Indexing and Abstracting

Vocabulary Control

Kuang-hua Chen
Department of Library and Information Science
National Taiwan University
khchen@ntu.edu.tw

Basic Concepts

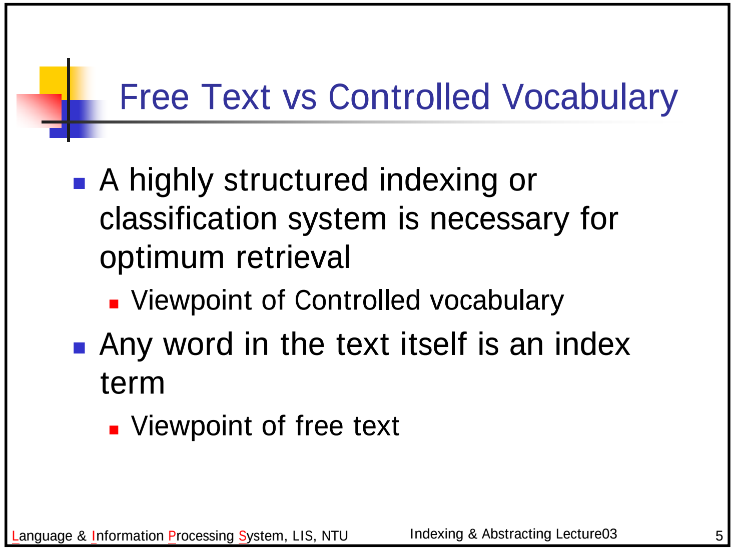
- A controlled vocabulary is a consistent set of words
- A controlled vocabulary is the terms or classification groups that have been created in order to make indexing consistent
- A natural language uncontrolled vocabulary used the words directly from the text written by the authors or the words from the indexer’s mind.

Enormous Choices of Word

- Same concept with different words
- Different concepts with same word
- The inconsistent use of words can lead to failure in searching

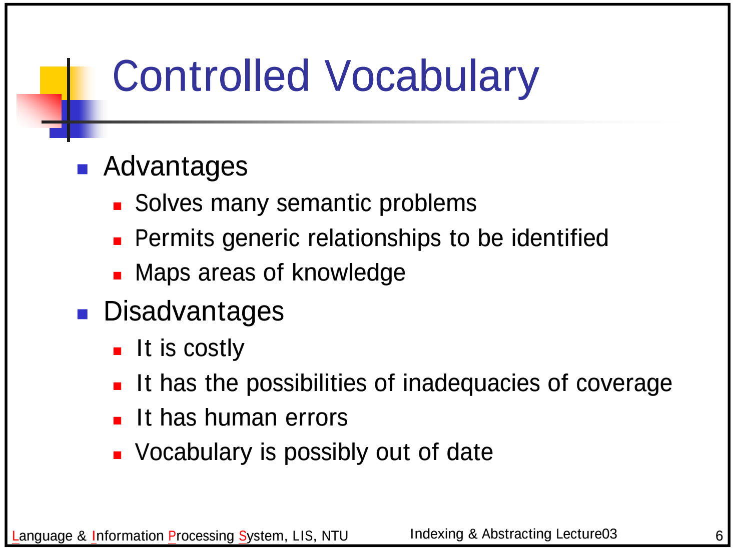
Examples of Controlled Vocabulary

- Classification schedules
- Subject authority files
- Thesauri
- …



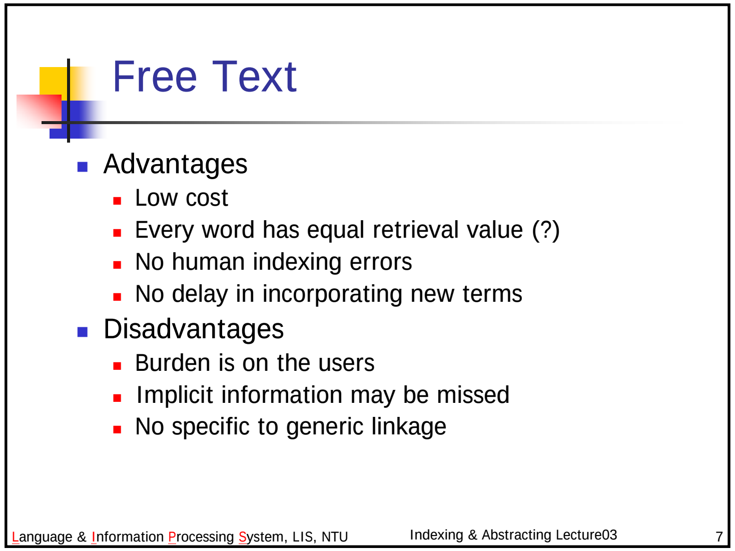
Free Text vs Controlled Vocabulary

- A highly structured indexing or classification system is necessary for optimum retrieval
  - Viewpoint of Controlled vocabulary
- Any word in the text itself is an index term
  - Viewpoint of free text



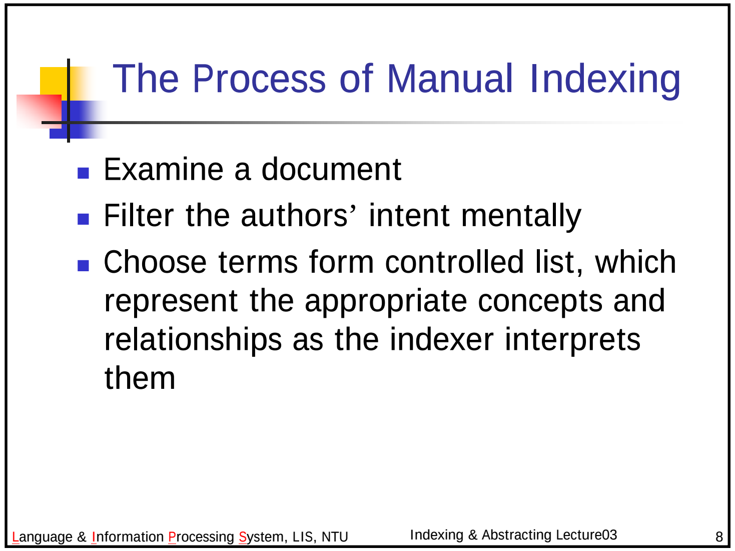
Controlled Vocabulary

- Advantages
  - Solves many semantic problems
  - Permits generic relationships to be identified
  - Maps areas of knowledge
- Disadvantages
  - It is costly
  - It has the possibilities of inadequacies of coverage
  - It has human errors
  - Vocabulary is possibly out of date



Free Text

- Advantages
  - Low cost
  - Every word has equal retrieval value (?)
  - No human indexing errors
  - No delay in incorporating new terms
- Disadvantages
  - Burden is on the users
  - Implicit information may be missed
  - No specific to generic linkage



The Process of Manual Indexing

- Examine a document
- Filter the authors' intent mentally
- Choose terms form controlled list, which represent the appropriate concepts and relationships as the indexer interprets them

Indexer vs User

- The function of control mechanism is to eventually lead both the indexer and the user to the same point

Characteristics of Controlled Vocabulary

- It represents the general conceptual structure of a subject area and presents a guide to the user of the index
- The terms are derived as nearly as possible from the vocabulary of use, that is, they closely reflect the literature vocabulary and the user’s own technical usage
- Where necessary it defines ambiguous terms
- Through cross-references it shows horizontal and vertical relationships among terms

Characteristics of Controlled Vocabulary
(continued)

- It supplies a standard vocabulary by controlling synonyms and near-synonyms in order to increase consistency
  - Only one term from a list of similar terms will be used in indexing a given concept
- It employs a considerable number of pre-coordinated phrases to reduce false drops to a minimum
  - Pre-coordinate *Venetian blind* there will not be a false drop of *blind Venetian* papers from the document file

Diversified Meanings

- Base
  - Military meaning
  - Mathematical meaning
  - Chemical meaning
- Model
  - Pattern, form, idea, measure, image, reproduction, mannequin, paragon

Controlled Vocabulary

- Generic vocabularies
  - Show vertical arrangements of words within classes
  - The related terms are seen in the context in scanning up or down (will not be scattered)
- Authority List
  - Alphabetical arrangements
  - Connect terms with relationships

Sample Generic Vocabularies

- Weapons
  - Nonconventional
    - Blowpipes
    - Boomerangs
    - Spears
    - Big rocks
  - Conventional
    - Guns
    - Rifles
    - Pistols
    - Shotguns
    - Bowie knives

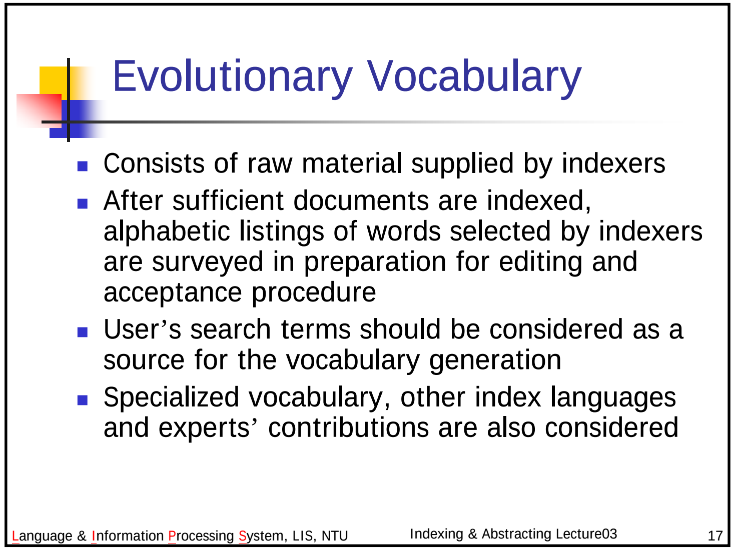
Authority List

- An authority list is a related group of words or phrases adopted by a particular group of people to be used in an indexing activity
- An authority list is a formal list of the words in the controlled vocabulary, showing the formal relationships between words and spelling out how they are to be used
- Ambiguity is solved by referring to the authority list as the final arbitrator in vocabulary control

- Subject headings
- Thesaurus

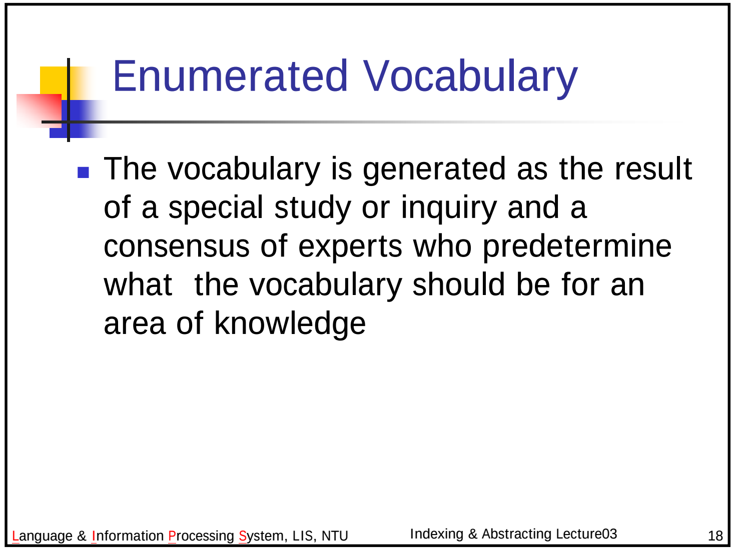
Construct Authority Lists

- Evolutionary Vocabulary
- Enumerated Vocabulary



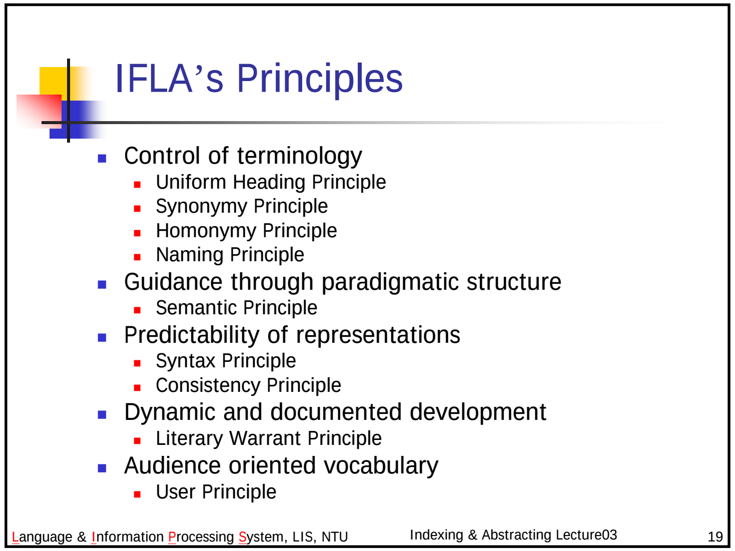
Evolutionary Vocabulary

- Consists of raw material supplied by indexers
- After sufficient documents are indexed, alphabetic listings of words selected by indexers are surveyed in preparation for editing and acceptance procedure
- User's search terms should be considered as a source for the vocabulary generation
- Specialized vocabulary, other index languages and experts' contributions are also considered



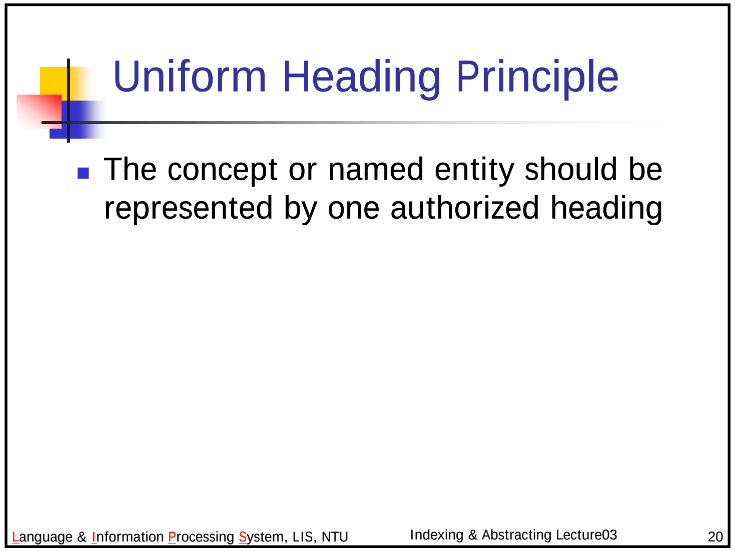
Enumerated Vocabulary

- The vocabulary is generated as the result of a special study or inquiry and a consensus of experts who predetermine what the vocabulary should be for an area of knowledge



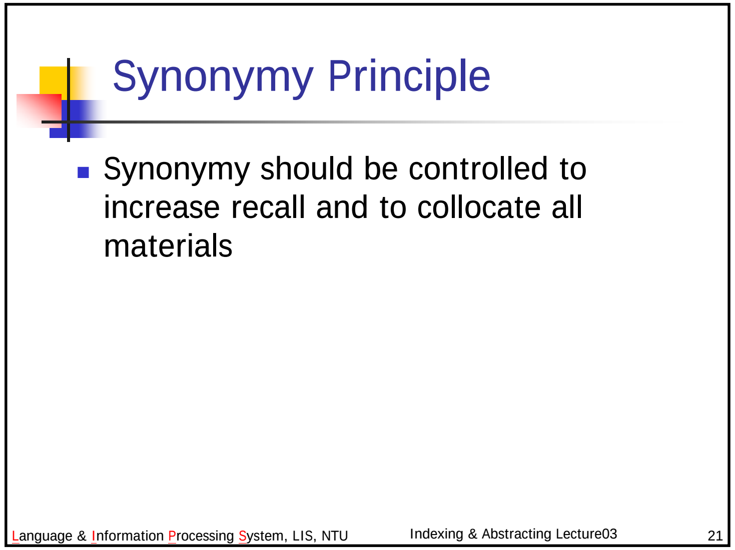
IFLA's Principles

- Control of terminology
  - Uniform Heading Principle
  - Synonymy Principle
  - Homonymy Principle
  - Naming Principle
- Guidance through paradigmatic structure
  - Semantic Principle
- Predictability of representations
  - Syntax Principle
  - Consistency Principle
- Dynamic and documented development
  - Literary Warrant Principle
- Audience oriented vocabulary
  - User Principle



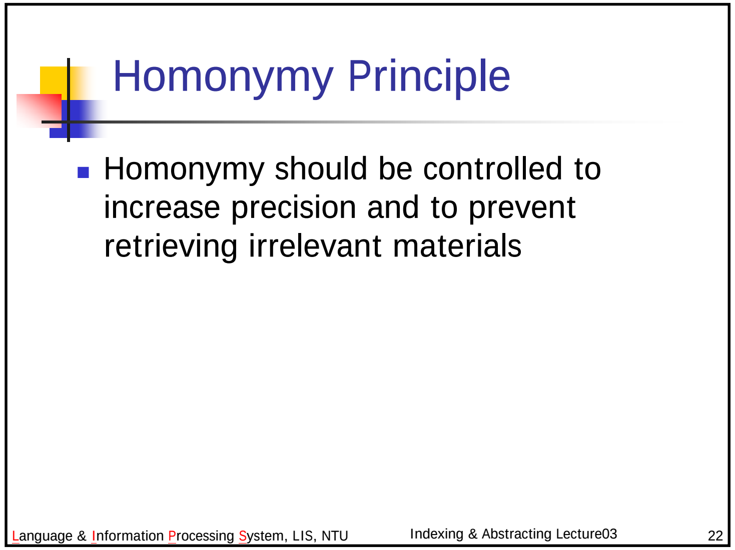
Uniform Heading Principle

- The concept or named entity should be represented by one authorized heading



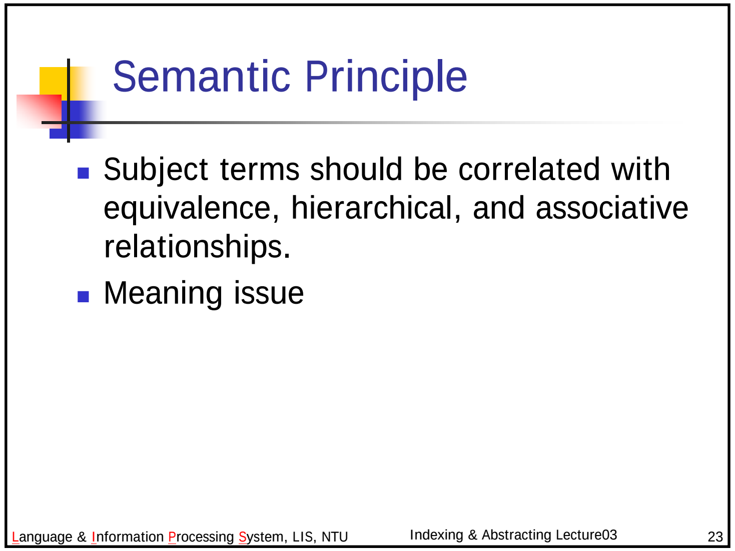
Synonymy Principle

- Synonymy should be controlled to increase recall and to collocate all materials



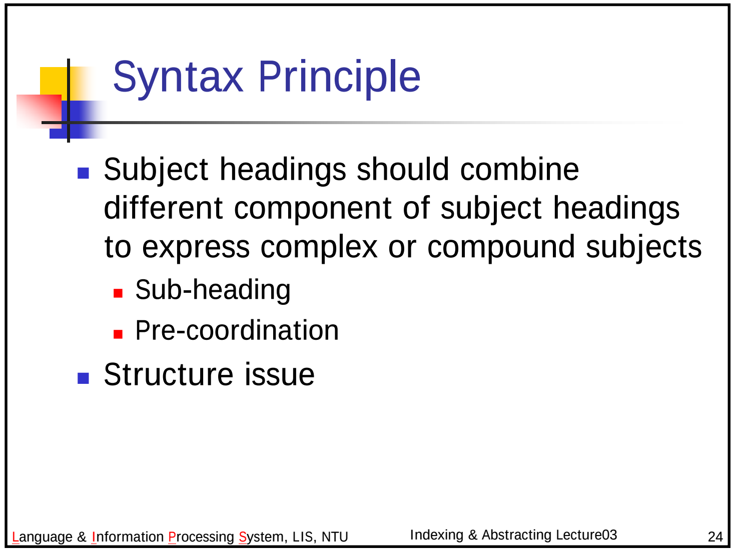
Homonymy Principle

- Homonymy should be controlled to increase precision and to prevent retrieving irrelevant materials



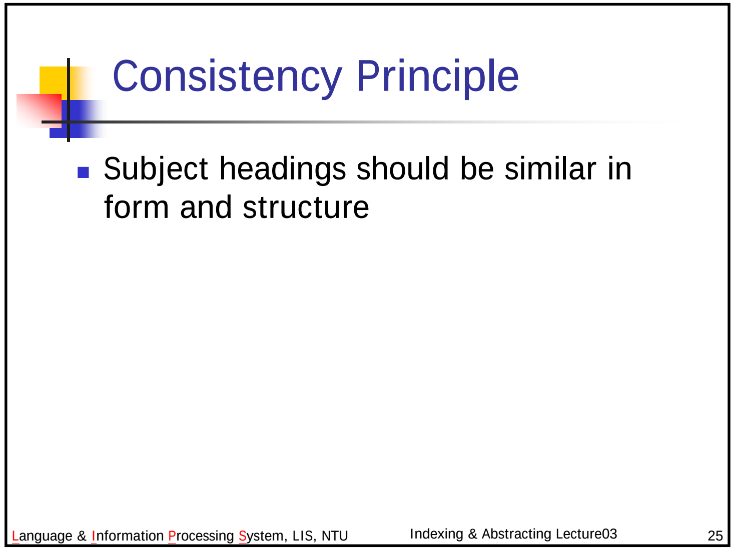
Semantic Principle

- Subject terms should be correlated with equivalence, hierarchical, and associative relationships.
- Meaning issue



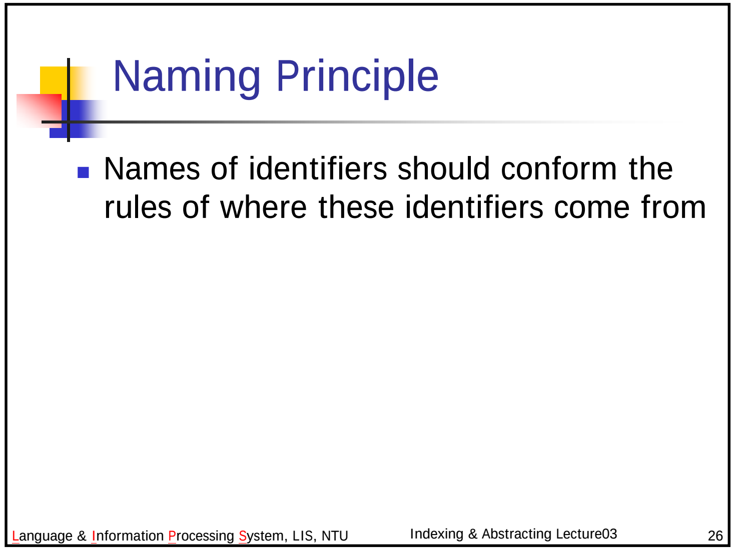
Syntax Principle

- Subject headings should combine different component of subject headings to express complex or compound subjects
  - Sub-heading
  - Pre-coordination
- Structure issue



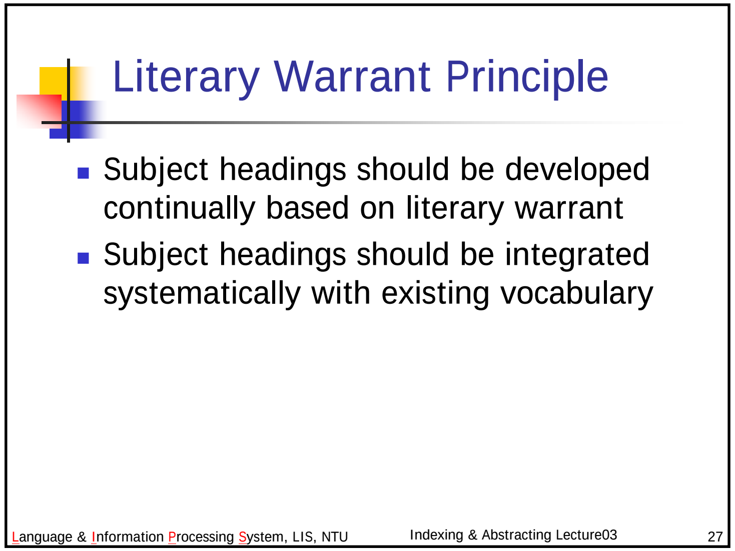
Consistency Principle

- Subject headings should be similar in form and structure



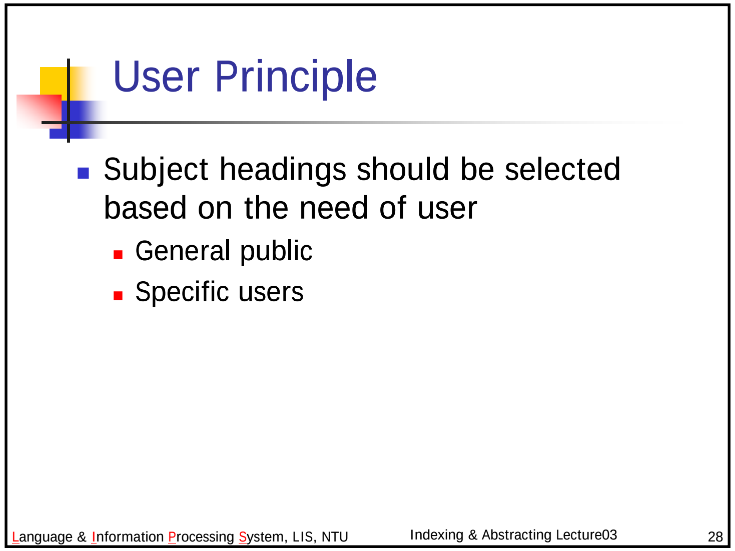
Naming Principle

- Names of identifiers should conform the rules of where these identifiers come from



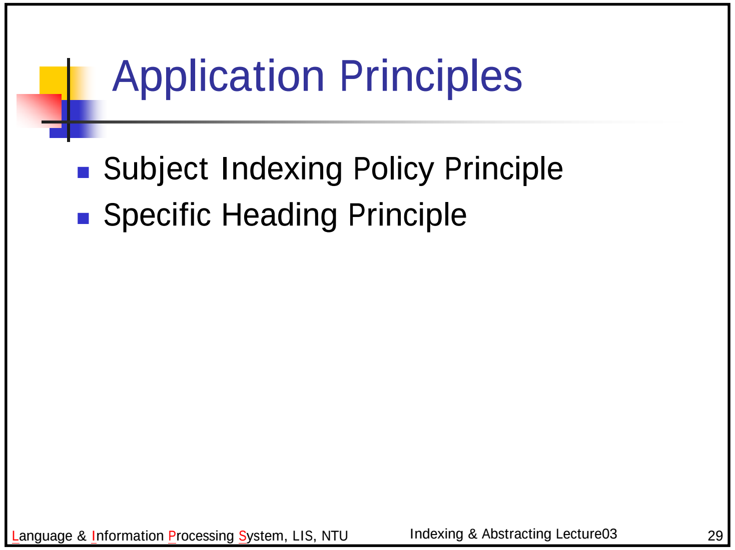
Literary Warrant Principle

- Subject headings should be developed continually based on literary warrant
- Subject headings should be integrated systematically with existing vocabulary



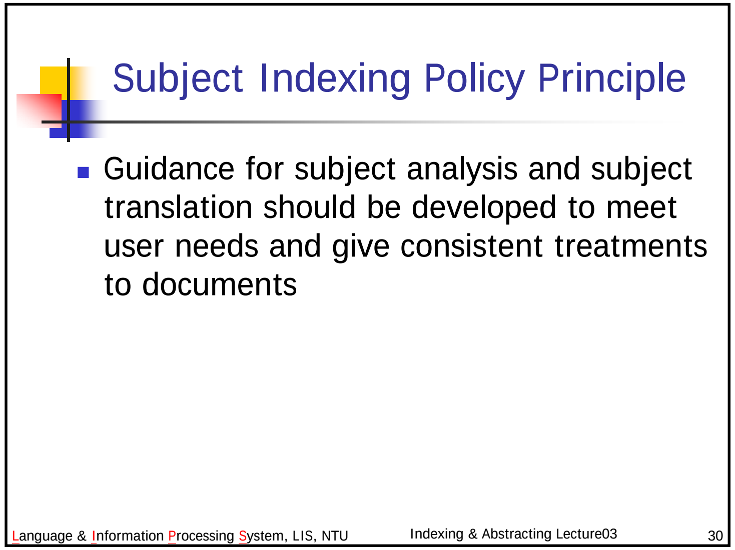
User Principle

- Subject headings should be selected based on the need of user
  - General public
  - Specific users



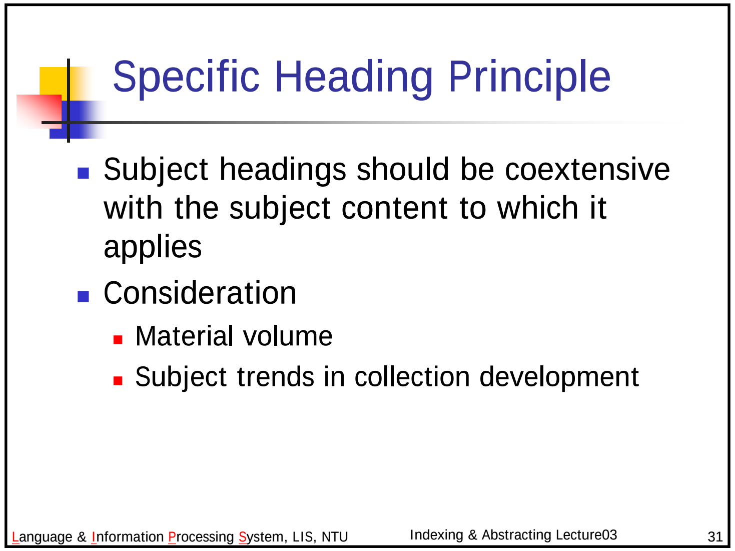
Application Principles

- Subject Indexing Policy Principle
- Specific Heading Principle



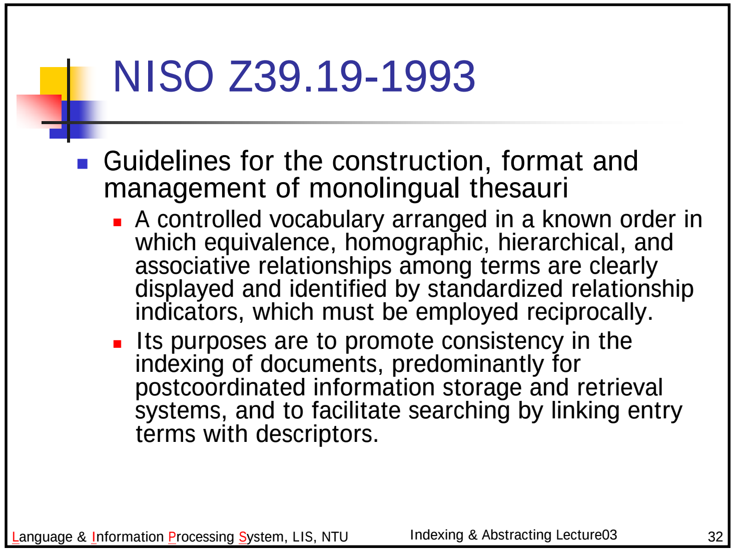
Subject Indexing Policy Principle

- Guidance for subject analysis and subject translation should be developed to meet user needs and give consistent treatments to documents



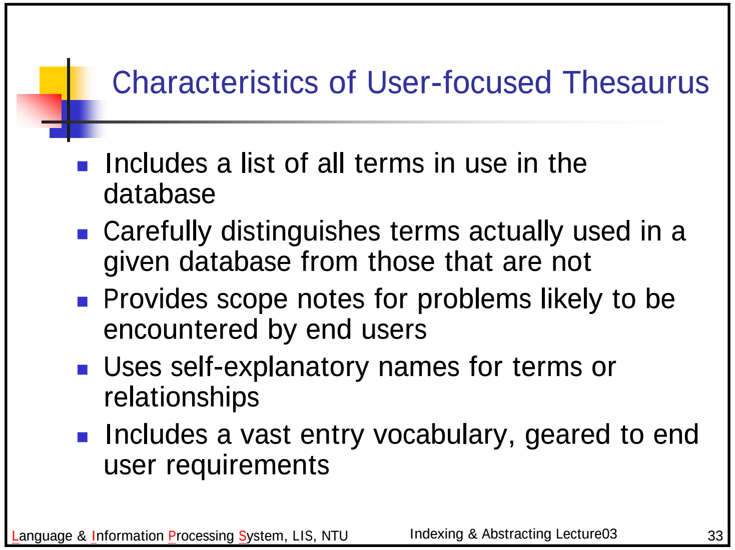
Specific Heading Principle

- Subject headings should be coextensive with the subject content to which it applies
- Consideration
  - Material volume
  - Subject trends in collection development



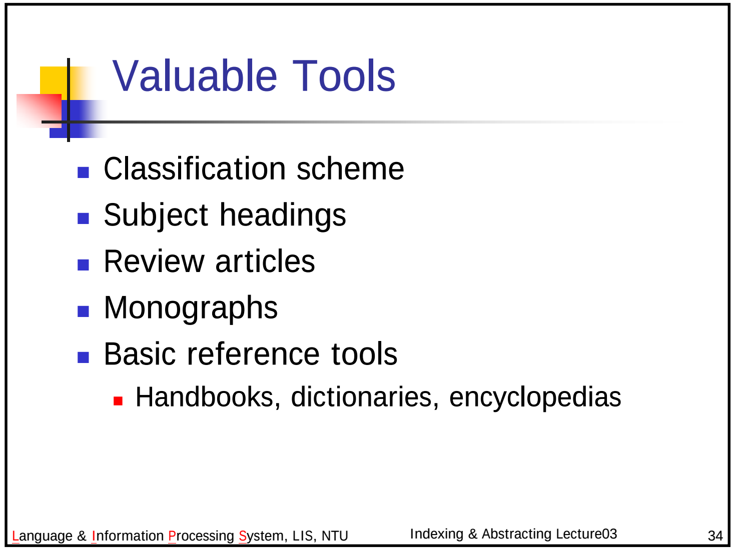
NISO Z39.19-1993

- Guidelines for the construction, format and management of monolingual thesauri
  - A controlled vocabulary arranged in a known order in which equivalence, homographic, hierarchical, and associative relationships among terms are clearly displayed and identified by standardized relationship indicators, which must be employed reciprocally.
  - Its purposes are to promote consistency in the indexing of documents, predominantly for postcoordinated information storage and retrieval systems, and to facilitate searching by linking entry terms with descriptors.



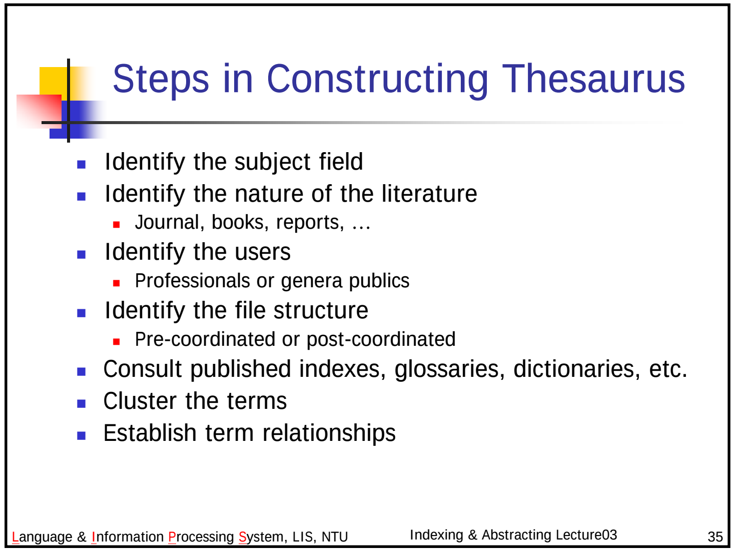
Characteristics of User-focused Thesaurus

- Includes a list of all terms in use in the database
- Carefully distinguishes terms actually used in a given database from those that are not
- Provides scope notes for problems likely to be encountered by end users
- Uses self-explanatory names for terms or relationships
- Includes a vast entry vocabulary, geared to end user requirements



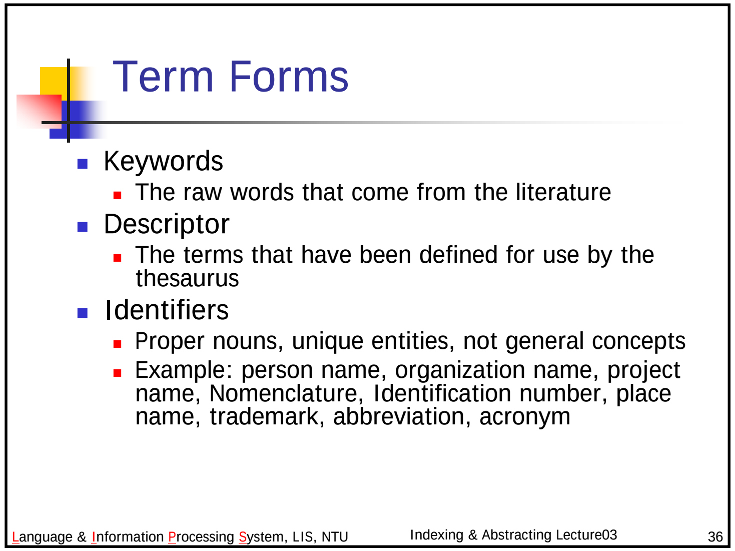
Valuable Tools

- Classification scheme
- Subject headings
- Review articles
- Monographs
- Basic reference tools
  - Handbooks, dictionaries, encyclopedias



Steps in Constructing Thesaurus

- Identify the subject field
- Identify the nature of the literature
  - Journal, books, reports, …
- Identify the users
  - Professionals or genera publics
- Identify the file structure
  - Pre-coordinated or post-coordinated
- Consult published indexes, glossaries, dictionaries, etc.
- Cluster the terms
- Establish term relationships



Term Forms

- Keywords
  - The raw words that come from the literature
- Descriptor
  - The terms that have been defined for use by the thesaurus
- Identifiers
  - Proper nouns, unique entities, not general concepts
  - Example: person name, organization name, project name, Nomenclature, Identification number, place name, trademark, abbreviation, acronym

Term Forms (Continued)

- Preferred terms
  - The words chosen for the thesaurus to represent a class of synonymous words
- Entry terms
  - Words that allow the user to enter the vocabulary structure
  - If an entry term is an allowable descriptor it will refer user to a term that is acceptable
  - A strong, full entry vocabulary will enhance the user’s chances for finding the right words in the search

Decision for Term Forms

- Descriptor should be nouns, either single nouns, noun phrases or nouns with qualifiers indicated in parentheses
- Multiword terms may be either pre-coordinated or formed by post-coordination of existing terms
- Singular form for processes and properties; Plural form for classes of people
  - Liquidation, Indexing – processes
  - Teachers, preachers, candlestick makers -- classes
- Multiword terms should be entered in their natural word order with see cross-references to the inverted forms
- Abbreviations should be used if the users know their meaning

Term Relationships

- Equivalence
  - USE
  - UF
- Hierarchical
  - Broader term (BT)
  - Narrower terms (NT)
- Associative
  - Related term (RT)

USE and Use For

- Use (USE)
  - Refer to a preferred descriptor from a nonusable term
  - Is a reciprocal of a USE FOR (UF)
- Use for (UF)
  - Deal primarily with synonyms or variant forms of the preferred descriptor
  - Also be used to lead the indexer to more general term

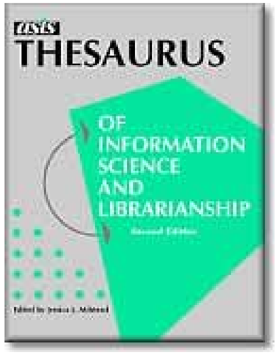
Examples of USE and UF

- Pecan trees
  - USE TREES
- Oak trees
  - USE TREES
- TREES
  - UF Pecan trees
- PROMOTION POLICIES
  - UF Automatic promotion
- Automatic promotion
  - USE PROMOTION POLILCIES

Scope Note (SN)

- Brief description of the sense or framework in which the terms should be used
- Restrict the usage of a description
- Clarify the ambiguity
- Example
  - CULTURAL BACKGROUND
    - SN The total social heritage and experience of an individual or group including institutions, folkways, literature, mores, and communal experience

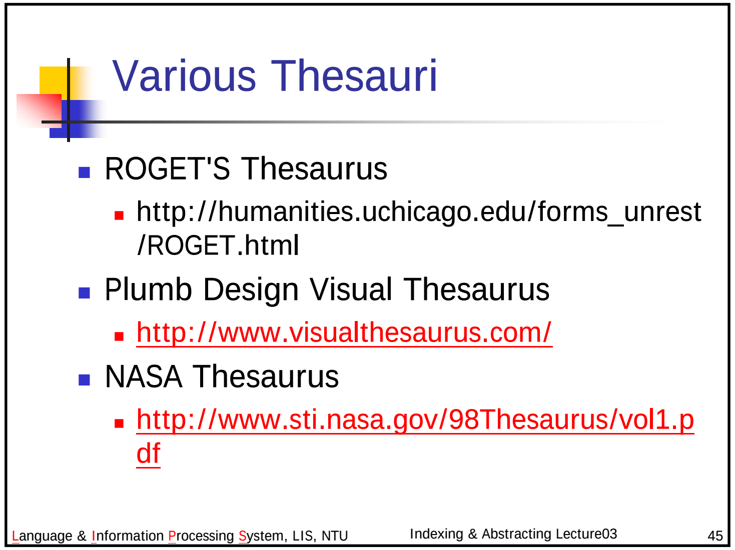
ASIS Thesaurus


ASIS
THESAURUS
OF INFORMATION SCIENCE AND LIBRARIANSHIP
Second Edition
Edited by Jessica L. Milstead

Example of UNESCO Thesaurus

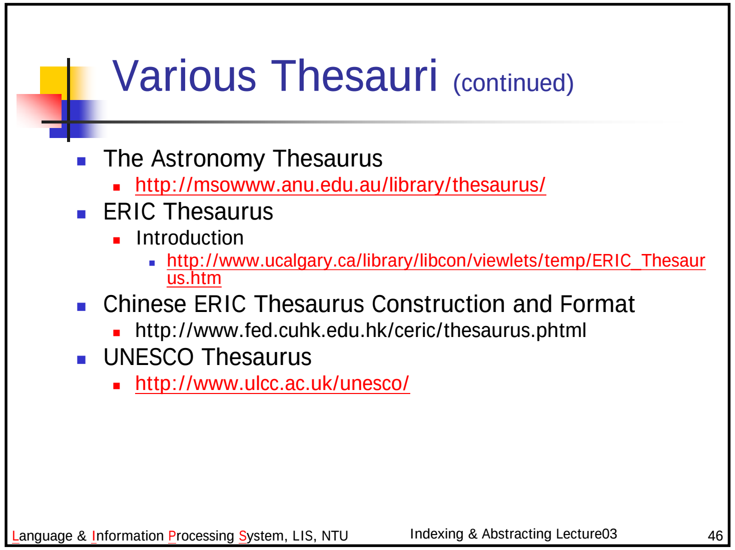
Computer programming

Narrower Term
NT1 Computer languages
UF Programming languages
NT1 Computer software
UF Computer programs, Software packages
NT2 Application software
NT2 Basic software
UF Operating systems
NT1 File organization
UF Computer storage organization
NT2 Data formats
UF Data layout
NT2 Random access
UF Direct access
NT1 Multiuser systems
UF Timesharing systems
NT1 Online systems
UF Interactive online systems, Realtime systems



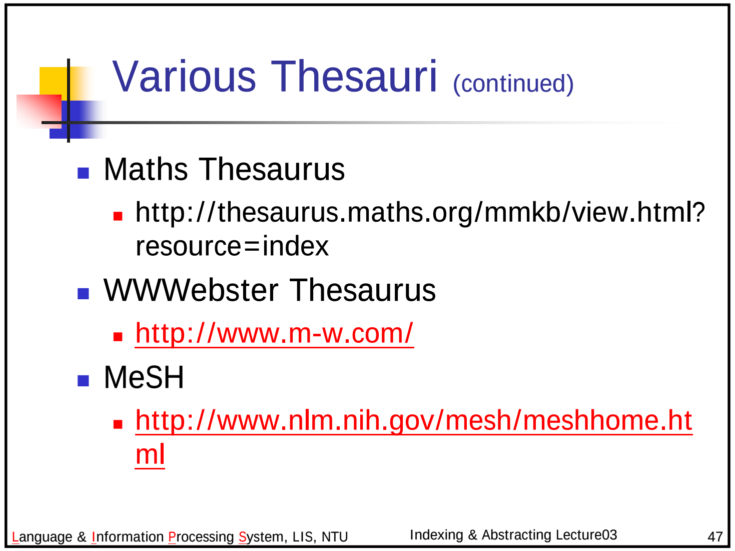
Various Thesauri

- ROGET'S Thesaurus
  - http://humanities.uchicago.edu/forms_unrest/ROGET.html
- Plumb Design Visual Thesaurus
  - http://www.visualthesaurus.com/
- NASA Thesaurus
  - http://www.sti.nasa.gov/98Thesaurus/vol1.pdf



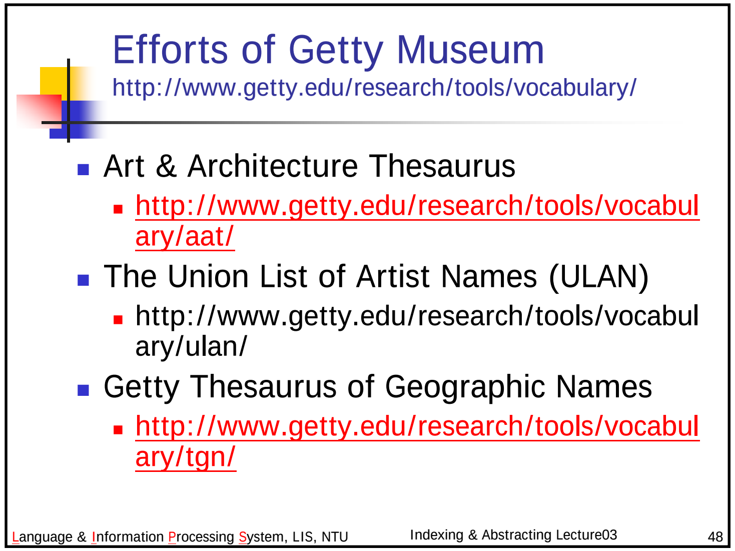
Various Thesauri (continued)

- The Astronomy Thesaurus
  - http://msowww.anu.edu.au/library/thesaurus/
- ERIC Thesaurus
  - Introduction
    - http://www.ucalgary.ca/library/libcon/viewlets/temp/ERIC_Thesaurus.htm
- Chinese ERIC Thesaurus Construction and Format
  - http://www.fed.cuhk.edu.hk/ceric/thesaurus.phtml
- UNESCO Thesaurus
  - http://www.ulcc.ac.uk/unesco/



Various Thesauri (continued)

- Maths Thesaurus
  - http://thesaurus.maths.org/mmkb/view.html?resource=index
- WWWebster Thesaurus
  - http://www.m-w.com/
- MeSH
  - http://www.nlm.nih.gov/mesh/meshhome.html



Efforts of Getty Museum

http://www.getty.edu/research/tools/vocabulary/

- Art & Architecture Thesaurus
  - http://www.getty.edu/research/tools/vocabulary/aat/
- The Union List of Artist Names (ULAN)
  - http://www.getty.edu/research/tools/vocabulary/ulan/
- Getty Thesaurus of Geographic Names
  - http://www.getty.edu/research/tools/vocabulary/tgn/

Near Comprehensive list

- Web Thesaurus Compendium
  - http://www.ipsi.fraunhofer.de/~lutes/thesoecd.html

Lexical Freenet http://www.lexfn.com/

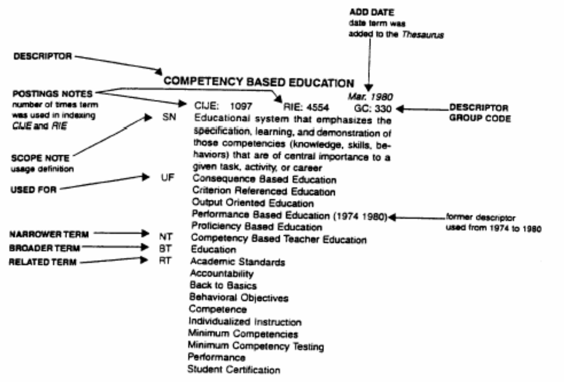
Relation	Example		
Allow trigger links	Clinton	→	Whitewater
Allow synonym links	bike	↔	bicycle
Allow generalization links	tree	→	acacia
Allow specialization links	shoe	←	footwear
Allow comprises links	Turkey	←	Istanbul
Allow part-of links	CPU	→	computer
Allow antonym links	opaque	↔	clear
Allow rhyme links	Reno	↔	casino
Allow sounds-like links	candle	↔	cancel
Allow anagram links	Geraldine	↔	realigned
Allow occupation links	Leonardo da Vinci	←	painter
Allow nationality links	Martin Luther	→	German
Allow birth year links	Orville Wright	→	1871
Allow death year links	Gilda Radner	→	1989
Allow biographical trigger links	Jesse Louis Jackson	→	rainbow
Allow *also known as* links	John Ono Lennon	↔	John Lennon

Display Formats of Thesaurus

- Alphabetical Descriptor Display
- Rotated Descriptor Display
- Hierarchical Descriptor Display
- Descriptor Group Display

  (Classified Descriptor Display)

Alphabetical Descriptor Display



COMPETENCY BASED EDUCATION

Mar. 1980

CIJE: 1097 RIE: 4554 GC: 330

SN Educational system that emphasizes the specification, learning, and demonstration of those competencies (knowledge, skills, behaviors) that are of central importance to a given task, activity, or career

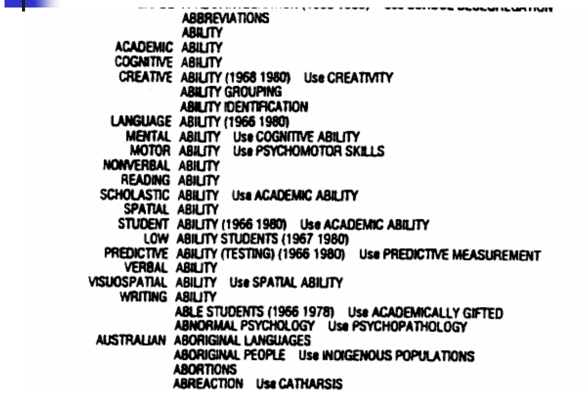
UF Consequence Based Education
Criterion Referenced Education
Output Oriented Education
Performance Based Education (1974 1980)
Proficiency Based Education

NT Competency Based Teacher Education

BT Education

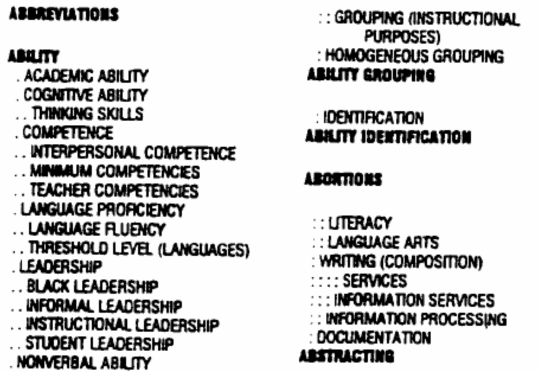
RT Academic Standards
Accountability
Back to Basics
Behavioral Objectives
Competence
Individualized Instruction
Minimum Competencies
Minimum Competency Testing
Performance
Student Certification

Rotated Descriptor Display



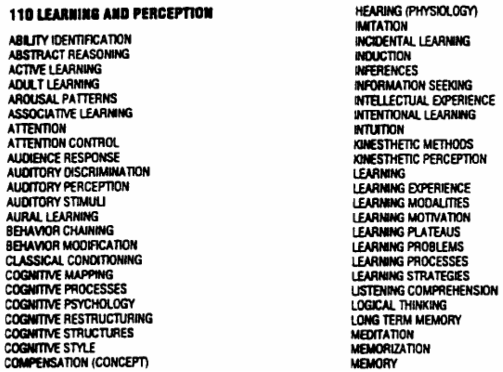
```
            ABBREVIATIONS
            ABILITY
   ACADEMIC ABILITY
  COGNITIVE ABILITY
   CREATIVE ABILITY (1968 1980)   Use CREATIVITY
            ABILITY GROUPING
            ABILITY IDENTIFICATION
   LANGUAGE ABILITY (1966 1980)
     MENTAL ABILITY   Use COGNITIVE ABILITY
      MOTOR ABILITY   Use PSYCHOMOTOR SKILLS
  NONVERBAL ABILITY
    READING ABILITY
 SCHOLASTIC ABILITY   Use ACADEMIC ABILITY
    SPATIAL ABILITY
    STUDENT ABILITY (1966 1980)   Use ACADEMIC ABILITY
        LOW ABILITY STUDENTS (1967 1980)
 PREDICTIVE ABILITY (TESTING) (1966 1980)   Use PREDICTIVE MEASUREMENT
     VERBAL ABILITY
VISUOSPATIAL ABILITY   Use SPATIAL ABILITY
    WRITING ABILITY
            ABLE STUDENTS (1966 1978)   Use ACADEMICALLY GIFTED
            ABNORMAL PSYCHOLOGY   Use PSYCHOPATHOLOGY
 AUSTRALIAN ABORIGINAL LANGUAGES
            ABORIGINAL PEOPLE   Use INDIGENOUS POPULATIONS
            ABORTIONS
            ABREACTION   Use CATHARSIS
```

Hierarchical Descriptor Display



```
ABBREVIATIONS

ABILITY
. ACADEMIC ABILITY
. COGNITIVE ABILITY
. . THINKING SKILLS
. COMPETENCE
. . INTERPERSONAL COMPETENCE
. . MINIMUM COMPETENCIES
. . TEACHER COMPETENCIES
. LANGUAGE PROFICIENCY
. . LANGUAGE FLUENCY
. . THRESHOLD LEVEL (LANGUAGES)
. LEADERSHIP
. . BLACK LEADERSHIP
. . INFORMAL LEADERSHIP
. . INSTRUCTIONAL LEADERSHIP
. . STUDENT LEADERSHIP
. NONVERBAL ABILITY

 : : GROUPING (INSTRUCTIONAL
       PURPOSES)
 : HOMOGENEOUS GROUPING
ABILITY GROUPING

 : IDENTIFICATION
ABILITY IDENTIFICATION

ABORTIONS

 : : LITERACY
 : : LANGUAGE ARTS
 : WRITING (COMPOSITION)
 : : : : SERVICES
 : : : INFORMATION SERVICES
 : : INFORMATION PROCESSING
 : DOCUMENTATION
ABSTRACTING
```

Descriptor Group Display



110 LEARNING AND PERCEPTION

ABILITY IDENTIFICATION
ABSTRACT REASONING
ACTIVE LEARNING
ADULT LEARNING
AROUSAL PATTERNS
ASSOCIATIVE LEARNING
ATTENTION
ATTENTION CONTROL
AUDIENCE RESPONSE
AUDITORY DISCRIMINATION
AUDITORY PERCEPTION
AUDITORY STIMULI
AURAL LEARNING
BEHAVIOR CHAINING
BEHAVIOR MODIFICATION
CLASSICAL CONDITIONING
COGNITIVE MAPPING
COGNITIVE PROCESSES
COGNITIVE PSYCHOLOGY
COGNITIVE RESTRUCTURING
COGNITIVE STRUCTURES
COGNITIVE STYLE
COMPENSATION (CONCEPT)
HEARING (PHYSIOLOGY)
IMITATION
INCIDENTAL LEARNING
INDUCTION
INFERENCES
INFORMATION SEEKING
INTELLECTUAL EXPERIENCE
INTENTIONAL LEARNING
INTUITION
KINESTHETIC METHODS
KINESTHETIC PERCEPTION
LEARNING
LEARNING EXPERIENCE
LEARNING MODALITIES
LEARNING MOTIVATION
LEARNING PLATEAUS
LEARNING PROBLEMS
LEARNING PROCESSES
LEARNING STRATEGIES
LISTENING COMPREHENSION
LOGICAL THINKING
LONG TERM MEMORY
MEDITATION
MEMORIZATION
MEMORY

To sum up, Thesaurus

- Provide the control of terminology by
  - showing a structural display of concepts
  - Supplying for each concept all terms that might express that concept
  - Presenting the associate and hierarchical relationships of vocabulary
- An alphabetical list of all the words and phrases making up the controlled vocabulary

General Rules of Thesaurus Evaluation

- Authority
- Proven usefulness
- Regular revision
- Ease of use

Questions for Thesaurus

- How good is the subject coverage of the concepts displayed? Is it adequate to allow proper indexing and searching
- How well does the thesaurus handle broader terms, narrower terms, and related terms? In other words, are all the structural relationships between terms treated adequately?
- How adequate is the display of the thesaurus? Is it easy to see, understand, and follow through on? Does it lead to efficient and effective indexing and searching?

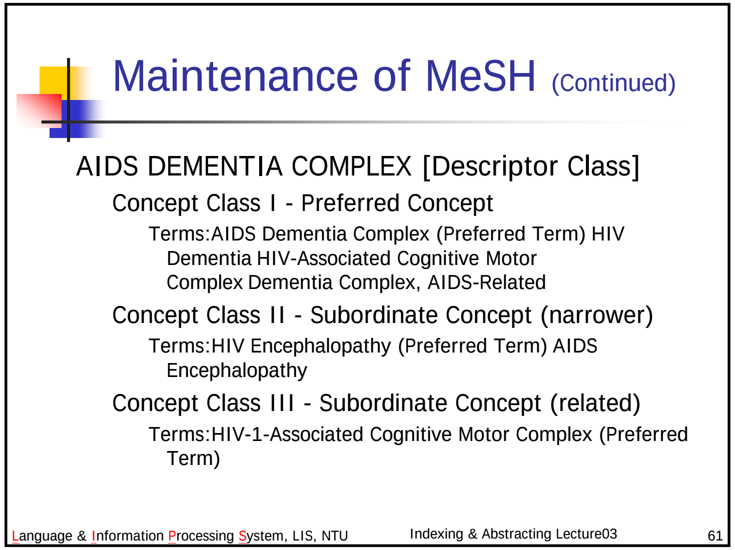
Maintenance of Thesaurus

- Cost
- Labor-intensive
- The work on a particular thesaurus is never finished

Maintenance of MeSH

AIDS DEMENTIA COMPLEX

HIV Dementia (equivalent)
HIV-Associated Cognitive Motor Complex (equivalent)
Dementia Complex, AIDS-Related (equivalent)
HIV Encephalopathy (narrower)
AIDS Encephalopathy (narrower)
HIV-1-Associated Cognitive Motor Complex (related)



Maintenance of MeSH (Continued)

AIDS DEMENTIA COMPLEX [Descriptor Class]

- Concept Class I - Preferred Concept
  - Terms:AIDS Dementia Complex (Preferred Term) HIV Dementia HIV-Associated Cognitive Motor Complex Dementia Complex, AIDS-Related
- Concept Class II - Subordinate Concept (narrower)
  - Terms:HIV Encephalopathy (Preferred Term) AIDS Encephalopathy
- Concept Class III - Subordinate Concept (related)
  - Terms:HIV-1-Associated Cognitive Motor Complex (Preferred Term)