

**Eötvös Loránd University  
Budapest (Hungary)**

---

**Indexing languages:  
comparative analysis between categorial  
languages (classifications) and  
vocabular languages (combinatorial)**

**30th April 2015**

**FERNANDA RIBEIRO**  
**[fribeiro-flup@gmail.com](mailto:fribeiro-flup@gmail.com)**

# Indexing Languages

- **SUBJECT INDEXING** – procedure that consists in describing and characterizing a document in terms of its subject content, that is, concepts are extracted from documents by a process of analysis and then transcribed into the elements of an indexing language, such as thesauri, classification schemes, etc.  
(In : *Indexing Principles* / UNISIST)

From the definition of indexing we can understand that:

- - indexing implies a process of analysis
- - in the information exist concepts (ideas), which are represented through an indexing language (lead terms, access points)

We still add:

- - indexing has a purpose: to enable an effective research of information
- - search is done through an “indexing language” (recorded in the indexing tools) and supposes also the indexing of the question

# Indexing Languages

---

- **SUBJECT INDEXING STAGES**

- 1<sup>st</sup> – Concepts Recognition / Identification

- 2<sup>nd</sup> - Concepts Representation

- Concept → Indexing term

  - {Keyword

- {Descriptor

- {Non-descriptor

- {Classification code

Indexing term – it is the vocabular expression that translates the concept; it can be simple or compound

# Indexing Languages

---

- **2<sup>nd</sup> STAGE OF INDEXING - Concepts representation**

- The representation of concepts is done through the indexing language and with the aid of the instruments that support indexing in which it is registered.

- **Indexing language:**

- - It is a condensation and simplification of natural language
- - It is a conventional, controlled, intermediate language between the user and the information
- - It is a system of representation of the informational content and of user queries, to facilitate the search of information
- - It is formed by a set of indexing terms whose logical organization or orderly presentation varies according to the type of language used

# Indexing Languages

---

The use of indexing language is a necessity given the characteristics of the natural language itself

In the natural language there are implicit relationships between words and terms:

- 1 - equivalence relationships
- 2 - hierarchical relationships
- 3 - associative relationships

# Indexing Languages

## Equivalence relationship

- Examples:
- ancient term / current term      butcher / butchery
- ordinary term / erudit term      sore throat / tonsillitis
- usual term / scientific term      salt / sodium chloride
- synonymous      pastry / confectionery
- almost synonymous      house / residential building

## Hierarchical relationships

- Examples:
- genre / specimen      vegetables / lettuce, tomato...
- whole / part      Hungary / Budapest

## Associative relationships

- Examples:
- school / professor / student
- fair / market
- industry / factory / industrial product

# Indexing Languages

---

## Characteristics of the indexing language:

- - It is a controlled vocabulary formed by a limited set of terms that imposes itself as a standard
- - It is recorded in an instrument of support for the indexer and the user
- - Consists of the terms used in indexing (preferred terms) and the unused (non-preferred), which direct to the adopted [equivalence relationships]
- - Presents explicitly, hierarchical semantic relations and associative relations, which gives it the form of a structured vocabulary

# Indexing Languages

---

## Constitutive elements of the indexing language:

- - Terms that are access points to information
- - Non-preferred terms, which are not access points to information and which lead to the preferential ones
- - Relations between indexing terms
- - Notations (symbols)
- - Scope notes
- - Syntax elements (abbreviations, graphic signs, ...) that help to explain the relationships between indexing terms

# Indexing Languages

---

Essential notions:

- - **Coordination** - syntactic relationship between the various concepts that constitute the informational content
- - **Authority control** - a process that allows to maintain consistency in access points, by connecting to a file (or list) of authority, constituted by the set of "authority records"
- - **Authorized term** - access point to information, included in an indexing controlled vocabulary and which represents a concept

# Indexing Languages

---

## Types of indexing languages

### Pre-coordinated

Categorical languages / classifications  
classification schemes  
(coordination at indexing time)

### Post-coordinated

combinatory languages  
thesauri, structured lists  
(coordination at searching time)

Categorical languages - subjects are logically organized in a structure made *a priori* (pre-coordination), i.e., independent of informational content; the indexer has to "fit" the subjects extracted from the information in these logical frameworks

Combinatorial languages - are formed by separate terms, extracted from natural language capable of varying combinations *a posteriori* (post-coordination), arising from the informational content

# Indexing Languages

---

## CATEGORIAL LANGUAGE

- - principle of classification
- - subjects are integrated in a rigid structure that constitutes the universe of knowledge
- - this universe is organized in a fix structure divided into classes and subclasses – classification scheme
- - each item of the classification scheme identifies a subject (from the information or from the query)
- - the items are exclusive and in construction are hierarchical; the hierarchical relationships can not be changed without changing the indexing of content already done
- - a classification scheme is based on Aristotelian logic and is deductive (from general to particular)

# Indexing Languages

---

- - Indexing with a categorical language means: identifying the main subject, assigning the most appropriate notation to that topic
- - in search: search in the shelves (systematic arrangement); search in the systematic catalog
- - an indexed subject can be found in all components of the same classification hierarchy (superordinate)
- Example: Cats' feed with canned products
- UDC notation - 599.742.7:613.2:664.8 (only 1 access point)
- Search:
- Animals
- - Mammals
- -- Cats
- --- Cats' feed
- ---- Cats' feed with canned products

# Indexing Languages

---

## COMBINATORIAL LANGUAGE

- - combinatorial principle
- - subjects are analyzed and represented in concepts, whose evolution characterizes the evolution of knowledge
- - Concepts are combined in a thesaurus, with a flexible structure, which is more devoted to searching than fixing the relations between indexing terms
- - each item identifies a concept (from the information or from the query)
- - items are exclusive in construction and are independent of each other, the relationships between them can easily be changed without changing the indexing already done
- - a thesaurus is constructed by inductive and analytical via (collect terms from the reality)

# Indexing Languages

---

- - Indexing with a combinatorial language means: identify concepts; represent them by the most appropriate indexing terms
- - Search: search in the ideographic catalogue in all indexing terms representative of the various concepts and their combinations
- - an indexed document can be found in all combinations of its concepts
- Example: Cats' feed with canned products
- Indexing terms – Feed / Cats / Canned products (3 access points)
- Search:
  - Feed / Cats / Canned products
  - Feed / Cats
  - Feed / Canned products
  - Cats