (including criticisms and collections), government policy, etc.

Sears provides suggestions for working with its headings in relation to these material forms and concepts. However, for those items and topics that cannot be adequately described using the Sears List of Subject Headings (or other “controlled/authority file-monitored “ subject thesauri such as the Library of Congress Subject Headings — LCSH), strategies exist for incorporating local or “uncontrolled/non-authority file-monitored” headings into metadata records created using the MARC metadata standard. One of these strategies is to enter these headings into the MARC field tag 653 – Uncontrolled Heading.

Another great improvement to the Sears 19th ed. is the streamlining of several subject headings that may have evolved or needed structural modification to clarify their meaning. For example, the heading Stereotype (Psychology) was discontinued and replaced with Stereotype (Social Psychology). Likewise, the heading Libraries and the Elderly is now Elderly — Library Services.

Of course, the structural flexibility of Sears remains. Users can create a variety of authorized headings based on one template heading or guideline, such as adding specific breeds of dogs “as needed.” While some may view structural flexibility as an advantage, others consider this aspect of Sears its “missing link,” desiring more concise entry rules.

To address those criticisms, Sears includes some specific guidelines in the aforementioned “Principles of the Sears List” chapter. Once again, this chapter can be used as a pocket catalogueer’s guide. It

**Notes**

provides a background on Charles A. Cutter’s specific and direct entry—the idea that if a specific subject exists for an entry, it should be added. Additionally, the “Principles of the Sears List” chapter explains how the teacher-librarian may create a heading according to guidelines if the desired heading does not exist. Furthermore, the inclusion of suggested Dewey numbers for each subject heading continues in Sears 19, providing yet another aid for the school librarian to use when evaluating the catalogue for both location and contextual usefulness/accuracy.

**Features “Principles of the Sears List”**

“Principles of the Sears List’ is one of the reasons why all institutions that teach subject analysis and all libraries that have untrained technical services staff should buy Sears 19, even if the library assigns Library of Congress subject headings. Sears’ explanation of subject headings is written in clear English and is an excellent description of the function and construction of subject headings. Libraries that assign Library of Congress subject headings can easily adapt the ‘Principles of the Sears List’ to refer to the LC subject list because the fundamental principles of subject analysis are the same.”—Technicalities.

Sears had a long career as a cataloguer and bibliographer at a variety of libraries (Bryn Mawr College, University of Minnesota, New York Public Library), before she joined the publishing company H. W. Wilson Company in 1923 to publish her List of Subject Headings for Small Libraries. The book provides a list of subject headings for small libraries to use in lieu of Library of Congress Subject Headings. Library of Congress headings are often not as useful for small libraries because they are too detailed. Sears’ List of Subject Headings also offers small libraries guidance on how to create their own new subject headings consistently when necessary.

In order to create her subject headings, Sears consulted small and medium sized libraries throughout the country to discern patterns of usage. She then developed her own system, based in part on the Library of Congress Subject Headings, but with a simplified subject vocabulary. In Sears’ system, common terms are much preferred over scientific and technical terms. Her system also allowed individual libraries the authority to create their own subject headings. The Sears model is not meant to serve as a standardized bridge for union catalogs, but rather as a model “for the creation of headings as needed”.

|  |   |
|--|---|
| <br><i>Task</i> | Classify the term sears list subject headings and principles of the sears list. |
|--|---|

Like the Library of Congress Subject Headings, Sears’ system is a subject list arranged in alphabetical order, making use of overarching subject categories and hierarchical subject subdivisions. However, Sears’ headings favor natural language. Her headings make use of only four types of headings: topical, form, geographic, and proper names. She also tended to convert inverted headings into direct entries.

In the third edition of the book (1933), Ms. Sears added a section called, “Practical Suggestions for the Beginner in Subject Heading Work”. These “Principles of the Sears List” were eventually published as a separate document and became a widely used teaching tool for library schools. In subsequent editions of the List, Sears’ subject headings were also linked to appropriate Dewey Decimal numbers.

In addition to creating the List, Sears edited the Standard Catalog for Public Libraries of the American Library Association, and an edition of the Standard Catalog for High School Libraries. She eventually left H.W. Wilson to teach at Columbia University’s School of Library Science, where she started the first graduate course in cataloging. Sears also remained an active participant in the American Library Association and the New York Library Association. After her death in 1933 at age 60, the book was eventually renamed in her honor to The Sears List of Subject Headings. The List is currently in its 20th edition.

## Principles

## Notes

Like other subject heading lists, Sears requires the cataloguer to arrive at a clear concept of the “aboutness” of a work as the first step toward selecting suitable headings. Sears subscribes to the principle of specific and direct entry so that headings that are co-extensive and specific to the main subject content of the work are assigned.

Direct entries ensure that the headings are entered on its own rather than subsumed under a subdivision. Although technically, as many headings as needed could be assigned in case of multiple-subject and complex works, Sears sees the practice of giving excess entries as a disservice to the user, and abides by the Rule of Three, stated as follows: “As many as three specific subject headings in a given area may be assigned to a work, but if the work treats of more than three subjects, then a broader heading is used instead and the specific headings are omitted.”

## Types of Headings

In favour of direct entry and natural language used by the user, Sears has converted all of its former inverted headings into direct entries. There are four types of headings:

**Topical headings:** words or phrases for common things or concepts; American usage and spellings are used, e.g., Elevators.

**Form headings:** reveals the form of the work, which is defined as referring to the intellectual form of materials rather than the physical format of the item, e.g., Encyclopedias and dictionaries; Fiction; Children’s plays.

**Geographic headings:** names of geographic areas, countries, cities, etc.

**Proper names:** personal, corporate names and uniform titles, e.g., Shakespeare, William, 1564-1616.

When more specific and narrow headings are needed, the scope of Sears can be expanded through the use of subdivisions. These include: topical, geographic, chronological and form subdivisions. As in the practice of using LCSH, Sears follows the LC recommendation for the standard order of [Topical] — [Geographic] — [Chronological] — [Form] in applying subdivisions to main headings.

## Application

The most specific heading authorized to represent each identified subject should be applied.

Use the most specific heading directly, not indirectly as a subdivision of a broader heading.

Generally, treat items first by topic, then by geographical focus or form, although some types of subject are exceptions.

Use subject headings that represent major literary forms or genres such as POETRY and FICTION for collections by several authors; do not use these headings for individual examples of the genre or for collections by a single author.

Treatment of literary works:

Works about any of the literary forms consist of their names: POETRY, DRAMA, FICTION, ESSAY.



*Example:* Newfoundland—Fiction.

Whenever the work is historical or critical, or both, the subdivision —HISTORY AND CRITICISM is added to the main heading.



*Example:* Science fiction—History and criticism.



## Notes

Works by one author, especially individual examples of an author's works are not usually assigned any headings at all.

Subject headings representing examples of literary forms or genres usually are reserved for collections of many authors.

Treatment of biographies:

Biographies about one, two, or three individuals (called the "biographees") are considered "individual" biographies. For such works, each biographee is assigned a subject heading consisting of that biographee's name in authorized AACR2R form. A second subject heading can be assigned if the item contains a great deal of information about the person's work or field of interest.



*Example:* Kennedy, John F. (John Fitzgerald), 1917-1963.

When a work is about four or more persons, it is deemed a "collective" biography. The heading reflects the common focus of the biographees.



*Example:*

- Hispanic Americans—Biography.
- Computer industry—Biography.

Treatment of works with geographical focus:

Many of the subject headings for topics in the fields of history, geography, and politics put the name of the location first and topical subdivisions following it.

Treatment of materials other than books:

Use the same principles to assign subject headings to videos, electronic resources, and other material formats.

"Videorecording" is not an authorized subdivision and do not use form subdivisions to describe physical format.

## Sears List

Minnie Earl Sears prepared the first edition of this work in response to demands for a list of subject headings.

### The Scope of the Sears List

New topics appear every day, and books on those topics require new subject headings. Headings for new topics can be developed from the Sears List in two ways, by establishing new terms as needed and by subdividing the headings already in the List.

### Form of Headings

The Sears List still reflects the usage of the Library of Congress unless there is some compelling reason to vary, but those instances of variation have become numerous over the years.

### Scope Notes

All the new and revised headings in this edition have been provided with scope notes where such notes are required. Scope notes are intended to clarify the specialized use of a term or to distinguish between terms that might be confused. If there is any question of what a term means, the cataloger should simply consult a dictionary.

## Classification

## Notes

The classification numbers in this edition of Sears are taken from the Abridged WebDewey, the continuously updated online version of the Abridged Dewey Decimal Classification. The numbers are intended only to direct the cataloger to a place in the DDC schedules where material on that subject is often found.

### Style, Filing, etc.

For spelling and definitions the editor has relied upon Webster's Third New International Dictionary of the English Language, Unabridged (1961) and the Random House Webster's Unabridged Dictionary, 2nd ed., revised and updated (1997). Capitalization and the forms of corporate and geographic names used as examples are based on the Anglo-American Cataloguing Rules, 2nd ed., 2002 revision. The filing of entries follows the ALA Filing Rules (1980).

Every term in the List that may be used as a subject heading is printed in boldface type whether it is a main term; a term in a USE reference; a broader, narrower, or related term; or an example in a scope note or general reference. If a term is not printed in boldface type, it is not used as a heading.

## What's New in the 19th Edition of Sears

### Over 400 new subject headings

#### New coverage of topics in the headlines

The growing interest in Islam among the general public and in school curriculum is reflected in the new materials published, for which the Sears List now provides heading such as Islam and politics, Islamic music, Muslim women, Shiites, Sunnis, and Dervishes.

#### Coverage of new forms of literature

The extraordinary growth in the publication of graphic novels prompts the timely addition of more than thirty new headings, among them: Adventure graphic novels, Romance graphic novels, Superhero graphic novels, Manga, Komodo, and Mecha. These new headings were suggested to us by Katherine L. Kan, a noted expert in the field. These headings are all genre headings and follow the patterns set by other literary form and genre headings already in the List.

New subject headings in a variety of other areas as well represent a major enhancement to the List in this edition. New headings have been added in the fields of science and technology, such as Computer animation, Open access publishing, and Stem cell research; in lifestyle and entertainment, such as Neopaganism, Reality television programs, and Body piercing; in politics and world affairs, such as War reparations, Suicide bombers, and Border patrols; and in literature and the arts, such as Urban fiction and Art pottery.

## Self Assessment

Multiple Choice Questions:

- Ms. Sears was died in ..... at age of 60.
 

|          |          |
|----------|----------|
| (a) 1913 | (b) 1923 |
| (c) 1933 | (d) 1943 |
- ..... provides suggestions for working with its headings in relation to these material forms and concepts.
 

|                     |                 |
|---------------------|-----------------|
| (a) Joseph miller   | (b) Ms. Sears   |
| (c) Barbara Bristow | (d) H.W. Wilson |

Notes

3. The ..... edition of the Sears list of subject headings melds the traditional with the new.
- |          |          |
|----------|----------|
| (a) 14th | (b) 16th |
| (c) 18th | (d) 19th |

### 13.2 Library of Congress Subject Headings

The Library of Congress Subject Headings (LCSH) comprises a thesaurus (in the information technology sense) of subject headings, maintained by the United States Library of Congress, for use in bibliographic records. LC Subject Headings are an integral part of bibliographic control, which is the function by which libraries collect, organize and disseminate documents. LCSHs are applied to every item within a library's collection, and facilitate a user's access to items in the catalogue that pertain to similar subject matter. If users could only locate items by 'title' or other descriptive fields, such as 'author' or 'publisher', they would have to expend an enormous amount of time searching for items of related subject matter, and undoubtedly miss locating many items because of the ineffective and inefficient search capability.

The Library of Congress subject headings system was originally designed as a controlled vocabulary for representing the subject and form of the books and serials in the Library of Congress collection, with the purpose of providing subject access points to the bibliographic records contained in the Library of Congress catalogues.

As an increasing number of other libraries have adopted the Library of Congress subject headings system, it has become a tool for subject indexing of library catalogs in general. In recent years, it has also been used as a tool in a number of online bibliographic databases outside of the Library of Congress.

A subject heading may consist of one or more words. A one-word heading represents a single concept, whereas a multiple-word heading may represent a single concept or multiple concepts.

A subject heading representing a single concept may appear as a single word or a multiple-word phrase, usually an adjectival phrase but occasionally a prepositional phrase. Each such heading represents a single object or idea (Examples include: Automobiles, Botany, Budget deficits, Electric interference, Boards of trade, Clerks of court).



*Task* How do you consider Library of Congress Subject Headings as an Art?

#### **An Art and a Science**

Subject heading classification is a human and intellectual endeavor, where trained professionals apply topic descriptions to items in their collections. Naturally, every library may choose to categorize the subject matter of their items differently, without a uniform consentaneous standard. The widespread use and acceptance of the Library of Congress Subject Headings facilitates the uniform access and retrieval of items in any library in the world using the same search strategy and LCSH thesaurus, if the correct headings have been applied to the item by the library.



*Did u know?* LCSH decisions involve a great amount of debate and even controversy in the library community.

Despite LCSH's wide-ranging and comprehensive scope, there are libraries where the use of LCSH is not ideal or effective. To deal with these types of collections and user communities, other subject headings may be required. The United States National Library of Medicine developed Medical Subject Headings (MeSH) to use on its many health science databases and collection. Many university

libraries may not apply both LCSH and MeSH headings to items. In Canada, the National Library of Canada worked with LCSH representatives to create a complementary set of Canadian Subject Headings (CSH) to access and express the topic content of documents on Canada and Canadian topics.

### 13.3 LCSH Policy Issues

Historically, issues have revolved around the terms employed to describe racial or ethnic groups. Notable has been the terms used to describe African-Americans. Until the 1990s, the LCSH administrators had a strict policy of not changing terms for a subject category. This was enforced to tighten and eliminate the duplication or confusion that might arise if subject headings were changed. Therefore, one term to describe African-American topics in LCSH was 'Afro-American' long after that term lost currency and acceptance in the population. LCSH decided to allow some alteration of terms in 1996 to better reflect the needs and access of library users. Nevertheless, many common terms, or 'natural language' terms are not used in LCSH, and may in effect limit the ability for users to locate items. There is a growing tradition of research in Library and Information Science faculties about the cultural and gender biases that affect the terms used in LCSH, which in turn may limit or deprive library users access to information stored and disseminated in collections. A notable American Library Science scholar on this subject is Sanford Berman.

#### Data Access

The Subject Headings are published in large red volumes (currently five), which are typically displayed in the reference sections of research libraries. They may also be searched online in the Library of Congress Classification Web, a subscription service, or free of charge (as individual records) at Library of Congress Authorities. The Library of Congress issues weekly updates. The data is published for a fee by the Cataloguing Distribution Service.



*Notes* A web service, [lcsch.info](http://lcsch.info), was set up by Ed Summers, a Library of Congress employee, circa April 2008,[1] using SKOS to allow for simple browsing of the subject headings. [lcsch.info](http://lcsch.info) was shut down by the Library of Congress's order on December 18, 2008.

This announcement was met with great dismay from the library science and semantic web communities, *e.g.*, Tim Berners-Lee and Tim Spalding of LibraryThing. After some delay, the Library did set up its own web service for LCSH browsing at [id.loc.gov](http://id.loc.gov) in April 2009.

#### Using LCSH

Once a library user has found the right subject heading(s), they are an excellent resource for finding relevant material in your library catalogue. Increasingly the use of hyperlinked, web-based Online Public Access Catalogues, or OPACs, allow users to hyperlink to a list of similar items displayed by LCSH once one item of interest is located. However, because LCSH are not necessarily expressed in natural language, many users may choose to search OPACs by keywords. Moreover, users unfamiliar with OPAC searching and LCSH, may incorrectly assume their library has no items on their desired topic, if they chose to search by 'subject' field, and the terms they entered do not strictly conform to a LCSH. For example 'body temperature regulation' is used in place of 'thermoregulation'. Thus the easiest way to find and use LCSH is to start with a 'keyword' search and then look at the Subject Headings of a relevant item to locate other related material.

The Library of Congress Subject Headings (LCSH) provides an alphabetical listing of authorized or preferred terms established by the Library of Congress. These "official" terms make searching for

**Notes**

music in the Friedheim Library more productive and efficient. LCSH is published annually in large volumes with red covers. In the Friedheim Library there is a copy on the index table in the Reference Area. Most names of places and people (i.e., proper nouns) are not listed in LCSH; however, they may be used as subject headings.

**USE References**

USE references are made FROM an unauthorized or non-preferred term TO an authorized or preferred term. They are made for synonyms and for older and variant forms of headings. For example,

- Spinnet: USE Harpsichord
- National anthems: USE National songs

**Components of LCSH Entries**

The following kinds of information may be found in subject headings:

- The authorized Subject heading is in boldface.
- The CODE (May Subd Geog) or (Not Subd Geog) (in italics) indicates whether or not the heading can be subdivided geographically.
- LC class numbers are often given when there is a close correspondence between the subject heading and the LC classification. Use this number to search the alphabetical Library of Congress call number index in the online catalog or browse the shelves for materials on your subject.
- Scope notes give guidance in the meaning or application of the heading.
- References express the relationship between terms:

|           |                |  |
|-----------|----------------|--|
| <b>UF</b> | Use For        | (Equivalency)  |
| <b>BT</b> | Broader Terms  | (Hierarchical)   |
| <b>NT</b> | Narrower Terms | (Hierarchical)   |
| <b>RT</b> | Related Terms  | (Associative)  |
| <b>SA</b> | See Also       | (A general reference to an entire group of headings or subdivisions rather than to individual headings or subdivisions.) |

**Example LCSH entry**

|   |                 |
|---|-----------------|
| Jazz                                    | Subject heading |
| (May Subd Geog)                         | Code            |
| [M1366 (Music)]                         | LC Class        |
| [ML3505.8-ML3509 History and criticism] | LC CLass        |

Here are entered jazz instrumental works for two or more performers. Songs performed in jazz style by a vocalist or vocal, with or without accompaniment, are entered under Jazz vocals.

|           |                                  |                |
|-----------|----------------------------------|----------------|
|           | Jazz — United States             |                |
| <b>UF</b> | Jive (Music)                     | <b>Use For</b> |
|           | Savophone and piano music (Jazz) |                |
|           | <i>[Former heading]</i>          |                |

|           |   | Broader Terms         | Notes |
|-----------|---|-----------------------|-------|
| <b>BT</b> | African Americans – Music   |                       |       |
| <b>RT</b> | Washboard band music  | <b>Related Term</b>   |       |
|           | <i>Headings for solo instrumental music followed by the parenthetical qualifier “(Jazz),” e.g. Piano music (Jazz); also headings that include the term “jazz ensemble” as a medium for performance,</i> |                       |       |
| <b>SA</b> | <i>e.g. Concertos (Flute with jazz ensemble); and headings for musical instruments with the subdivisions Methods (Jazz) or Studies and exercises (Jazz)</i>   |                       |       |
|           | Big band music  |                       |       |
|           | Boogie Woogie   |                       |       |
|           | Bop (Music)   |                       |       |
|           | Dixieland music   |                       |       |
| <b>NT</b> | Instrumentation and orchestraion (Dance Orchestra)  | <b>Narrower Terms</b> |       |
|           | Jazz vocals   |                       |       |
|           | Latin jazz  |                       |       |
|           | Swing (Music)   |                       |       |
|           | Western swing (Music)   |                       |       |
|           | Jazz — To 1921  |                       |       |
|           | • 1921–1930   |                       |       |
|           | • Interpretation (Phrasing, dynamics, etc.)   | <b>Subdivisions</b>   |       |
|           | • Lead sheets   |                       |       |
|           | • Religious aspects   |                       |       |

### Subdivisions

Subdivisions combine a number of different concepts into a single subject heading. Only a fraction of all possible heading and subdivision combinations are listed in LCSH.

There are four types of subdivisions:

1. Topical: *e.g.*, Musical theater—Production and direction
2. Form: *e.g.*, Music Bibliography ; Sonatas (Violin and piano) Scores and parts
3. Chronological: Jazz—1921-1930
4. Geographic: Folk music—Thailand

For details on what subdivisions are used for composers, go to the alphabetical list of subjects and look up the pattern heading “Wagner, Richard” for some of the possibilities; for musical instruments, look up “Piano”; for music compositions, look up “Operas”.

### Choice of Terms in LCSH

Frequently, LC subject headings are not the terms most commonly used. Examples of how different the terms follow:

| Common terms            | LCSH                    |
|-------------------------|-------------------------|
| Piano duets             | Piano music (4 hands)   |
| Flute and bassoon music | Bassoon and flute music |
| Film music              | Motion picture music    |

**Notes**

When searching for chamber music for specific combinations of instruments, it often helps to search the print volumes of LCSH to find the proper order in which to enter the instruments.

**Singular and Plural**

Often LCSH uses the singular of words to denote the treatment of the word as a subject and the plural of the same word to denote individual examples of that subject. So searching under the term Opera you will find books and other materials dealing with the subject of opera. Under the term Operas, you will find individual operas (full scores, vocal scores, recordings videos).

**Searching in the Online Catalogue**

When using the Johns Hopkins online catalogue, you may do an alphabetical search for the LC Subject Heading. When you find a useful item, follow the links for narrower or related subjects that seem helpful. If you are not sure of the official subject heading, check the volumes of LCSH located on the Index Table in the Reference area, or you may perform a Subject Keyword Search in the online catalogue to find valid headings. Again, when you find a useful item, follow the links for narrower or related subjects that seem helpful.

**13.4 Medical Subject Headings (MeSH)**

Medical Subject Headings (MeSH) is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences; it can also serve as a thesaurus that facilitates searching. Created and updated by the United States National Library of Medicine (NLM), it is used by the MEDLINE/PubMed article database and by NLM's catalogue of book holdings.



*Notes* MeSH can be browsed and downloaded free of charge on the Internet through PubMed. The yearly printed version was discontinued in 2007 and MeSH is now available online only.

Originally in English, MeSH has been translated into numerous other languages and allows retrieval of documents from different languages.

From 6-15 subject headings are assigned for each article, with up to 3 assigned for major emphasis of the article. Articles are indexed to the most specific term available to allow for very precise subject searching. Subheadings, terms which cover general, frequently discussed aspects of a subject such as adverse effects or therapy, are combined with MeSH terms to indicate the specific focus.

A particularly powerful feature designed into Medline allows users to “explode” a category of terms in a hierarchy from general to specific to retrieve all of the articles on the general term and all of the specific terms listed underneath. “Explode” is distinct from the concept of truncation in that the terms do not have to begin with the same string of characters to be retrieved. “Exploding” a term allows the information requestor to search a term and all levels of its narrower terms.

**Structure of MeSH**

The 2009 version of MeSH contains a total of 25,186 subject headings, also known as descriptors. Most of these are accompanied by a short description or definition, links to related descriptors, and a list of synonyms or very similar terms (known as entry terms). Because of these synonym lists, MeSH can also be viewed as a thesaurus.

## Descriptor Hierarchy

## Notes

The descriptors or subject headings are arranged in a hierarchy. A given descriptor may appear at several places in the hierarchical tree. The tree locations carry systematic labels known as tree numbers, and consequently one descriptor can carry several tree numbers. For example, the descriptor “Digestive System Neoplasms” has the tree numbers C06.301 and C04.588.274; C stands for Diseases, C06 for Digestive System Diseases and C06.301 for Digestive System Neoplasms; C04 for Neoplasms, C04.588 for Neoplasms By Site, and C04.588.274 also for Digestive System Neoplasms. The tree numbers of a given descriptor are subject to change as MeSH is updated. Every descriptor also carries a unique alphanumerical ID that will not change.

## Descriptions

Most subject headings come with a short description or definition. The explanatory text is written by the MeSH team based on their standard sources if not otherwise stated. References are mostly encyclopaedias and standard textbooks of the subject areas. References for specific statements in the descriptions are not given, instead readers are referred to the bibliography.

## Qualifiers

In addition to the descriptor hierarchy, MeSH contains a small number of standard qualifiers (also known as subheadings), which can be added to descriptors to narrow down the topic. For example, “Measles” is a descriptor and “epidemiology” is a qualifier; “Measles/epidemiology” describes the subheading of epidemiological articles about Measles. The “epidemiology” qualifier can be added to all other disease descriptors. Not all descriptor/qualifier combinations are allowed since some of them may be meaningless. In all there are 83 different qualifiers.

## Supplements

In addition to the descriptors, MeSH also contains some 139,000 Supplementary Concept Records. These do not belong to the controlled vocabulary as such and are not used for indexing MEDLINE articles; instead they enlarge the thesaurus and contain links to the closest fitting descriptor to be used in a MEDLINE search. Many of these records describe chemical substances.

## Use in Medline/PubMed

In MEDLINE/PubMed, every journal article is indexed with some 10-15 headings and subheadings, with some of them designated as major and marked with an asterisk. When performing a MEDLINE search via PubMed, entry terms are automatically translated into (= ‘mapped to’) the corresponding descriptors with a good degree of reliability; it is recommended to check the Details tab in PubMed to see how a search formulation was ‘translated’. By default a search will include all the descriptors that are located below the given one in the hierarchy.

The Medical Subject Headings (or MeSH for short) is designed to help you focus your search, and to avoid ambiguous terms or synonyms, *i.e.* where one word can mean many different things, or where different words are used for the same topic.

## Medical Subject Headings (MeSH) Indexing Tips

Medical Subject Headings (MeSH) is a list (thesaurus) of keywords or descriptors that describe articles in Index Medicus and MEDLINE. Indexers scan an entire article and assign up to twenty MeSH terms to each article. Terms are chosen to cover both the central aspects of an article (major headings) and other significant information discussed (minor headings).

By using terms from the MeSH thesaurus, all articles on a given topic can be found regardless of the terminology used by the authors.



|              |                     |  |
|--------------|---------------------|--|
| <b>Notes</b> | Specificity         | — Each article is indexed to the most specific MeSH terms available, <i>e.g.</i> an article on acne is indexed under acne, but not under skin diseases.  |
|              | Near Match          | — Articles with no exact match are indexed to the closest related MeSH term, <i>e.g.</i> seminal vesiculitis to seminal vesicles, pseudoappendicitis to appendicitis, nursing caps to clothing.  |
|              | Two Terms           | — The most precise way to cover a topic may be two MeSH terms in combination, <i>e.g.</i> jejunitis to jejunal diseases and enteritis.   |
|              | Textwords           | — It is assumed you will use textwords in some cases to define a subject, <i>e.g.</i> tobacco smoke pollution (MeSH term) and passive (textword) to retrieve passive smoking.  |
|              | Check Tags          | — Large-volume concepts are routinely “checked” for in each article by indexers. Check tags pinpoint specific age groups, males or females, humans or animals, publication types, etc.   |
|              | Drugs               | — Drugs are indexed under the generic name, <i>e.g.</i> valium is indexed to diazepam.   |
|              | Medical Specialty   | — There are separate terms for the medical specialty and the disease or organ, <i>e.g.</i> endocrinology is the specialty versus endocrine diseases or endocrine glands.   |
|              | Neoplasms           | — Neoplasms are indexed to site and histologic type, <i>e.g.</i> adenocarcinoma of the colon is indexed to both colonic neoplasms and adenocarcinoma.  |
|              | Relational Concepts | — Some relational concepts cannot be indexed precisely, <i>e.g.</i> , degrees of quality or quantity, specific time relationships, primary versus secondary except for neoplasms, general body positions. Try or experiment with textwords for these concepts. Even then you may not retrieve the relationship you wish. |

“Medical Subject Headings (MeSH), the hierarchical classification scheme of some 19,000 main headings and codes used for indexing databases produced by the National Library of Medicine, must be cited when looking for “best practices” in indexing. The Medline database is a premier biomedical database and is the electronic counterpart to Index Medicus, International Nursing Index, International Dental Literature. MeSH indexing available within Medline is a key feature of the database.

The Medical Subject Headings are continually revised and updated by subject specialists responsible for areas of the health sciences in which they have knowledge and expertise. The staff collects new terms as they appear in the scientific literature or in emerging areas of research; define these terms within the context of existing vocabulary; and recommend their addition to MeSH. They also receive suggestions from indexers and other professionals. This indexing structure has stood the test of time and is widely acclaimed for the accuracy and precision in retrieval that it allows.

MeSH should be considered the gold standard and a benchmark for evaluating indexing structures in other disciplines” (Sykes, 2001, 5-6).

Jenuwine & Floyd (2004) investigated the performance of two search strategies in the retrieval of primary research papers containing descriptive information on the sleep of healthy people from MEDLINE. Two search strategies—one based on the use of only Medical Subject Headings (MeSH), the second based on text-word searching—were evaluated as to their specificity and sensitivity in retrieving a set of relevant research papers published in the journal *Sleep* from 1996 to 2001 that were preselected by a hand search.

The subject search provided higher specificity than the text-word search (66% and 47%, respectively) but lower sensitivity (78% for the subject search versus 88% for the text-word search). Each search strategy gave some unique relevant hits. The paper concludes that the two search strategies complemented each other and should be used together for maximal retrieval. No combination of MeSH terms could provide comprehensive yet reasonably precise retrieval of relevant articles. The text-word searching had sensitivity and specificity comparable to the subject search. In addition, use of text words “normal.... healthy,” and “control” in the title or abstract fields to limit the final sets provided an efficient way to increase the specificity of both search strategies.



*Task* Explain the term Medical Subject Headings (MeSH).

### The MeSH Tree Structure

The MeSH vocabulary is organized by 16 main branches, including those listed below:

Anatomy

Organisms

Diseases

Chemical and Drugs

Analytical, Diagnostic and Therapeutic Techniques and Equipment

Psychiatry and Psychology

Biological Sciences

Natural Sciences

### MeSH facts at a glance

- 1874 - John Shaw Billings produces the Index Catalogue of the Library of the Surgeon-General.
- 1879 - Index Medicus is published as a monthly comprehensive index of medical journal articles.
- 1927 - Index Medicus is merged with the American Medical Association’s competing bibliography and renamed Quarterly Cumulative Index Medicus.
- 1951 - Colonel Frank Rogers produces Current List of Medical Literature, a standardized list of subject headings.
- 1960 - Medical Literature Analysis and Retrieval System (MEDLARS) is developed, as part of this effort a MeSH database is also developed with a new and thoroughly revised list of subject headings.
- 1963 - MeSH database is updated with “Tree Structure”.

### MeSH - Maintenance

- Medical Subject Headings staff continually revise and update the MeSH vocabulary.
- Staff collects new terms as they appear in the scientific literature or in emerging areas of research
- Staff defines terms within the context of existing vocabulary; and recommend their addition to MeSH.
- Consultants also advise and offer recommendations.

Notes

**MeSH- application - I**

- The 17,000 word MeSH vocabulary is divided into complex hierarchical tree structure.
- There are 16 main top level categories in the hierarchical structure which further give rise to branch like tree structure sub categories in order of their increasing specificity.
- The MeSH tree structure allows to find relevant MeSH terms and corresponding articles in PubMed even when only general concept areas are known.

**MeSH- application- II**

- The MeSH hierarchical tree structure allows to have a control on both precision and recall.
- Recall- The MeSH allows to broaden the domain and increase the recall by fetching the related articles as well.
- Precision- The MeSH tree structure leads to specificity and hence increases precision.
- For a proficient MEDLINE searcher - Right balance needed between precision and recall.

**MeSH- application- III**

- **Descriptors**- Also known as Main Headings, Descriptors are used to index citations in NLM's MEDLINE database, for cataloging of publications, and other databases.
- **Publication Characteristics**- Although MeSH Descriptors, these records are unlike other MeSH Descriptors in that they indicate what the indexed item is, i.e., its type, rather than what it is about, for example, *Historical Article, Editorial*
- **Geographics**-Descriptors which include continents, regions, countries, states, and other geographic subdivisions.
- **Qualifiers**-There are 83 topical Qualifiers (also known as Subheadings) used for indexing and cataloging in conjunction with Descriptors. Qualifiers afford a convenient means of grouping together those citations which are concerned with a particular aspect of a subject. For example- *Liver/Drug effects*.

**MeSH- relevance to Information Architecture - I**

- Mesh Tree structures successfully incorporate the following four major IA components:
  - Organization System
    - Whole tree organized into 16 main categories
    - Further organized into sub-categories
    - Each heading is labeled in a decreasing order of spectrum (range)
  - Labeling System
    - Labels contain terms which are directly/indirectly related
    - Poor labels can ruin a good organization or navigation system

**MeSH- relevance to Information Architecture - II**

- Navigation System
  - Global (what's important)
  - Local (what's nearby)
  - Contextual (related)

- Searching System  
Returns the corresponding MeSH heading  
Presents the exact tree structure of the subject

### Problem of Language

- It is the problem faced by inexperienced MEDLINE user while looking for appropriate MeSH terms in order to execute a query.
- Many biomedical terms point to the single canonical MeSH term.  
Example- Heart Failure, Congestive is the only MeSH term to represent all types of cardiac failures

#### Solutions-Problem of Language

- Use of entry terms- used to map non MeSH terms to MeSH terms.
- Use MeSH hierarchies to traverse to the articles.

### Research

#### NLM's Indexing Initiative (IND)

- Human indexing is an expensive and labor-intensive activity.
- IND project involves investigating automated methods to substitute for current indexing practices.
- More stress is being given to improve the current retrieval performance.
- New approaches like Machine Learning algorithms viz. Naïve. Bayes, Adaptive Boosting and Support Vector Machines are being tested.

### List of MeSH codes

The following is a list of the codes for MeSH. It is a product of the United States National Library of Medicine.

Source for content is 2006 MeSH Trees.

- A — Anatomy
- A01 —body regions (74 articles)
- A02 —musculoskeletal system (213 articles)
- A03 —digestive system (98 articles)
- A04 —respiratory system (46 articles)
- A05 —urogenital system (87 articles)
- A06 —endocrine system
- A07 —cardiovascular system
- A08 —nervous system
- A09 —sense organs
- A10 —tissues
- A11 —cells
- A12 —fluids and secretions

**Notes**

- A13 — animal structures
- A14 — stomatognathic system
- A15 — hemic and immune systems
- A16 — embryonic structures
- A17 — integumentary system
- B — Organisms
- B01 — animals
- B02 — algae
- B03 — bacteria
- B04 — viruses
- B05 — fungi
- B06 — plants
- B07 — archaea
- B08 — mesomycetozoa
- C — Diseases
- C01 — bacterial infections and mycoses
- C02 — virus diseases
- C03 — parasitic diseases
- C04 — neoplasms
- C05 — musculoskeletal diseases
- C06 — digestive system diseases
- C07 — stomatognathic diseases
- C08 — respiratory tract diseases
- C09 — otorhinolaryngologic diseases
- C10 — nervous system diseases
- C11 — eye diseases
- C12 — urologic and male genital diseases
- C13 — female genital diseases and pregnancy complications
- C14 — cardiovascular diseases
- C15 — hemic and lymphatic diseases
- C16 — congenital, hereditary, and neonatal diseases and abnormalities
- C17 — skin and connective tissue diseases
- C18 — nutritional and metabolic diseases
- C19 — endocrine system diseases
- C20 — immune system diseases
- C21 — disorders of environmental origin
- C22 — animal diseases
- C23 — pathological conditions, signs and symptoms
- D — Chemicals and Drugs

- D01 — inorganic chemicals
- D02 — organic chemicals
- D03 — heterocyclic compounds
- D04 — polycyclic compounds
- D05 — macromolecular substances
- D06 — hormones, hormone substitutes, and hormone antagonists
- D07 — none (enzymes and coenzymes)
- D08 — enzymes and coenzymes (carbohydrates)
- D09 — carbohydrates (lipids)
- D10 — lipids (amino acids, peptides, and proteins)
- D11 — none (nucleic acids, nucleotides, and nucleosides)
- D12/20 — amino acids, peptides, and proteins (complex mixtures)
- D13/23 — nucleic acids, nucleotides, and nucleosides (biological factors)
- D14/25 — biomedical and dental materials
- D15/26 — pharmaceutical preparations
- D16/27 — chemical actions and uses
- D20 — complex mixtures
- D23 — biological factors
- E — Analytical, Diagnostic and Therapeutic Techniques and Equipment
- E01 — diagnosis
- E02 — therapeutics
- E03 — anesthesia and analgesia
- E04 — surgical procedures, operative
- E05 — investigative techniques
- E06 — dentistry
- E07 — equipment and supplies
- F — Psychiatry and Psychology
- F01 — behavior and behavior mechanisms
- F02 — psychological phenomena and processes
- F03 — mental disorders
- F04 — behavioral disciplines and activities
- G — Biological Sciences
- G01 — biological sciences
- G02 — health occupations
- G03 — environment and public health
- G04 — biological phenomena, cell phenomena, and immunity

|              |     |   |
|--------------|-----|---|
| <b>Notes</b> | G05 | — genetic processes                                       |
|              | G06 | — biochemical phenomena, metabolism, and nutrition        |
|              | G07 | — physiological processes                                 |
|              | G08 | — reproductive and urinary physiology                     |
|              | G09 | — circulatory and respiratory physiology                  |
|              | G10 | — digestive, oral, and skin physiology                    |
|              | G11 | — musculoskeletal, neural, and ocular physiology          |
|              | G12 | — chemical and pharmacologic phenomena                    |
|              | G13 | — genetic phenomena                                       |
|              | G14 | — genetic structures                                      |
|              | H   | — Physical Sciences                                       |
|              | H01 | — natural sciences  |
|              | I   | — Anthropology, Education, Sociology and Social Phenomena |
|              | I01 | — social sciences   |
|              | I02 | — education   |
|              | I03 | — human activities  |
|              | J   | — Technology and Food and Beverages                       |
|              | J01 | — technology, industry, and agriculture                   |
|              | J02 | — food and beverages                                      |
|              | K   | — Humanities  |
|              | K01 | — humanities  |
|              | L   | — Information Science                                     |
|              | L01 | — information science                                     |
|              | M   | — Persons   |
|              | M01 | — persons   |
|              | N   | — Health Care   |
|              | N01 | — population characteristics                              |
|              | N02 | — health care facilities, manpower, and services          |
|              | N03 | — health care economics and organizations                 |
|              | N04 | — health services administration                          |
|              | N05 | — health care quality, access, and evaluation             |
|              | V   | — Publication Characteristics                             |
|              | V01 | — publication components (publication type)               |
|              | V02 | — publication formats (publication type)                  |
|              | V03 | — study characteristics (publication type)                |
|              | V04 | — support of research                                     |
|              | Z   | — Geographic Locations                                    |
|              | Z01 | — geographic locations                                    |

**Medical Subject Headings**

**Notes**

Alphabetic List

Abelmoschus

Abies

Abrus

Acanthaceae

Access to Information

Achyranthes

Acoraceae

Acorus

Actinidia

Actinidiaceae

Activin Receptors

Activin Receptors, Type I

Activin Receptors, Type II

Activins

Acupuncture

Adenomatous Polyposis Coli Protein

Adenoviruses, Bovine

Adenoviruses, Porcine

Adhatoda

Advance Directive Adherence

Advisory Committees

AGAMOUS Protein, Arabidopsis

Agavaceae

Agave

Alnus

Alpha Karyopherins

Alpha-Tocopherol

Alzheimer Vaccines

Amacrine Cells

Amaranthaceae

Amaranthus

Amino Acid Transport System A

Amino Acid Transport System ASC

Amino Acid Transport System L

Amino Acid Transport System X-AG

Amino Acid Transport System y+

Amino Acid Transport System y+L



Notes

Amino Acid Transport Systems  
Amino Acid Transport Systems, Acidic  
Amino Acid Transport Systems, Basic  
Amino Acid Transport Systems, Neutral  
Amino Acids, Acidic  
Amino Acids, Aromatic  
Amino Acids, Basic  
Amphibian Proteins  
Amsinckia  
Amsonia  
Amyloid Neuropathies, Familial  
Amyloidosis, Familial  
Anacardiaceae  
Anemia, Diamond-Blackfan  
Anemia, Hypoplastic, Congenital  
Anethum graveolens  
Aneugens  
Angelica  
Angelica archangelica  
Angelica sinensis  
Animal Migration  
Animals, Genetically Modified  
Anion Transport Proteins  
Anise  
Annonaceae  
Antigens, CD98  
Antigens, CD98 Heavy Chain  
Antigens, CD98 Light Chains  
Antineoplastic Protocols  
Apium graveolens  
Apocynaceae  
Apocynum  
Aquifoliaceae  
Arabidopsis Proteins  
Araceae  
Aralia  
Araliaceae  
Archaeal Viruses  
Arctostaphylos

Notes

Arecaceae  
Argas  
Argasidae  
Aristolochia  
Aristolochiaceae  
Arteriviridae  
Asarum  
Asclepiadaceae  
Ascoviridae  
Asfarviridae  
Asparagus  
Aspartate Aminotransferase, Cytoplasmic  
Aspartate Aminotransferase, Mitochondrial  
Aspidosperma  
Astragalus gummifer  
Astragalus membranaceus  
Astragalus Plant  
AT-Hook Motifs  
ATP Synthetase Complexes  
Atrial Myosins  
Atriplex  
Avian Proteins  
Bacterial Proton-Translocating ATPases  
Bacteriochlorophyll A  
Bacteriophage HK022  
Bacteriophage IKE  
Bacteriophage N4  
Bacteriophage Pf1  
Bacteriophage PRD1  
Balsaminaceae  
Bassia scoparia  
Batrachoidiformes  
Beloniformes  
Beneficence  
Berberidaceae  
Berberis  
Beta Karyopherins  
Beta vulgaris  
Beta-Tocopherol

**Notes**

- Betaretrovirus
- Betula
- Betulaceae
- Bignoniaceae
- Bioethical Issues
- Biomedical Enhancement
- Biotinidase Deficiency
- Black Pepper
- Blood Coagulation Disorders, Inherited
- Blotting, Far-Western
- Blotting, Southwestern
- Blueberry Plant
- Boraginaceae
- Borago
- Bornaviridae
- Borrelia burgdorferi
- Boswellia
- Brassica napus
- Brassica rapa
- BRCA2 Protein
- Bronchitis, Chronic
- Bryonia
- Bupleurum
- Burseraceae
- Bystander Effect
- Cactaceae
- Caenorhabditis elegans Proteins
- Caesalpinia
- Calluna
- Calophyllum
- Camellia
- Camellia sinensis
- Campanulaceae
- Camptotheca
- Canarypox virus
- Cannabaceae
- Capitalism
- Caprifoliaceae
- Capsella

Notes

Carbonic Anhydrase I  
Carbonic Anhydrase II  
Carbonic Anhydrase III  
Carbonic Anhydrase IV  
Carbonic Anhydrase V  
Carboxypeptidase U  
Cardiac Myosins  
Cardiomyopathy, Hypertrophic, Familial  
Carica  
Carum  
Caryophyllaceae  
Catha  
Catharanthus  
Cation Transport Proteins  
Cationic Amino Acid Transporter 1  
Cationic Amino Acid Transporter 2  
Caulophyllum  
Cedrus  
Celastraceae  
Cellular Apoptosis Susceptibility Protein  
Cementogenesis  
Centaurium  
Centella  
Cephaelis  
Cerclage, Cervical  
Cerebral Amyloid Angiopathy, Familial  
Chamaecyparis  
Chelidonium  
Chenopodium  
Chenopodium album  
Chenopodium ambrosioides  
Chenopodium quinoa  
Chive  
Chloride-Bicarbonate Antiporters  
Chloroplast Proton-Translocating ATPases  
Chromatography, Supercritical Fluid  
Chromosome Disorders  
Chromosome Pairing  
Chromosomes, Artificial, P1 Bacteriophage

**Notes**

Cicer  
Cichlids  
Cicuta  
Cinnamomum  
Cinnamomum camphora  
Cistaceae  
Cistus  
Citrullus  
Clinical Trials Data Monitoring Committees  
Closteroviridae  
Clove  
Clusiaceae  
Codonopsis  
Cola  
Collagen Type I  
Collagen Type II  
Collagen Type III  
Collagen Type IV  
Collagen Type IX  
Collagen Type V  
Collagen Type VI  
Collagen Type VII  
Collagen Type VIII  
Collagen Type X  
Collagen Type XI  
Collagen Type XII  
Collagen Type XIII  
Colonography, Computed Tomographic  
Combretaceae  
Combretum  
Commiphora  
Complicity  
Conium  
Convolvulaceae  
Coriandrum  
Cornaceae  
Cornus  
Coronary Restenosis  
Coronary Stenosis

Notes

Coronavirus 229E, Human  
Coronavirus OC43, Human  
Costus  
Crassulaceae  
Crataegus  
Crepis  
Crinivirus  
Crocus  
Crotalaria  
Croton  
Cucumis  
Cucumis melo  
Cucurbita  
Cuminum  
Cupressaceae  
Cupressus  
Curcuma  
Cycas  
Cyperaceae  
Cyperus  
Cytisus  
Cytochrome-c Oxidase Deficiency  
D-Aspartic Acid  
Databases, Genetic  
Databases, Nucleic Acid  
Databases, Protein  
DEFICIENS Protein  
Denys-Drash Syndrome  
Derris  
Dianthus  
Diarrhea Virus 1, Bovine Viral  
Diarrhea Virus 2, Bovine Viral  
Diazepam Binding Inhibitor  
Dicarboxylic Acid Transporters  
Dihydroergocornine  
Dihydroergocristine  
Dihydroergocryptine  
Dinosaurs  
Dioscorea

Notes

- Dioscoreaceae
- Diospyros
- Directly Observed Therapy
- Disclosure
- DNA Gyrase
- DNA Topoisomerase IV
- DNA Topoisomerases
- DNA Topoisomerases, Type I, Archaeal
- DNA Topoisomerases, Type I, Bacterial
- DNA Topoisomerases, Type I, Eukaryotic
- DNA Topoisomerases, Type II, Archaeal
- DNA Topoisomerases, Type II, Bacterial
- DNA Topoisomerases, Type II, Eukaryotic
- Dominance, Ocular
- Drosera
- Drosophila Proteins
- Drug Resistance, Bacterial
- Drug Resistance, Fungal
- Drug Resistance, Multiple, Bacterial
- Drug Resistance, Multiple, Fungal
- Drug Resistance, Multiple, Viral
- Drug Resistance, Viral
- E-Box Elements
- Ebola-like Viruses
- Ecdysteroids
- Echocardiography, Stress
- Eclecticism, Historical
- Egg Hypersensitivity
- Electric Capacitance
- Electrophoretic Mobility Shift Assay
- Elettaria
- Eleutherococcus
- Enterovirus A, Human
- Enterovirus B, Human
- Enterovirus C, Human
- Enterovirus D, Human
- Enterovirus, Bovine
- Ephedra
- Ephedra sinica

Epididymal Secretory Proteins

Epsilonretrovirus

Ericaceae

Erythroxylaceae

Escherichia coli Proteins

Esociformes

Estrous Cycle

Ethical Analysis

Ethical Relativism

Ethical Review

Ethical Theory

Ethicists

Ethics Committees, Clinical

Ethics Committees, Research

Ethics, Clinical

Euonymus

Euphorbia

Euthanasia, Active

Excitatory Amino Acid Transporter 1

Excitatory Amino Acid Transporter 2

Exercise Movement Techniques

F2-Isoprostanes

Factor XIIIa

Fagaceae

Fagus

Faith Healing

Feedback, Biochemical

Feedback, Psychological

Feline panleukopenia virus

Ferns

Fibril-Associated Collagens

Fibrillar Collagens

Ficus

Fish Proteins

3' Flanking Region

5' Flanking Region

Fluoroquinolones

Foeniculum

Food, Genetically Modified



Notes

- Foot-and-Mouth Disease Virus
- Fowl adenovirus A
- Fundulidae
- Fungal Components
- Galanthus
- Galega
- gamma-Tocopherol
- Garcinia
- Garcinia cambogia
- Garcinia kola
- Garcinia mangostana
- Gastroepiploic Artery
- GB virus A
- GB virus B
- GB virus C
- Gene, sry
- Genes, BRCA2
- Genetic Diseases, Inborn
- Genetic Enhancement
- Genetic Privacy
- Gentian
- Gentianaceae
- Glyceraldehyde 3 Phosphate-Dehydrogenase (NADP+)(Phosphorylating)
- Glyceraldehyde 3-Phosphate Dehydrogenase (NADP+)
- Glyceraldehyde-3-Phosphate-Dehydrogenase (Phosphorylating)
- Glycogen Phosphorylase
- Glycogen Phosphorylase, Brain Form
- Glycogen Phosphorylase, Liver Form
- Glycogen Phosphorylase, Muscle Form
- Glycyrrhiza uralensis
- Gonadal Dysgenesis, 46,XX
- Granulovirus
- Guttaviridae
- Gymnotiformes
- Gynostemma
- Halorhodopsins
- Hamamelidaceae
- Hamamelis
- Heliotropium

Hemorrhagic Syndrome, Bovine  
Hepatitis A virus  
Heracleum  
Herpesvirus 1, Meleagrid  
Herpesvirus 3, Gallid  
Herpesvirus 4, Bovine  
Herpesvirus 5, Bovine  
Hevea  
Heymann Nephritis Antigenic Complex  
Hip Injuries  
Hirudin Therapy  
HIV Fusion Inhibitors  
HMG-Box Domains  
HMGA Proteins  
HMGA1a Protein  
HMGA1b Protein  
HMGA1c Protein  
HMGA2 Protein  
HMGB Proteins  
HMGB1 Protein  
HMGB2 Protein  
HMGB3 Protein  
HMGN Proteins  
HMGN1 Protein  
HMGN2 Protein  
Holocarboxylase Synthetase Deficiency  
Huckleberry Plant  
Humulus  
Hydrangeaceae  
Hydrophobicity  
Hydrophyllaceae  
Hydroxymethylglutaryl-CoA Reductases, NAD-Dependent  
Hydroxymethylglutaryl-CoA-Reductases, NADP-dependent  
Hyperotreti  
Hyphae  
Hypoxis  
Ilex  
Ilex guayusa  
Ilex paraguariensis

Notes

Ilex vomitoria  
Illicium  
Impatiens  
Indigofera  
Indole Alkaloids  
Infectious hematopoietic necrosis virus  
Infectious Laryngotracheitis-like Viruses  
Influenzavirus A  
Influenzavirus B  
Inhibin-beta Subunits  
Ipomoea  
Ipomoea batatas  
Iridaceae  
Isoaspartic Acid  
Isoprostanes  
Ixodidae  
Jervell-Lange Nielsen Syndrome  
Juglandaceae  
Karyopherins  
Lagovirus  
Lanthanoid Series Elements  
Large Neutral Amino Acid-Transporter 1  
Larix  
Laser Therapy, Low-Level  
Lathyrus  
Laughter Therapy  
Laurus  
Lavandula  
LDL-Receptor Related Protein 1  
LDL-Receptor Related Protein 2  
LDL-Receptor Related Protein-Associated Protein  
LDL-Receptor Related Proteins  
Ledum  
Leeching  
Lentils  
Lepidium  
Lespedeza  
Levisticum  
Libocedrus

Notes

Ligusticum  
Lilium  
Limb Salvage  
Lindera  
Lipothrixviridae  
Lithospermum  
Lobelia  
Loganiaceae  
Loranthaceae  
Lotus  
Luffa  
Lycopodiaceae  
Lythraceae  
Maackia  
MADS Domain Proteins  
Magnoliaceae  
Mahonia  
Maize streak virus  
Malpighiaceae  
Malus  
Malva  
Mandragora  
Manipulation, Chiropractic  
Manipulation, Osteopathic  
Marburg-like Viruses  
Marek's Disease-like Viruses  
Marrubium  
Maytenus  
MCM1 Protein  
Medicago  
Medicine, Herbal  
Meliaceae  
Melilotus  
Membrane Transport Proteins  
Menispermaceae  
Mentally Ill Persons  
Mentha  
Metabolic Syndrome X  
Metapneumovirus

**Notes**

Mikamycin  
Mind-Body and Relaxation Techniques  
Mitochondrial Diseases  
Mitochondrial Proteins  
Mitochondrial Proton-Translocating ATPases  
Molecular Diagnostic Techniques  
Momordica  
Momordica charantia  
Monimiaceae  
Monocarboxylic Acid Transporters  
Moraceae  
Moral Obligations  
Moringa  
Moringa oleifera  
Multidrug Resistance-Associated Proteins  
Murine pneumonia virus  
Musa  
Musaceae  
Muscle Development  
Musculoskeletal Manipulations  
Mushroom Bodies  
Mycelium  
Myosin Type I  
Myosin Type II  
Myosin Type III  
Myosin Type IV  
Myosin Type V  
Myricaceae  
Myristica fragrans  
Myristicaceae  
Myroxylon  
Myrsinaceae  
Myrtaceae  
Nairobi sheep disease virus  
Nanotechnology  
Nanovirus  
National Socialism  
Nepeta

Nerium  
Neurofibromin 1  
Neurofibromin 2  
Nidovirales  
Nidovirales Infections  
Nitregic Neurons  
Nodaviridae  
Non-Fibrillar Collagens  
Nonmuscle Myosin Type IIA  
Nonmuscle Myosin Type IIB  
Norwalk-like Viruses  
Novirhabdovirus  
Nuclear Pore Complex Proteins  
Nucleocytoplasmic Transport Proteins  
Nurse's Role  
Nut Hypersensitivity  
Nymphaea  
Nymphaeaceae  
Nyssaceae  
Ochrosia  
Ocotea  
Oleaceae  
Oleavirus  
Oligosaccharides, Branched-Chain  
Onagraceae  
Oncogene Proteins v-raf  
Open Bite  
Oplopanax  
Optic Atrophy, Autosomal Dominant  
Optic Atrophy, Hereditary, Leber  
Opuntia  
Orchidaceae  
Organic Anion Transport Polypeptide C  
Organic Anion Transport Protein 1  
Organic Anion Transporters  
Organic Anion Transporters, ATP-Dependent  
Organic Anion Transporters, Sodium-Dependent  
Organic Anion Transporters, Sodium-Independent

**Notes**

- Organic Cation Transport Proteins
- Organic Cation Transporter 1
- Organisms, Genetically Modified
- Ornithodoros
- Orthoreovirus, Avian
- Orthoreovirus, Mammalian
- Oxyphil Cells
- p14ARF Protein
- Pachyrhizus
- Paeonia
- Palyam Virus
- Papaveraceae
- Papillomaviridae
- Papillomavirus Infections
- Parainfluenza Virus 3, Bovine
- Parainfluenza Virus 4, Human
- Parechovirus
- Parvovirus, Porcine
- Passiflora
- Pastinaca
- Paternalism
- Patient Education Handout [Publication Type]
- Patient Rights
- Peanut Hypersensitivity
- Pedaliaceae
- Perineuronal Satellite Cells
- Permethrin
- Peroxynitrous Acid
- Persea
- Personal Autonomy
- Personhood
- Peste-des-Petits-Ruminants
- Petroselinum
- Phaseolus
- Phoradendron
- Phosphate Transport Proteins
- Phosphate-Binding Proteins
- Phosphofructokinase-1, Liver Type

Notes

Phosphofructokinase-1, Muscle Type  
Phosphofructokinase-1, Type C  
Phosphofructokinase-2  
Phosphofructokinases  
Photography, Dental  
Phyllanthus  
Phyllanthus emblica  
Physical Therapy (Specialty)  
Physical Therapy Techniques  
Physostigma  
Phytolacca  
Phytolacca americana  
Phytolacca dodecandra  
Phytolaccaceae  
Picea  
Pimenta  
Pinaceae  
Pinellia  
Pinus  
Piper nigrum  
Piperaceae  
Pistacia  
Plant Bark  
Plant Preparations  
Plants, Genetically Modified  
Plumbaginaceae  
Podophyllum peltatum  
Polyadenylation  
Polygalaceae  
Polygonum  
Polyomaviridae  
Polyomavirus Infections  
Polypodiaceae  
Polyubiquitin  
Portulacaceae  
Potassium Channel Blockers  
Potassium Channels, Calcium-Activated  
Potassium Channels, Inwardly Rectifying



**Notes**

- Potassium Channels, Tandem Pore Domain
- Potassium Channels, Voltage-Gated
- Potassium-Hydrogen Antiporters
- Primate T-lymphotropic virus 1
- Primate T-lymphotropic virus 2
- Primate T-lymphotropic virus 3
- Primulaceae
- Principal Component Analysis
- Principle-Based Ethics
- Pristinamycin
- Professional Misconduct
- Professional Role
- Prosopis
- Prostatic Secretory Proteins
- Protein D-Aspartate-L-Isoaspartate Methyltransferase
- Protein Interaction Mapping
- Proton-Phosphate Symporters
- Prunus
- Pseudotsuga
- Psoralea
- Pteridaceae
- Pueraria
- Pulmonaria
- Pulmonary Disease, Chronic Obstructive
- Puumala Virus
- Pyruvate Dehydrogenase (Lipoamide)
- Quercus
- Ranunculaceae
- Reactive Nitrogen Species
- Reproductive Techniques, Assisted
- Reptilian Proteins
- Rhamnaceae
- Rhinomanometry
- Rhinometry, Acoustic
- Rhizome
- Rhodiola
- Rhododendron
- Rhodopsins, Microbial

Rhus  
RNA 3' End Processing  
Romano-Ward Syndrome  
Rosa  
Rosaceae  
Rosmarinus  
Rudiviridae  
Rutaceae  
S-Nitroso-N-Acetylpenicillamine  
S-Nitrosoglutathione  
S-Nitrosothiols  
Saccharomyces cerevisiae Proteins  
Salicaceae  
Salvia  
Salvia miltiorrhiza  
Salvia officinalis  
Sambucus  
Sambucus nigra  
Sandfly fever Naples virus  
Sanicula  
Santalaceae  
Sapindaceae  
Saponaria  
Sapotaceae  
Sapporo-like Viruses  
Sarcoma, Granulocytic  
Sassafras  
Saussurea  
Schisandraceae  
Schizosaccharomyces pombe Proteins  
Scrophulariaceae  
Second-Look Surgery  
Seminal Plasma Proteins  
Seminal Vesicle Secretory Proteins  
Sendai virus  
Senna  
Sensory Art Therapies  
Sensory Rhodopsins

**Notes**

Seoul virus  
Septo-Optic Dysplasia  
Serenoa  
Serum Response Element  
Serum Response Factor  
Sex Chromosome Disorders  
Shallots  
Sialic Acid Storage Disease  
Silene  
Simaroubaceae  
Simian T-lymphotropic virus 2  
Sin Nombre virus  
Singlet Oxygen  
Skeletal Muscle Myosins  
Small Ubiquitin-Related Modifier Proteins  
Smegmamorpha  
Smilacaceae  
Smooth Muscle Myosins  
Sodium Channel Blockers  
Sodium-Bicarbonate Symporters  
Sodium-Potassium-Chloride Symporters  
Somatic Hypermutation, Immunoglobulin  
Sophora  
Soybean Proteins  
Spiritual Therapies  
Spirituality  
Spondylarthritis  
Spondylarthropathies  
Starch Phosphorylase  
Steam Bath  
Stellaria  
Sterculiaceae  
Streptogramin A  
Streptogramin B  
Streptogramin Group A  
Streptogramin Group B  
Streptogramins  
Strophanthus

Notes

Styracaceae  
SUMO-1 Protein  
Suregada  
Surgery, Computer-Assisted  
Swertia  
Symporters  
Synteny  
Tabebuia  
Tabernaemontana  
Tai Ji  
Takifugu  
Tamarindus  
Tamus  
Tauopathies  
Tenuivirus  
Terminalia  
Tetraodontiformes  
Theaceae  
Theology  
Thuja  
Thyme  
Thymelaeaceae  
Tiliaceae  
Tissue Engineering  
Tobacco necrosis satellite virus  
Tocopherols  
Tocotrienols  
Transcription Initiation Site  
Transition Elements  
Trichosanthes  
Trifolium  
Trigonella  
Tripterygium  
Tsuga  
Tumor Suppressor Proteins  
Turnera  
Ubiquitin  
Ubiquitin C

Notes

Umbellularia  
Uniparental Disomy  
Urtica dioica  
Urticaceae  
Vaccinium  
Vaccinium macrocarpon  
Vaccinium myrtillus  
Vaccinium vitis-idaea  
Vacuolar Proton-Translocating ATPases  
Valerianaceae  
Value of Life  
Ventricular Myosins  
Verbenaceae  
Vernamycin B  
Vesivirus  
Viburnum  
Violaceae  
Virtues  
Viscaceae  
Viscum  
Viscum album  
Vitaceae  
Vitamin B 6  
Vitamin B 6 Deficiency  
Vitamin D Response Element  
Vitamin K 2  
Vitamin K 3  
Vitis  
Wheat Hypersensitivity  
Withholding Treatment  
WT1 Proteins  
Xanthophylls  
Xenopus Proteins  
Yarrowia  
Yucca  
Zebrafish Proteins  
zeta Carotene  
Zingiberaceae  
Zygophyllaceae

**Self Assessment****Notes**

Fill in the blanks:

4. .... a term allows the information requestor to search a term and all levels of its narrower terms.
5. The descriptors or subject headings are arranged in a ..... .
6. Most subject headings come with a short definition or ..... .
7. In addition to the descriptors, MeSH also contains some ..... supplementary concept records.
8. Medical subject Headings is a list of keywords or descriptors that describe articles in index medicus and ..... .

**13.5 Summary**

- Medical Subject Headings (MeSH) is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences; it can also serve as a thesaurus that facilitates searching.
- The Library of Congress Subject Headings (LCSH) comprises a thesaurus (in the information technology sense) of subject headings, maintained by the United States Library of Congress, for use in bibliographic records.
- Online Public Access Catalogues, or OPACs, allow users to hyperlink to a list of similar items displayed by LCSH once one item of interest is located.
- Joseph Miller and Barbara Bristow maintain the primary mission of Sears by using simplified vocabulary specifically geared for the school and small public library setting.
- The 19th edition of the Sears List of Subject Headings melds the traditional with the new.
- Several notable 19th edition features include over 440 completely new subject headings and two totally new categories — “Islam” and “Graphic Novels.” Expanded coverage of headings in the science/technology, lifestyle/entertainment, politics/world affairs, and literature/arts categories are also noteworthy additions.

**13.6 Keywords**

*MeSH* : Medical Subject Headings also serves as a thesaurus that facilitates searching.

*OPAC* : Online Public Access Catalogue, allows users to hyperlink to a list of similar items displayed by LCSH once one time of interest is located.

**13.7 Review Questions**

1. Write the principles of the sears list.
2. What comprises Library of Congress Subject Headings (LCSH)?
3. Mention the structure of MeSH.

**Answers: Self Assessment**

1. (c)
2. (b)

Notes

3. (d)
4. Exploding
5. Hierarchical
6. Description
7. 139,000
8. MEDLINE

### 13.8 Further Readings



Books

Deegan, M. and Simon Tanner. *Digital futures*. London. LA, 2002.

Maltby, A., ed. *Sayer's manual of classification for libraries*. 5th. Ed. London: Andre Deutsch, 1975.

Oddy, P. *Future libraries, future catalogs*. London: LA, 1996.



Online links

<http://home.olemiss.edu/~tharry/SH/lcshguide.pdf>

<http://blis10kristel.blogspot.com/2010/11/>

[http://hlwiki.slais.ubc.ca/index.php/Medical\\_Subject\\_Headings\\_%28MeSH%29](http://hlwiki.slais.ubc.ca/index.php/Medical_Subject_Headings_%28MeSH%29)

## Unit 14: ERIC and Thesaurofacet

### CONTENTS

Objectives

Introduction

14.1 ERIC (Educational Resources Information Centre) Thesaurus

14.2 Thesaurofacet

14.3 Summary

14.4 Keywords

14.5 Review Questions

14.6 Further Readings

### Objectives

After studying this unit, you will be able to:

- Define keyword vs. description searching
- Describe UF and RT
- Explain thesaurofacet.

### Introduction

The Thesaurus of ERIC Descriptors (Thesaurus) is a controlled vocabulary - a carefully selected list of education-related words and phrases assigned to ERIC records to organize them by subject and make them easier to retrieve through a search. Searching by Descriptors involves selecting relevant terms from this controlled vocabulary to locate information on your topic.

The inadequacy of conventional information systems to meet effectively the challenge posed by the accelerated rate of growth of information in science and technology has given rise to modern high speed retrieval systems. Proper functioning of these systems depends upon the existence of certain basic tools. One such tool is the thesaurus.

### 14.1 ERIC (Educational Resources Information Center) Thesaurus

The thesaurus terms are used by indexers to describe the contents of publications in a consistent, comprehensive and concise manner. These terms are listed in the Descriptors field (DE=) of each record added to the database.

### Keyword vs. Descriptor Searching

While you can also search ERIC using keywords of your choosing, you will get more precise search results if you use Thesaurus terms. That's because searching by keywords requires matching the



**Notes**

exact words found in a record, while searching by Descriptors allows you to locate records indexed by subject, regardless of the terminology the author may have used. The ERIC Thesaurus will allow you to conduct systemic searches and save time by reducing guesswork and trial-and-error methods.

**Getting Started - Search ERIC Using ERIC Descriptors**

To plan your search using ERIC Descriptors, we recommend that you follow these steps:

Describe the topic in your own words.

Divide the topic into major concepts.

Use the Thesaurus to locate the appropriate Descriptors for each concept of the topic.

Add the Descriptor(s) to your search.

If you'd like to take a shortcut, you might try simply doing a keyword search, retrieving a record that looks relevant to your topic, and then examining the Descriptors. You can even click on one of the Descriptors to start a new search.



*Did u know?* The Thesaurus of ERIC Descriptors, 13th Edition, contains an alphabetical listing of terms used for indexing and searching in the ERIC database.

This word-by-word alphabetical display is probably the most familiar since it provides a variety of information (a “display”) for each Descriptor. This includes a Scope Note, Use For (UF) and Use (USE) references, Narrower Terms (NT), Broader Terms (BT), and Related Terms (RT). Each of these segments of the Thesaurus display is explained in detail below.

**Scope Note**

A Scope Note is a brief statement of the intended usage of a Descriptor. It may be used to clarify an ambiguous term or to restrict the usage of a term. Special indexing notes are often included.

**TESTS**

Devices, procedures, or sets of items that are used to measure ability, skill, understanding, knowledge, or achievement.



*Notes* Use a more specific term if possible—this broad term corresponds to pubtype code 160 and should not be used except as the subject of a document.

**Oral Interpretation**

The oral interpretation and presentation of a work of literature to an audience Alerts indexers and searchers to an earlier Thesaurus instruction.



*Notes* Prior to mar80, the instruction “oral interpretation, use interpretive reading” was carried in the thesaurus.

**Non-formal Education**

Organized education without formal schooling or institutionalization in which knowledge, skills, and values are taught by relatives, peers, or other community members.

## Notes



*Notes* Do not confuse with “nonschool educational programs” or the identifier “informal education”.

**UF (Use For)**

The “UF” reference is employed generally to solve problems of synonymy occurring in natural language. Terms following the UF notation are not used in indexing. They most often represent either synonymous or variant forms of the main term, or specific terms that, for purposes of storage and retrieval, are indexed under a more general term. The examples below illustrate both types of UFs:

**MAINSTREAMING**

Use For

Desegregation (Disabled Students)

Integration (Disabled Students)

Least Restrictive Environment (Disabled)

Regular Class Placement (1968–1978)

**LIFELONG LEARNING**

Use For

Continuous Learning (1967–1980)

Education Permanente

Lifelong Education

Life Span Education

Permanent Education

Recurrent Education

**LABOR FORCE DEVELOPMENT**

Use For

Human Resources Development (Labor)

Manpower Development(1966–1980)

**PHYSICAL DISABILITIES**

Use For

Crippled Children (1968–1980)

Orthopedically Handicapped (1968–1980)

Physical Handicaps (1966–1980)

A former Descriptor that has been downgraded to the status of a UF term is accompanied by a “life span” notation in parentheses: *e.g.*, (1966 1980). This indicates the time period during which the term was used in indexing. It provides useful information for searching older printed indexes, or computer files that have not been updated.



*Task* Compare the traditional subject headings with the latest ones.

**Notes**

**Use**

The USE reference, the mandatory reciprocal of the UF, refers an indexer or searcher from a nonusable (nonindexable) term to the preferred indexable term or terms.

In the examples below, there is only one USE term for each entry. This means that there is a direct, one-to-one correlation in the ERIC system from the UF to the USE term.

REGULAR CLASS PLACEMENT (1968–1978)

USE MAINSTREAMING

CONTINUOUS LEARNING (1967–1980)

USE LIFELONG LEARNING

ORTHOPEDICALLY HANDICAPPED (1968–1980)

USE PHYSICAL DISABILITIES

MANPOWER

USE LABOR FORCE

A coordinate or multiple USE reference looks a little different. The following example illustrates the use of two main terms together to represent a single concept, both for indexing and searching:

FOLK DRAMA (1969–1980)

USE DRAMA AND FOLK CULTURE

BT (Broader Term) and NT (Narrower Term)

These indicate the existence of a hierarchical relationship between a class and its subclasses. Narrower terms are included in the broader class represented by the main entry. The [+] symbol beside a term indicates that there are further narrower terms.

**LIBRARIES**

Narrower Terms

Academic Libraries [+]

Branch Libraries

Childrens Libraries

Depository Libraries

Public Libraries [+]

Research Libraries

School Libraries

Special Libraries [+]

**MODELS**

Narrower Terms

Causal Models [+]

Mathematical Models

Role Models [+]

Student Writing Models

Teaching Models

The Broader Term (BT) is the mandatory reciprocal of the NT. Broader Terms include as a subclass the concept represented by the main (narrower) term.

Notes

## SCHOOL LIBRARIES

Broader Terms

Libraries [+]

## MATHEMATICAL MODELS

Broader Terms

Models [+]

Sometimes a term may have more than one Broader Term:

## REMEDIAL READING

Broader Terms

Reading [+]

Reading Instruction [+]

Remedial Instruction [+]



*Notes* In ERIC, computer searching of a Broad Term will not automatically retrieve documents representing the concepts of its Narrower Terms, unless those NTs have also been assigned to the documents in indexing (*e.g.*, searching LIBRARIES will not automatically also retrieve literature on SCHOOL LIBRARIES or any of the other NTs to LIBRARIES). To search automatically the broad term and all its narrower terms, use the Explode function.

## RT (Related Term)

Related terms have a close conceptual relationship to the main term, but not the direct class/subclass relationship described by BTs/NTs. Part-whole relationships, near-synonyms, and other conceptually related terms, which might be helpful to the user, appear as RTs.

## HIGH SCHOOL SENIORS

Related Terms

College Bound Students

Grade 12

High School Freshmen

High School Graduates

Noncollege Bound Students

## MINIMUM COMPETENCY TESTING

Related Terms

Academic Achievement [+]

**Notes**

Academic Standards [+]  
Basic Skills [+]  
Competence [+]  
Competency Based Education [+]  
Mastery Tests  
Minimum Competencies  
National Competency Tests  
Student Certification  
Test Score Decline  
Parenthetical Qualifiers

A Parenthetical Qualifier is used to identify a particular indexable meaning of a homograph. In other words, it discriminates between terms (either Descriptors or USE references) that might otherwise be confused with each other. Examples include LETTERS (ALPHABET) and LETTERS (CORRESPONDENCE); SELF EVALUATION (INDIVIDUALS) and SELF EVALUATION (GROUPS).



*Notes* The Qualifier is considered an integral part of the Descriptor and must be used with the Descriptor in indexing and searching.

## **Self Assessment**

Fill in the blanks:

1. A ..... is a brief statement of the intended usage of a descriptor.
2. ....., procedures, or sets of items that are used to measure ability, skill, understanding, knowledge, or achievement.
3. While you can also search ERIC using keywords of your choosing, you will get more precise search results if you use ..... terms.

Multiple Choice Questions:

4. A former descriptor that has been downgraded to the status of a UF term is accompanied by a ..... notation in parentheses.  
(a) Lifelong (b) Life span  
(c) Recurrent (d) Permanent
5. .... have a close conceptual relationship to the main term, but not the direct class/subclass relationship described by BTs/NTs.  
(a) Narrow Terms (b) Broader Terms  
(c) Related Terms (d) None of These

## **14.2 Thesaurofacet**

The term "thesaurofacet" was coined by Aitchison *et al.* (1969) as the combination of a faceted classification and a thesaurus.

Aitchison (1970) describes a faceted classification and thesaurus covering engineering and related scientific, technical, and management subjects. A novel feature of the system is the integration of

the classification schedules and thesaurus. Each term appears both in the thesaurus and in the schedules.

In the schedules the term is displayed in the most appropriate facet and hierarchy: the thesaurus supplements this information by indicating alternative hierarchies and other relationships which cut across the classified arrangement. The thesaurus also controls word forms and synonyms and acts as the alphabetical index to the class numbers.

The thesaurus is one of the most commonly used controlled vocabulary indexing tools - the aim of the FACET project is to investigate the retrieval potential of thesauri. FACET investigates the closer integration of the thesaurus into the interface and search techniques that do not require the user to exactly match how an item has been indexed.

FACET collaborates with the J. Paul Getty Trust in exploring the retrieval potential of its vocabularies, in particular the Art and Architecture Thesaurus (AAT), and with the National Museum of Science and Industry (NMSI) in its attempts to promote wider access to its collections database. The aim is to complement NMSI's development of major areas of 'rich content'. Railway/Locomotive History has been selected as one area particularly appropriate for the project due to its AAT coverage and synergy with ongoing work at the National Railway Museum on extending the AAT with railway terms. The mda and CHIN act as advisors to the project.

### **What is a thesaurus and why FACET?**

Thesauri and classifications are types of controlled indexing vocabulary, in which index terms are restricted to a controlled set of terms. A large number of systems exist, covering a variety of subject domains, for example the Medical Subject Headings, the Art and Architecture Thesaurus and the Dewey Decimal Classification. These controlled vocabularies have long been part of standard cataloguing practice in libraries and museums and are now being applied to digital hypertexts via thematic keywords in metadata resource descriptors. Metadata sets for the WWW, such as Dublin Core and the Resource Description Framework (RDF) typically include the more complex notion of the Subject of a resource in addition to elements for Title, Creator, Date, etc.

It is recommended that, where possible, the Subject element be taken from a relevant controlled vocabulary. This semantic index approach offers the potential for searcher and indexer to speak the same language, and for a user to be guided to fruitful terms when searching a particular collection for a particular purpose. Links between concepts in the subject domain can be expressed by the semantic relationships in a thesaurus (or classification). The three main thesaurus relationships are Equivalence (equivalent terms), Hierarchical (broader/narrower terms), and Associative (more loosely Related Terms). Specialisations of the three main relationships offer possibilities for semantic web applications.

Facet analysis is a key technique in thesaurus construction; concepts are decomposed into elemental classes, or facets, which form homogenous mutually exclusive groups. The faceted approach to subject analysis began in 1933 with Ranganathan's Colon Classification (Personality, Matter, Energy, Space and Time) and was subsequently elaborated by the British Classification Research Group. Faceted thesauri or classification systems include MESH, BLISS, PRECIS and the main thesaurus used in the project, the Art and Architecture Thesaurus (AAT).

The AAT is a large, evolving thesaurus (nearly 120,000 terms), organised into 7 facets (and 33 hierarchies as subdivisions) according to semantic role: Associated concepts, Physical attributes, Styles and periods, Agents, Activities, Materials, Objects and optional facets for time and place.

Faceted thesauri are similar in structure to faceted classifications but explicitly represent equivalence, hierarchical and associative links between concepts. A thesaurus can be used as a search thesaurus

## Notes

for refining or expanding a free text query (either interactively or automatically). Alternatively a thesaurus can be used both in searching and indexing with controlled vocabulary indexed datasets and this latter use is the immediate application of our current work (although we also see the techniques as useful with free text searching).

In retrieval, thesaurus relationships are conventionally used to expand synonyms and sometimes narrower query terms but the FACET system also performs more general semantic term expansion (to broader and to related concepts). Reasoning over the semantic relationships in the thesaurus permits imprecise matching between query and index terms. This allows the ranking of matching items in a result list or a 'More like this' option for similar but not necessarily identically indexed items.

Faceted systems are based on a primary division of terminology into fundamental, high-level categories, or facets. A knowledge system can be considered as enumerative, when all possible simple and compound terms are explicitly listed in their hierarchical position, or as synthetic. Faceted systems are normally synthetic; they do not attempt to include the vast number of possible multi-concept headings or descriptors in a domain, but combine terms from a limited number of fundamental facets, as needed when indexing or querying. This flexibility allows highly specific, nuanced metadata descriptions (or annotations). Matching such compound descriptors poses significant challenges when searching and the full potential for retrieval has remained untapped.

## Objectives

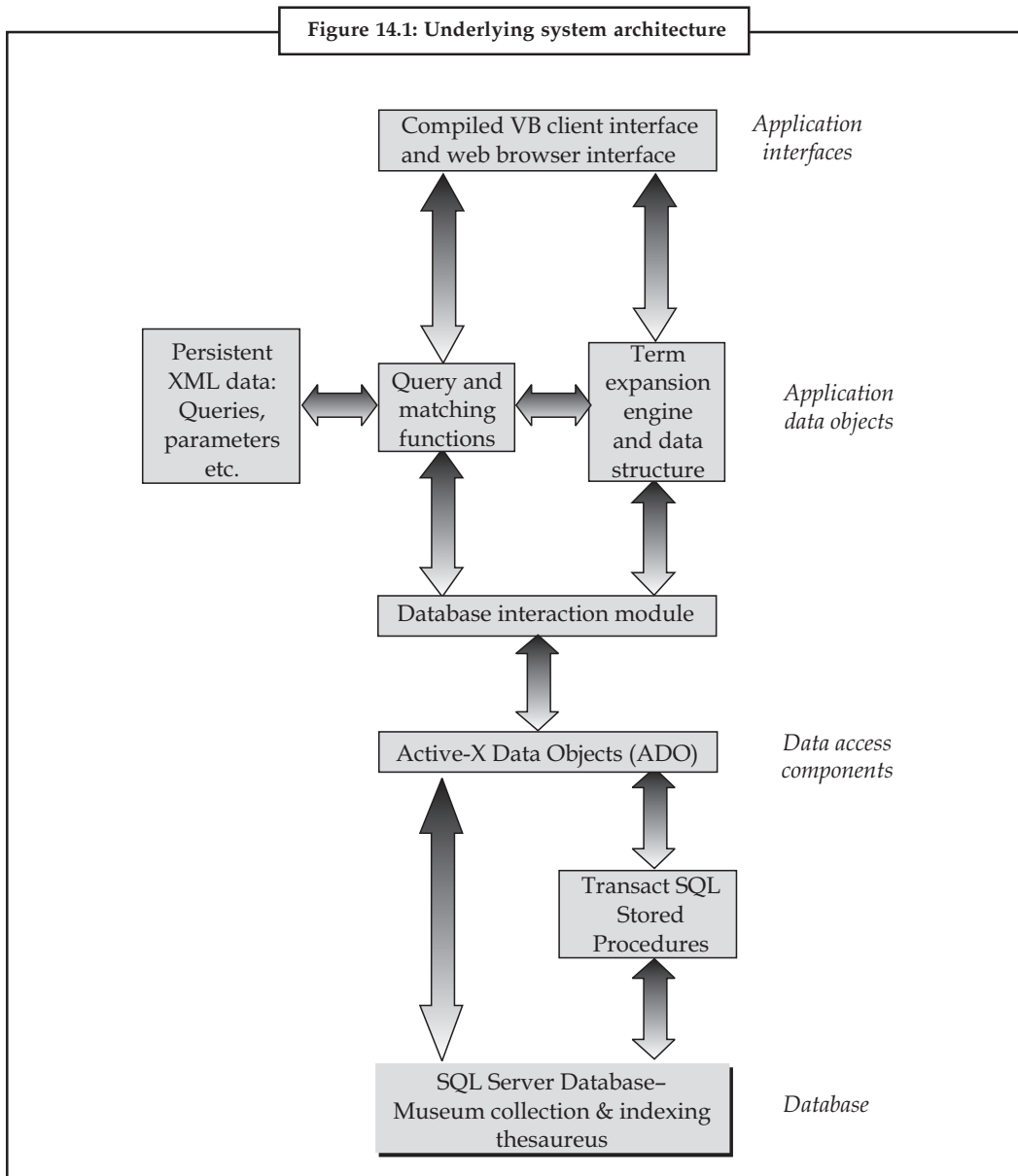
- The overarching objective of the research was to.
- Develop and evaluate retrieval tools based on a matching function incorporating thesaurus semantic closeness measures.
- Derive heuristics to guide automatic and interactive expansion/refinement of strings of thesaurus terms, taking advantage of the context provided by facets.
- Experiment with techniques for creating complex queries using a query editor with knowledge of the semantic roles of thesaurus facets. This will draw on previous work in the cultural heritage domain.
- Design and implement semantic closeness measures based on thesaurus relationships.

## Beneficiaries of the research

The research is directly relevant to cultural heritage organisations and the users of their digital collections, also to collection management vendors and commercial image providers. Thesauri are one of the most common Knowledge Organisation Systems and frequently underpin higher level schemas and ontologies. Initiatives to update international thesaurus standards are currently underway and various groups are working on XML/RDF representations for thesauri. Thesauri and faceted approaches have been applied to website architecture and hierarchical browsing interfaces to web databases.

## FACET Architecture and Interfaces

The final FACET system comprises a tiered component-based architecture (Fig. 14.1), accessing a SQL Server relational database. Queries with associated results are stored persistently using XML format data.



This architecture has enabled the reuse of key underlying components in the development of two main client interfaces - the first a compiled standalone VB 'fat' client, the second a browser based ASP web application. Intrinsic to both systems is the C++ semantic expansion engine operating over the in-memory directed graph structure populated from the relational tables representing the thesaurus.

### Controlled Vocabularies: A Glosso-Thesaurus

"There is a singular lack of vocabulary control in the field of controlled vocabularies. To help you cut through the maze of verbiage often found in this field, we have created a glossary of terms.

The glossary reflects our usage of terms in the articles of this series. But this glossary is more than just a list of terms. We wanted it to serve as an illustration of what a controlled vocabulary looks like (we are fond of killing multiple birds with multiple stones).



**Notes**

Accordingly, the glossary is itself a controlled vocabulary, more specifically a thesaurus. So you will find all of the standard features of any thesaurus: broader, narrower, and variant term indicators, as well as scope notes. In this case, however, the scope notes provide the definition of the particular glossary term being presented.

Glosso-Thesaurus

The following standard abbreviations are used in the glosso-thesaurus.

**BT** = Broader Term

**NT** = Narrower Term

**RT** = Related Term ("See also")

**SN** = Scope Note

**UF** = Used For

**USE** = "See" (Refers reader from variant term to vocabulary term.)

**Alternate Term**

**USE Variant Term**

**Associative Relationship**

SN The connection between related vocabulary terms. That is, related terms are connected through an associative relationship.

BT Term Relationship

RT Equivalence Relationship

Hierarchical Relationship

Related Term

**Authority File**

SN A flat (non-hierarchical) list containing preferred terms. May include variant terms. Essentially, an authority file is a synonym ring with the preferred term identified for each concept.

BT Controlled Vocabulary

RT Synonym Equivalence List

**Broader Term**

SN The superordinate word in an inclusion or hierarchical relationship. A class or category term. Abbreviated in displays as "BT." The inversion of broader term is narrower term. For example, "shoe" is a broader term than "running shoe." Broader terms are sometimes referred to as "parent" terms.

**UF Parent Term**

RT Hierarchical Relationship

Narrower Term

Related Term

**Card Sorting****Notes**

SN An exercise that can be used to help create a controlled vocabulary. In a card sort, users are asked to group cards into like categories or to name categories of like items. Card sorting can be used to compile lists of variant terms or to verify the relationships in a hierarchy.

RT Free Listing

Hierarchy

**Child Term**

USE Narrower Term

**Controlled Vocabulary**

SN A subset of natural language that is used to tag documents and then to find content through navigation or search. Use of a controlled vocabulary increases consistency in tagging and can help match users' natural language with preferred terms. Abbreviated as "CV."

Controlled vocabularies exhibit the following relationships:

Synonym ring

+ Preferred terms =

Authority file

+ Broader and narrower terms =

Taxonomy

+ Related terms =

Thesaurus

NT Authority File

Faceted Classification

Synonym Equivalence List

Synonym Ring

Taxonomy

Thesaurus

RT Natural Language

**Entry Term**

**USE Variant Term**

**Equivalence Relationship**

SN The connection between terms in a synonym ring, or between preferred terms and variant terms. Terms that exhibit an equivalence relationship refer to the same concept. For example, "cat" and "feline" are often considered as being equivalent.

BT Term Relationship

RT Associative Relationship

Hierarchical Relationship

Variant Term

Notes

**Exhaustivity**

SN The range of concept coverage of vocabulary terms in a controlled vocabulary. If the vocabulary terms cover all of the concepts included in the content under consideration, then the controlled vocabulary is exhaustive.

**RT Specificity**

**Facet**

SN A fundamental category by which an object or concept may be described. For example, a child's ball may be described using the facets of size, weight, shape, color, texture, material and price.

RT Facet Analysis

Faceted Classification

**Facet Analysis**

SN The process of analyzing content to determine appropriate facets and vocabulary term relationships, using "one characteristic of division at a time, to produce homogeneous, mutually-exclusive groups."

RT Facet

Faceted Classification

**Faceted Classification**

SN A controlled vocabulary that divides vocabulary terms into facets.

BT Controlled Vocabulary

**Free Listing**

A method of vocabulary development in which users are asked to "name all the [x] you know." Free listing can identify core terms in a controlled vocabulary, as well as variant terms.

RT Card Sorting

User Warrant

**Granularity**

SN The level of specificity with which content is described. The more granular, the more specific.

RT Specificity

**Hierarchy**

SN A collection of vocabulary terms that show levels of superordination and subordination. Hierarchies comprise broader terms and narrower terms. Hierarchies may be tested using card sorting.

**RT Card Sorting**

Polyhierarchy

Taxonomy

**Hierarchical Relationship**

SN The connection between broader and narrower terms in a taxonomy or thesaurus.

BT Term Relationship

RT Associative Relationship

Broader Term

Equivalence Relationship

Narrower Term

**Literary Warrant****Notes**

SN The inclusion of a vocabulary term in a controlled vocabulary based on its appearance in one or more content items. For example, a medical text may use the term "oncology." Based on literary warrant, that term would be included in the controlled vocabulary even though the general public uses the term "cancer."

RT User Warrant

Vocabulary Term

**Narrower Term**

SN The subordinate word in an inclusion or hierarchical relationship. A member or part. Abbreviated in displays as "NT." For example, "running shoe" is a narrower term than "shoe." Narrower terms are sometimes referred to as "child" terms. The inversion of narrower term is broader term.

UF Child Term

RT Broader Term

Hierarchical Relationship

Related Term

**Natural Language**

SN Language as it is spoken; language in everyday use.

RT Controlled Vocabulary

User Warrant

**Non-preferred Term**

USE Variant Term

**Parent Term**

USE Broader Term

**Polyhierarchy**

SN A hierarchy in which some vocabulary terms have more than one broader term. For example, "Rome" might be a narrower term under both "European capitals" and "Italian cities" in a geographic vocabulary.

RT Hierarchy

Taxonomy

**Precision**

SN A ratio that measures the success of a search. Precision is defined mathematically as the number of relevant items returned by a search divided by the total number of items returned by the search. Thus, a search that returned only relevant items would have a precision of 1.0.

Precision usually has an inverse relationship to recall. That is, increasing the precision of a search usually decreases the recall. Precision can be increased by increasing the specificity of vocabulary terms. For more information, see:

IAWiki: "Recall vs. Precision"

**Notes**

Ongoing: "On Search: Precision and Recall"

RT Recall

Specificity

**Preferred Term**

SN The vocabulary term in a controlled vocabulary used to tag content.

RT Broader Term

Narrower Term

Variant Term

**Recall**

SN A ratio that measures the success of a search. Recall is defined mathematically as the number of relevant items returned by a search divided by the total number of relevant items in the collection. Thus, a search that returned all the relevant items in a collection would have a recall of 1.0.

Recall can be increased by the use of synonym rings and variant terms. Recall usually has an inverse relationship to precision. That is, increasing the recall of a search usually decreases the precision. For more information, see:

IAWiki: "Recall vs. Precision"

Ongoing: "On Search: Precision and Recall"

RT Precision

Variant Terms

**Related Term**

SN Vocabulary terms in a controlled vocabulary that are closely related. That is, they refer to closely related concepts. Abbreviated in displays as "RT." Related terms may, for example, exhibit the following relationships:

field of study/objects studied

operation/agent

action/product of action

concepts/properties

agent/counter-agent

concept/opposite

**UF "See Also" Term**

RT Associative Relationship

**Scope Note**

SN (1) A definition of a preferred term in a controlled vocabulary. (2) An indication of restrictions in meaning or other clarification needed for the proper use of the preferred term. Abbreviated in displays as "SN." Examples of scope notes are provided throughout this glossary.

RT Preferred Term

USE Related Term

**Specificity****Notes**

SN The exactness with which a vocabulary term covers a concept. Thus, in considering the concept "dog," the term "canine" is more specific than "animal." Increasing specificity of vocabulary terms increases precision and granularity, but may decrease recall.

RT Exhaustivity

Granularity

**Synonym Equivalence List**

SN A synonym ring or an authority file.

BT Controlled Vocabulary

RT Authority File

Synonym Ring

**Synonym Ring**

SN One of the simplest of controlled vocabularies. Includes only a list of equivalent terms. When one of the terms is searched, the synonym ring returns results as if the complete set of terms was searched.

BT Controlled Vocabulary

RT Equivalence Relationship

Synonym Equivalence List

**Taxonomy**

SN A controlled vocabulary, the preferred terms of which are all connected in a hierarchy or polyhierarchy. Terms in a taxonomy may exhibit equivalence or hierarchical relationships.

BT Controlled Vocabulary

RT Hierarchical Relationship

Hierarchy

Polyhierarchy

**Term**

USE Vocabulary Term

**Term Relationship**

SN The type of association between vocabulary terms. Terms may be broader, narrower, related or variant, exhibiting hierarchical, associative or equivalence relationships.

NT Associative Relationship

Equivalence Relationship

Hierarchical Relationship

RT Broader Term

Narrower Term

Notes

Related Term

Variant Term

**Thesaurus; pl. Thesauri**

SN A controlled vocabulary that indicates preferred terms and variant terms. In addition to the equivalence relationship, vocabulary terms in a thesaurus exhibit both hierarchical and associative relationships. These three relationships are called "standard thesaural relationships." Thesauri are usually considered the most complex of controlled vocabularies.

BT Controlled Vocabulary

RT Associative Relationship

Equivalence Relationship

Hierarchical Relationship

**User Warrant**

SN The inclusion of a vocabulary term in a controlled vocabulary based on use by users. Such terms can be identified through search log analysis or free listing.

RT Free Listing

Literary Warrant

Vocabulary Term

**Variant Term**

SN A vocabulary term that means nearly the same thing as a preferred term. Variant terms are used in the controlled vocabulary to provide entry terms that lead to preferred terms. Variant terms may include synonyms, lexical variants, quasi-synonyms and abbreviations. Variant terms are sometimes referred to as "entry terms." The collection of all variant terms may be referred to as the "entry vocabulary."

UF Alternate Term

Entry Term

Non-preferred Term

RT Equivalence Relationship

Preferred Term

**Vocabulary Term**

SN A word or phrase in a controlled vocabulary. It may be a preferred term or variant term. Vocabulary terms may exhibit several types of term relationships.

**UF Term**

NT Preferred Term

Variant Term

The resulting tool is multipurpose, as easily applicable to shelf arrangement and conventional classified card catalogues as to co-ordinate indexing and computerized retrieval systems. The reasons are given for modifying certain traditional facet techniques, including the choice of traditional disciplines for main classes, the lack of a 'built-in' preferred order, and the use, in certain instances,

of enumeration rather than synthesis to express multi-term concepts.. Methods of application of the Thesaurofacet in pre-coordinate and post-coordinate systems are discussed and brief account is given of the techniques employed in its compilation.

Notes

### 14.3 Summary

- The term “thesaurofacet” was coined by Aitchison.
- A Parenthetical Qualifier is used to identify a particular indexable meaning of a homograph
- A Scope Note is a brief statement of the intended usage of a Descriptor.
- This word-by-word alphabetical display is probably the most familiar since it provides a variety of information (a “display”) for each Descriptor.
- The Thesaurus of ERIC Descriptors, 13th Edition, contains an alphabetical listing of terms used for indexing and searching in the ERIC database.

### 14.4 Keywords

*ERIC Thesaurus* : Educational Resources Information Centre Thesaurus is a controlled vocabulary  
*UF* : Use for, is employed generally to solve problems of synonymy occurring in natural language.

### 14.5 Review Questions

1. What is ERIC Thesaurus?
2. Who coined the term “thesaurofacet”?
3. Define Scope Note.

### **Answers: Self Assessment**

1. Scope Note
2. Devices
3. Thesaurus
4. (b)      5. (c)

### 14.6 Further Readings



Books

Cooke, A : *A guide to finding quality Information on the Internet*. 2<sup>nd</sup> Edition. London: Facet Publishing, 2001.

Oddy, P. *Future libraries, future catalogs*. London: LA, 1996.

Chowdhary, GG: *Introduction to Modern Infomation Retrieval*. London: LA, 1999.



Online links

[http://www.enotes.com/topic/Education\\_Resources\\_Information\\_Center](http://www.enotes.com/topic/Education_Resources_Information_Center)

<http://besser.tsoa.nyu.edu/impact/f95/Papers-projects/Papers/perles.html>





**LOVELY PROFESSIONAL UNIVERSITY**

Jalandhar-Delhi G.T. Road (NH-1)

Phagwara, Punjab (India)-144411

For Enquiry: +91-1824-300360

Fax.: +91-1824-506111

Email: [odl@lpu.co.in](mailto:odl@lpu.co.in)