

Chapter IR:III

III. Indexing

- ❑ Indexing Basics
- ❑ Inverted Index
- ❑ Query Processing I
- ❑ Query Processing II
- ❑ Index Construction
- ❑ Index Compression

Indexing Basics

Definition 1 (Index [\[ANSI/NISO 1997\]](#))

An index is a systematic guide designed to indicate topics or features of documents in order to facilitate retrieval of documents or parts of documents.

The function of an index is to provide users with an effective and systematic means for locating documentary units (complete documents or parts of documents) that are relevant to information needs or requests.

Indexing Basics

Definition 1 (Index [\[ANSI/NISO 1997\]](#))

An index is a systematic guide designed to indicate topics or features of documents in order to facilitate retrieval of documents or parts of documents.

The function of an index is to provide users with an effective and systematic means for locating documentary units (complete documents or parts of documents) that are relevant to information needs or requests.

- thesauri
 - definition : 12
- titles of documents : 6.2.9, 6.2.9.4
 - capitalization : 6.2.3
 - initial articles in alphanumeric arrangement : 9.4
- topical headings
 - see also: entries, headings, terms**
 - initial articles in alphanumeric arrangement : 9.4
- topics
 - definition : 12
 - major versus minor topics : 7.3
- transcription
 - definition : 12
- transliteration : 6.2.10
 - definition : 12
- truncation
 - definition : 12
 - in searching : 7.5.3
- turnover lines : 8.2.5.1, 8.2.5.3
 - definition : 12

- vectors
 - definition : 12
 - in searching : 7.5.2
- vertical spacing
 - in indexes : 8.2.4
- video recordings
 - locators : 7.4.2b
- visual indexes
 - see: displayed indexes**
- vocabulary : 6
 - see also: descriptors; terminology of indexing; terms**
 - control, tracking, management : 3h, 5.13; as essential process : *preface*; definition : 12
 - display in displayed indexes : 6.8.1; non-displayed electronic search indexes : 6.8.2
 - entry. definition : 12
 - lead-in : 5.13
 - sources : 3d-e, 6.1

Indexing Basics

Requirements [\[ANSI/NISO 1997\]](#)

1. identify documentary units that treat particular topics or possess particular features.
2. indicate all important topics or features of documentary units in accordance with the level of exhaustivity appropriate for the index.
3. discriminate between major and minor treatments of particular topics or manifestations of particular features.
4. provide access to topics or features using the terminology of prospective users.
5. provide access to topics or features using the terminology of verbal texts being indexed whenever possible.
6. use terminology that is as specific as documentary units warrant and the indexing language permits.

Indexing Basics

Requirements [\[ANSI/NISO 1997\]](#)

1. identify documentary units that treat particular topics or possess particular features.
2. indicate all important topics or features of documentary units in accordance with the level of exhaustivity appropriate for the index.
3. discriminate between major and minor treatments of particular topics or manifestations of particular features.
4. provide access to topics or features using the terminology of prospective users.
5. provide access to topics or features using the terminology of verbal texts being indexed whenever possible.
6. use terminology that is as specific as documentary units warrant and the indexing language permits.

Indexing Basics

Requirements [\[ANSI/NISO 1997\]](#)

1. identify documentary units that treat particular topics or possess particular features.
2. indicate all important topics or features of documentary units in accordance with the level of exhaustivity appropriate for the index.
3. discriminate between major and minor treatments of particular topics or manifestations of particular features.
4. provide access to topics or features using the terminology of prospective users.
5. provide access to topics or features using the terminology of verbal texts being indexed whenever possible.
6. use terminology that is as specific as documentary units warrant and the indexing language permits.

Indexing Basics

Requirements [\[ANSI/NISO 1997\]](#)

1. identify documentary units that treat particular topics or possess particular features.
2. indicate all important topics or features of documentary units in accordance with the level of exhaustivity appropriate for the index.
3. discriminate between major and minor treatments of particular topics or manifestations of particular features.
4. provide access to topics or features using the terminology of prospective users.
5. provide access to topics or features using the terminology of verbal texts being indexed whenever possible.
6. use terminology that is as specific as documentary units warrant and the indexing language permits.

Indexing Basics

Requirements [\[ANSI/NISO 1997\]](#)

1. identify documentary units that treat particular topics or possess particular features.
2. indicate all important topics or features of documentary units in accordance with the level of exhaustivity appropriate for the index.
3. discriminate between major and minor treatments of particular topics or manifestations of particular features.
4. provide access to topics or features using the terminology of prospective users.
5. provide access to topics or features using the terminology of verbal texts being indexed whenever possible.
6. use terminology that is as specific as documentary units warrant and the indexing language permits.

Indexing Basics

Requirements (continued) [\[ANSI/NISO 1997\]](#)

6. provide access through synonymous and equivalent terms.
7. guide users to terms representing related concepts (narrower terms, other related terms, and if possible, broader terms).
8. provide for the combination of terms to facilitate the identification of particular types or aspects of topics or features and to eliminate unwanted types or aspects.
9. provide a means for searching for particular topics or features by means of a systematic arrangement of entries in displayed indexes or, for non-displayed indexes, by means of a clearly documented and displayed method for entering, combining, and modifying terms to create search statements and for reviewing retrieved items.

Indexing Basics

Requirements (continued) [\[ANSI/NISO 1997\]](#)

6. provide access through synonymous and equivalent terms.
7. guide users to terms representing related concepts (narrower terms, other related terms, and if possible, broader terms).
8. provide for the combination of terms to facilitate the identification of particular types or aspects of topics or features and to eliminate unwanted types or aspects.
9. provide a means for searching for particular topics or features by means of a systematic arrangement of entries in displayed indexes or, for non-displayed indexes, by means of a clearly documented and displayed method for entering, combining, and modifying terms to create search statements and for reviewing retrieved items.

Indexing Basics

Requirements (continued) [\[ANSI/NISO 1997\]](#)

6. provide access through synonymous and equivalent terms.
7. guide users to terms representing related concepts (narrower terms, other related terms, and if possible, broader terms).
8. provide for the combination of terms to facilitate the identification of particular types or aspects of topics or features and to eliminate unwanted types or aspects.
9. provide a means for searching for particular topics or features by means of a systematic arrangement of entries in displayed indexes or, for non-displayed indexes, by means of a clearly documented and displayed method for entering, combining, and modifying terms to create search statements and for reviewing retrieved items.

Remarks:

- ❑ Several standards worldwide govern the (manual) construction and formatting of indexes.
[\[Gibbs 2015\]](#)
- ❑ ANSI = American National Standards Institute
- ❑ NISO = National Information Standards Organization

Indexing Basics

Definition 2 (Document, Documentary Unit [\[ANSI/NISO 1997\]](#) (Indexing Unit, Unit of Retrieval))

A document is a medium on or in which a message is encoded; the combination of message and medium. A documentary unit is a document, document segment, or collection of documents to which entries in an index refer.

Indexing Basics

Definition 2 (Document, Documentary Unit [\[ANSI/NISO 1997\]](#) **(Indexing Unit, Unit of Retrieval))**

A document is a medium on or in which a message is encoded; the combination of message and medium. A documentary unit is a document, document segment, or collection of documents to which entries in an index refer.

Relation between documents and files as digital media:

- ❑ One file, one document.

Examples: web page, PDF file, Word file.

- ❑ One file, many documents.

Examples: archive files, email threads and attachments, Sammelbände.

- ❑ Many files, one document.

Examples: paginated web pages, e.g., news, slide decks, forum threads.

Indexing Basics

Definition 2 (Document, Documentary Unit [\[ANSI/NISO 1997\]](#) **(Indexing Unit, Unit of Retrieval))**

A document is a medium on or in which a message is encoded; the combination of message and medium. A documentary unit is a document, document segment, or collection of documents to which entries in an index refer.

Relation between documents and documentary units:

- ❑ One document, one unit.
The default in most large-scale search engines.
- ❑ One document, many units.
Examples: comments, reviews, posts, arguments, chapters, sentences, words, etc.
- ❑ Many documents, one unit.
Examples: corpora and datasets, Sammelbände, libraries.

Indexing Basics

Definition 3 (Vocabulary, Terminology, Term [\[ANSI/NISO 1997\]](#))

A vocabulary is a collection of words and phrases.

The terminology is a collection of index terms used in an index.

An index term, or term, is (a derivation of) a word or phrase selected from a vocabulary, used to represent a topic or feature of a documentary unit in an index.

Indexing Basics

Definition 3 (Vocabulary, Terminology, Term [\[ANSI/NISO 1997\]](#))

A vocabulary is a collection of words and phrases.

The terminology is a collection of index terms used in an index.

An index term, or term, is (a derivation of) a word or phrase selected from a vocabulary, used to represent a topic or feature of a documentary unit in an index.

Sources of vocabulary:

- ❑ Documents to be indexed.
Applies only to text documents.
- ❑ Users of the index.
Typically available only after the index is in use.
- ❑ Experts and human indexers.
Domain experts or librarians. Typically not applied for search engines.
- ❑ Compilations of vocabularies.
Dictionaries, thesauri, etc. Possibly applied in a search engines' backend.

Indexing Basics

Definition 3 (Vocabulary, Terminology, Term [\[ANSI/NISO 1997\]](#))

A vocabulary is a collection of words and phrases.

The terminology is a collection of index terms used in an index.

An index term, or term, is (a derivation of) a word or phrase selected from a vocabulary, used to represent a topic or feature of a documentary unit in an index.

The terminology of an index **should match** the vocabulary of users.

Aspects of terms:

- ☐ Parts of speech
- ☐ Spelling
- ☐ Capitalization
- ☐ Singular and plural forms
- ☐ Articles
- ☐ Compound terms
- ☐ Antonyms and associated terms
- ☐ Names
- ☐ Homographs
- ☐ Synonyms and equivalent terms
- ☐ Hierarchical relationships among terms
- ☐ Other relationships
- ☐ Changes over time
- ☐ **Weighted terms**

Indexing Basics

Definition 3 (Vocabulary, Terminology, Term [\[ANSI/NISO 1997\]](#))

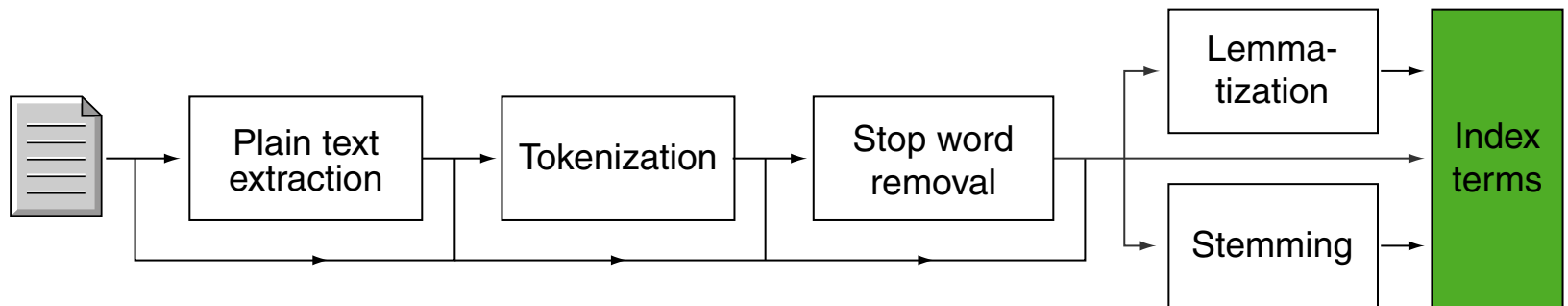
A vocabulary is a collection of words and phrases.

The terminology is a collection of index terms used in an index.

An index term, or term, is (a derivation of) a word or phrase selected from a vocabulary, used to represent a topic or feature of a documentary unit in an index.

The terminology of an index **should match** the vocabulary of users.

Automatic derivation of terminology through natural language processing:



Excursus: Tokenization, Stopping, Stemming, Lemmatization.

Remarks:

- ❑ The terms “vocabulary” and “terminology” are often used interchangeably.

Indexing Basics

Querying an Index

Queries are users' formulations of information needs in a search engine's language:

- ❑ Keyword queries
- ❑ Question queries
- ❑ Query by example