

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

"He dieth not who giveth life to learning"
Muhammad (P.B.U.H)

Subject: Online Information Retrieval (LISC-43203)

Topic: *Controlled Vocabulary vs. Natural Language and Authority Control*



Contents

1. Introduction & Definition
2. Idea of Controlled Vocabulary
3. Purpose of Controlled Vocabularies
4. Purpose of Controlled Vocabulary in LIS
5. Kinds of Controlled Vocabularies
6. Principle for Creation & Application of Controlled Vocabulary
7. Problems with Controlled Vocabulary
8. Natural Language & Vocabularies System
9. Natural Languages and Indexing
10. Importance of Natural Languages
11. What is Authority Control & Authorities
12. Authority Control
13. Authority Records & “Cards”
14. An Authority Card Contains
15. Benefits of Authority Control

Introduction & Definition

A controlled vocabulary is an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms and has a defined scope or describes a specific domain.

- *“Organized lists of words and phrases, or notation systems, that are used to initially tag content, and then to find it through navigation (map-reading) or search.” (Amy Warner)*
- *“a controlled vocabulary is a type of metadata that functions as a “subset of natural language.”*
- A controlled vocabulary is a list of terms or other symbols used in indexing.

Idea of Controlled Vocabulary

1. The idea of a controlled vocabulary is to reduce the inconsistency of expressions used to characterize the document being indexed, e.g. by avoiding *synonyms* and *remove ambiguity*.
2. By principle one is only allowed to use terms from the controlled vocabulary in the indexing process.
3. If a relevant term is missing from the controlled vocabulary, the indexer might suggest that the term is added to the list.

The need for vocabulary control arises from two basic features of natural language, namely:

- Two or more words or terms can be used to represent a single concept

Example: VHF/Very High Frequency

- ▶ Two or more words that have the same spelling can represent different concepts

Example:

Base (air base)

Base (civil engineers)

Base (computer science)

Base (beautician)

Purpose of Controlled Vocabularies

Controlled vocabularies serve five purposes:

- 1. Translation:** Provide a means for converting the natural language of authors, indexers, and users into a vocabulary can be used for indexing and retrieval.
- 2. Consistency:** Promote uniformity in term format and in the assignment of terms.
- 3. Indication of relationships:** Indicate semantic relationships among terms.
- 4. Label and browse:** Provide consistent and clear hierarchies in a navigation system to help users locate desired content objects.
- 5. Retrieval:** Serve as a searching aid in locating content objects.

Purpose of Controlled Vocabulary in Library & Information Science

1. 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.
2. Controlled vocabularies solve the problems of *homographs* and *synonyms* between concepts and authorized terms.
3. In short, controlled vocabularies reduce ambiguity natural in normal human languages where the same concept can be given different names and ensure consistency.

Kinds of Controlled Vocabularies

1. Subject Heading

A subject heading is part of a systematic list of terms that describe a given subject matter, e.g. like in a library catalogue.

2. Thesauri

a book that lists words in groups of synonyms and related concepts

3. Ontology

In computer science and information science, an ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really exist in a particular domain of discourse.

Principle for Creation & Application of Controlled Vocabulary (CV)

- ▶ **Specificity**- the level of hierarchical depth in the concepts
- ▶ **Literary warrant** - terminology is added to a subject heading list or thesaurus when a new concept shows up in the information resources that need organizing and therefore needs to have specific terminology assigned to it.
- ▶ **Direct entry** - a concept should be entered into a vocabulary using the term that names it, rather than treating that concept as a subdivision of a broader concept.

Application

- ▶ **Specific Entry** - this allow the user to know when to stop searching for an appropriate controlled vocabulary term.
- ▶ **Number of Terms Assigned** - There should not be any limits on the number of terms or descriptors assigned to the concepts.
- ▶ **Concepts not in CV** - If a concept is not present in the controlled vocabulary, it should be represented temporarily by a more general concept, rather than simply adding unauthorized terms to the record.

Problems with Controlled Vocabulary

- ▶ There are a lot of work
- ▶ There are often difficult and time consuming to maintain
- ▶ Authors have freedoms in choice of terms

Natural Language & Vocabularies System

- i. Natural language system *gives the power to users to enter their own search terms.*
- ii. The problem is that **not all roads lead to Rome** when you conduct information search.
- iii. **The recall is high**, but generally **the precision is low**.
- iv. The users might have to conduct further searches to eliminate irrelevant information.

Vocabularies System

1. The natural language vocabularies system **works great for subject field** that often **generate new words and terms**.
2. It is **difficult to control these vocabularies** when they are often new and not well known.

Natural Languages and Indexing

- ▶ Derived term system or any information retrieval system without vocabulary control, is referred to as a “Natural - Language” or sometimes, as a “free - text”, system because the system allows the indexer to select the term to be used directly from the text being indexed.

EXAMPLE: The uniterm systems developed in the early days of information retrieval are the example of natural language system in which index terms were extracted from documents by human indexers with the application of computers.

Importance of Natural Languages

According to **F.W. Lancaster** “the future will see increased emphasis on the use of natural language in information retrieval.

1. The community growth in the availability of machine readable data bases
2. The continuity expansion of on line systems
3. A number of evaluation studies have indicate the natural language offers several advantage over controlled vocabulary
4. Natural language systems have been shown to work, and work well.
5. New development in computer storage devices will make the storage of very large text file increasing feasible

What is Authority Control & Authorities?

- ▶ In library science, **authority control** is a process that organizes bibliographic information, for example in library catalogs by using a single, distinct spelling of a name (heading) or a numeric identifier for each topic.

https://en.wikipedia.org/wiki/Authority_control

Authorities

- Authority **Control** governs usage of a controlled vocabulary. This is managed with>
- Authority **Files**, that consist of>
- Authority **Records**, each of which records a term and its variants as well as evidence. They are created using>
- Authority **Work**, bibliographic detective work usually.

Authority Control

- ▶ Maintains consistency of usage of names of individuals, corporate bodies, and titles of works.
- ▶ Always:
 - ▶ Smiraglia, Richard P., 1952-
 - ▶ Not Smiraglia, R.P.
 - ▶ Not Smiraglia, Richard
- ▶ Always:
 - ▶ Taylor, Arlene G., 1941-
 - ▶ Not Dowell, Arlene Taylor, 1941-

Authority Records & “CARD”

- ▶ Authority control works through the use of authority records
- ▶ Authority records record:
 - ▶ Authority *work*—the actual decision-making process of the cataloger
 - ▶ Variant forms found along the way
 - ▶ References in the catalog from recognized variant forms

- Originally authority cards were kept in a file in the cataloging department
- Hand-written emendations made over time

Smiraglia, Richard P., 1952-

His Shelflisting music, 1981: t.p. (Richard P. Smiraglia) CIP data (b. 3-18-52)

81031242

AACR2

An Authority Card Contains

- ▶ The authorized “heading”
- ▶ The authorized references
- ▶ Evidence about the forms found in print
- ▶ The cataloging rules used to formulate the heading
- ▶ The date, number, creator, etc. of the authority record

Benefits of Authority Control

- ▶ Makes searching more predictable
- ▶ Better researching
- ▶ Consistency of records
- ▶ Efficiency for Cataloguers
- ▶ Easier to maintain the Catalog
- ▶ Fewer Errors

References

Croft, W.B., Metzler, D., and Strohman, T. (2009). Information retrieval in practice. New Jersey: Pearson Education.

Manning, C.D., Raghavan, P., and Schutze, H. (2008). Introduction to information retrieval. New York: Cambridge University Press.

Grossman, D.A. and Frieder, O. (2004). Information retrieval: algorithms and heuristics. Dordrecht: Springer.

Chowdhury, G.G. (2010). Introduction to modern information retrieval. 3rd edition. London: Facet Publishing.