

Controlled Vocabulary

Definitions

- 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 list of terms or other symbols used in indexing.
- A controlled vocabulary is a type of metadata that functions as a “subset of natural language.
- 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.

Terms not belonging to a controlled vocabulary are called free text terms, natural language terms, and sometimes keywords.

Controlled vocabularies solve the problems of homographs and synonyms between concepts and authorized terms. 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 Vocabulary

Three types of controlled vocabulary are used, they are as follows:

- **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. These are standardized words assigned to a concept. Subject headings are assigned to an article or a book by a person, rather than a computer. Subject headings are also assigned based on the *topic* of the article, rather than just the words that appear in the text.

- **Thesauri**

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

- **Ontology**

An ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really exist in a particular domain. Thus, it is a practical application of philosophical ontology, with a taxonomy.

The fields of artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture all create ontologies to limit complexity and organize information. The ontology can then be applied to problem solving.

Why Controlled Vocabulary?

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.

By principle one is only allowed to use terms from the controlled vocabulary in the indexing process. If a relevant term is missing from the controlled vocabulary, the indexer might suggest that the term is added to the list.

Vocabulary control is used to improve the effectiveness of information storage and retrieval systems, Web navigation systems. The primary purpose of vocabulary control is to achieve consistency in the description of content objects and to facilitate retrieval.

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

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

Example: VHF/Very High Frequency

- II. 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 Vocabulary

Controlled vocabulary serves the following five purposes:

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

How Vocabulary Control is achieved

Vocabulary control is achieved by three principal methods:

- Defining the scope or meaning of terms
- Using the equivalence relationship to link synonymous and nearly synonymous terms
- Distinguishing among homographs.

General Principles for Creating Controlled Vocabulary

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.

Specific Entry – this allows 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

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

Natural Language Vocabularies System

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

Natural Language 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 Language

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

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

- ❖ May be extracted or derived from document text:
- ❖ Natural language system **gives the power to users to enter their own search terms.**
- ❖ The problem is that **not all roads lead to Rome** when you conduct information search.
- ❖ **The recall is high**, but generally **the accuracy is low.**
- ❖ The users might have to conduct further searches to eliminate irrelevant information.

Difference among Natural language, indexing language and controlled vocabulary

- All indexing languages originate as ***natural language, or the language found in documents. Natural language does not refer to writing style, but to the fact that the language is not under authority control.***
- Language under authority control is called ***controlled vocabulary. There is nothing special about the words in controlled vocabulary except the fact that they are standardized for use in certain systems.***
- Natural indexing languages are also called ***derived-term approaches***

Processes For Natural Language

- ☐ Terms are based on existing vocabulary of documents (which may be inconsistent).
- ☐ catalogers extract terms from documents and enter them (or their own terms) in various subject fields
- ☐ Searchers may enter any search terms that are likely to occur in natural language

It is **difficult to control these vocabularies** when they are often new and not well known.

Importance of Natural Language

1. The natural language vocabularies system **works great for subject field** that often **generate new words and terms.**
2. The community growth in the availability of machine readable data bases
3. The continuity expansion of online systems
4. A number of evaluation studies have indicate the natural language offers several advantage over controlled vocabulary
5. New development in computer storage devices will make the storage of very large text file increasing feasible

Boolean operators

Boolean Searching is a database search method based on the principles of Boolean logic, originally developed by the British mathematician George Boole in the mid 19th century. Boolean searching allows you to combine search terms in specific ways for effective matches.

| Operator | Examples | Explanation |
|--|-----------------------------------|--|
| AND Document must have both words (or both phrases) | copyright AND moral rights | Will <u>narrow</u> your search because search results will include all documents that contain both the first term and the second term. |
| OR Document can have either word (or either phrase) | tobacco OR cigarettes | Will <u>expand</u> your search because search results will include all documents that contain either the first term or the second term or both. |
| NOT Document must have first term. Must not have second | love NOT war | Will exclude an idea/concept from your search because it will find documents that contain the first word, but do not contain the second. |
| ADJ Forces the computer to search for words in a specified order | obsessive adj compulsive | Will find results where <i>obsessive</i> immediately precedes <i>compulsive</i> |

| | | |
|---|---|--|
| <p>NEAR</p> <p>Retrieves items that have both terms in the same sentence. You can add a number to near to instruct the computer to find results within those numbers of words in any order</p> | <p>alcohol near abuse</p> <p>alcohol near3 abuse.</p> | <p>This will retrieve results where alcohol is within 3 words of abuse.</p> <p>i.e., "Men who <i>abuse</i> their wives after<i>alcohol</i> consumption" or "<i>Alcohol</i>consumption leads to <i>abuse</i>"</p> |
|---|---|--|

Boolean Strategies

- If you are retrieving too many records on your topic, try adding another search term with the connector AND.
- If you are retrieving too few records on your topic, try adding another search term with the connector OR.
- If you are retrieving too many records on an unrelated topic, try eliminating a word with the connector NOT.

Reference

- 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.