Indexing languages

6.2.1. Indexing concepts

Overview

In order to understand indexing languages and how they are used in subject description, it is first necessary to understand the concept of indexing. The term "indexing" is used two ways:

- In a broad sense, indexing is a **generic concept** that refers to processes of creating and using verbal representations, including both physical and subject description, that take place in a variety of professional settings.

- In a narrow sense, indexing is a major **area of practice** in which specific types of indexes (and often abstracts) are created. As indicated in the module on subject description processes, cataloging is another area of practice in which specific types of cataloging records (and often classification) are created. While the principles of verbal representation are the same in indexing and cataloging, the terminology differs somewhat.

This module introduces concepts for indexing in general and then focuses on subject indexing.

What are indexing languages?

In order to define indexing languages, the first thing to understand, again, is indexing in its broad sense. **Indexing** in general . . .

- Is the process of creating an index for purposes of representing and providing access to information objects

- May be performed by humans or computers

An index . . .

- Is an organized list of pointers or access points to concepts and items in collection, or document*

- May be called index, catalog, or information system

- May be print or electronic

**Document** can refer to any text or nontext information object. Here it refers to target documents, not search queries.
Many types of indexes exist, such as . . .

- Back-of-the-book index
- Periodical index for specific periodical(s)
- Index to the literature of a discipline
- Catalog of materials in collection or library
- Citation index of connections among documents based on authors' citations
- Directory of persons, businesses, organizations, etc.
- Inverted file/index of data values drawn from main file of computer database for the purpose of facilitating matches with search terms for retrieval

An index entry . . .

- Is any pointer or indicator included in an index

An index term . . .

- Is any word/phrase/number used for physical or subject representation in an information system
- Is any word/phrase/number used for searching and retrieving representations
- May describe any attribute of a document

Indexing languages, then (the original question!) . . .

- Are terms or vocabularies used to represent document container or content
- Serve as access points for searching
- Vary from one index or system to the next
- May be extracted or derived from document text: natural language
- May be assigned from authority control list: controlled vocabulary

These are important concepts to know throughout the lessons and assignments on indexing languages.
How are indexing languages created?

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.

This diagram illustrates the processes involved in translating natural language (NL) terms into controlled vocabulary (CV) terms for entry in database records. The diagram helps explain why . . .

- Natural indexing languages are also called **derived-term** approaches
- Controlled indexing languages are also called **assigned-term** approaches

You may notice that these processes encompass the steps of subject analysis described in another module. To review, subject analysis requires you to (1) become familiar with document content; (2) extract significant concepts and terms; (3) translate extracted terms into the language—often controlled—of the system; and (4) formalize the terms (format them, etc.) according to input rules.

Remember, though, that the analysis process is used for physical description well as subject description. Further, individual fields often have their own indexing languages within the same system. For example, name authority control is often used for physical description, as in authors' names.
How does indexing relate to searching?

Indexing and searching are complementary processes:

1. Indexers or catalogers enter terms in database record fields.
2. Searchers enter terms in fields of database search interface.
3. If search terms match terms in records exactly, the system retrieves the records.

Here’s how the processes differ for natural language and controlled vocabulary:

<table>
<thead>
<tr>
<th></th>
<th>Natural language</th>
<th>Controlled vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terms are . . .</td>
<td>based on existing vocabulary of documents (which may be inconsistent)</td>
<td>based on standardized vocabulary intended to describe concepts consistently</td>
</tr>
<tr>
<td>Indexers/catalogers . . .</td>
<td>extract terms from documents and enter them (or their own terms) in various subject fields</td>
<td>extract terms from documents, choose appropriate authorized terms from controlled vocabulary list, and enter terms in designated controlled vocabulary field</td>
</tr>
<tr>
<td>Searchers . . .</td>
<td>may enter any search terms that are likely to occur in natural language</td>
<td>must enter search terms that are in controlled vocabulary</td>
</tr>
</tbody>
</table>

What is subject indexing?

LIS focuses the most attention on languages for subject description because these are the most complex. Basically, subject indexing is the process of creating an index for the purpose of representing and providing access to intellectual content.

A subject term is . . .

- Any term used to describe document content
- Any term used to search for and retrieve document based on its content
- A compact surrogate for subject representation

For subject terms under authority control (or vocabulary control), a subject authority file or list . . .

- Is a list of terms authorized for use in representing subjects
- Standardizes which of two synonyms to assign for a given topic
- Determines the preferred term when more than one term expresses a single topic
- Provides cross references to terms in equivalent, hierarchical, and associative relationships
Cataloging and indexing professionals have created different subject authority control structures:

- Catalogers use a **subject headings list** in which subject terms are called **subject headings**.
- Indexers use a **thesaurus** in which subject terms are called **descriptors**.

Another module explains the differences between these files.

---

**Project Alert!** For IOP section 4.2, all of your **subject fields** are natural language except the one field that contains controlled vocabulary from your thesaurus (required) and any field for which you specify a validation list in Inmagic (optional).