6.3.1. How to use indexing languages in searching

Overview

This module explains how you can become a better searcher by exploiting the power of indexing and indexing languages. Two keys are knowing how fields are indexed and using boolean operators. Although the emphasis here is on subject searching, the same principles apply to all searching.

Knowing how fields are indexed

Successful searching requires an understanding of the indexing language and where it is applied. A brief review of the concepts:

- **Natural language** is the existing language of the document or language freely chosen by the indexer.

- **Controlled vocabulary** is drawn from an authority file and assigned to a specific field by the indexer.

- **Subject-description fields** contain only terms describing the aboutness of a document.

- **Physical-description fields** are all other fields, including fields such as title that may contain clues to aboutness.

To put these concepts together, the searcher who knows that a field contains controlled vocabulary has a better idea of which terms to enter to get an exact match for retrieval.

The fields searched for subject concepts depend on the fields that are appropriate (contain subject terms) and that are available and searchable in the record structure.

Any query term can be entered in any field, but this hardly guarantees a match. The searcher should ask:

- Which field is likely to hold subject information?

- What type of indexing language is used for this field?
Subject searching exercise

Here good is your subject searching skill? Here's a simple exercise to find out. In the tables below are selected record fields from two information retrieval systems. Your task is to decide which fields contain subject data and whether the subject fields contain controlled vocabulary or natural language. The exercise works best if you print out the tables and mark on them.

### Selected fields from MARC record in library catalog

<table>
<thead>
<tr>
<th>MARC fields*</th>
<th>Contains subject terms</th>
<th>Contains controlled vocabulary</th>
<th>Contains natural language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publisher</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Note</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Call no.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note contains any kind of comment; it may contain a summary or abstract. Subject contains Library of Congress Subject Headings. Call number is based on Dewey Decimal or Library of Congress Classification code.

### Selected fields from record in ERIC index to education resources

<table>
<thead>
<tr>
<th>ERIC fields*</th>
<th>Contains subject terms</th>
<th>Contains controlled vocabulary</th>
<th>Contains natural language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accession no.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Journal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abstract</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Descriptors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identifiers</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Accession number contains unique number assigned when document is acquired. Descriptors contains terms from the Thesaurus of ERIC Descriptors. Identifiers contains terms from the document that are not in the thesaurus.

Spend a little time on this—it's not as easy as it looks!
When you are done with the exercise, click here for the answers!

How did you do? Do you agree that understanding indexing principles can improve your search skills?

### Using Boolean operators

Indexing languages provide ways to combine terms to describe documents that have composite or multiple subjects in order to improve retrieval. To review:

- **Postcoordination** is the combination of indexing terms at the time of searching. This is a decision made at the time of creating a controlled vocabulary. Postcoordination is contrasted with precoordination, in which terms are combined at the time of indexing to form subject headings.

- **Boolean query formulation** is the use of boolean operators to combine the terms.

- **Boolean operators** (named for mathematician George Boole, 1815-1864) are mathematical symbols that signify a logical operation to be performed on a set:
  
  **AND**: product; conjunctive; coincidence of terms; decreases output set size

  **OR**: sum; additive; alternative terms; increases output set size

  **NOT**: difference; subtractive; excludes terms; decreases output set size

At first glance, the definitions of boolean operators may seem obvious. Note, however, that the operators AND and OR are opposite what most people expect. For instance, if you ask for an apple and an orange, you want both, right? Conversely, if you ask for an apple or an orange, you want only one. So why do we say the boolean operator AND decreases output, while OR increases output?

The reason is that this is not about real objects, but about representations of the objects—terms—in documents and queries. In the Venn diagram below, AND retrieves only documents containing both terms (fewer documents) . . .

**Example**

Apple AND orange
... while OR retrieves all documents with either or both terms (more documents).

Apple OR orange

How do you remember this? Simple—**OR gets more!**

**Other techniques**

With **Boolean operators**, word order (syntax) may or may not matter, depending on the database. The results are usually the same for:

- gardening AND organic
- organic AND gardening

An alternative to boolean operators is **proximity operators**, often used in natural-language searching of full-text documents. These include NEAR, ADJACENT, FAR, and BEFORE. The exact operators and their abbreviations and syntax depend on the database. The NEAR operator abbreviated "5N" below means that two terms must occur within five words of each other.

- gardening 5N organic
- organic ADJ gardening

Another alternative is **phrase searching**, also used in natural-language searching of full-text documents. The terms are typically bound by quotation marks that signify their exact order.

- "organic gardening"

All of these techniques for combining terms add enormous power to online searching.

Boolean and proximity operators work with controlled vocabulary or natural language. If the terms are part of precoordinated subject headings, the searcher can enter just one or two of the terms. Boolean operators also work across fields, as in a simultaneous search for subject and author.

Phrase searching has become popular on the Web, where almost all searching is natural language in full-text documents. Some search sites now automatically bind a phrase without requiring quotation marks.
Summary

This module brings together the major principles of indexing and searching in practice. The difference between uninformed end users and information professionals lies in their understanding of these principles of information retrieval.

Users may not be aware of:

- the exact-match principle of IR; the fact that no hits may only mean no matches
- the record fields that are appropriate for certain kinds of searches
- the existence of controlled vocabularies and how to use them in searching
- the fact that actual subjects of documents are not searched, but rather representations of subjects
- the fact that representations of subjects often are not predictable, accurate, or consistent

Information professionals should be aware of:

- approaches to searching based on knowing which fields contain which kinds of data, how to exploit controlled vocabulary, how to combine terms, etc.
- the fact that, while the emphasis in indexing languages is on subject description, the vocabulary principles and challenges are the same for physical description

This is why we say at the beginning of the course that understanding information organization will make you a better searcher!