Chapter IR:III

III. Indexing

- Indexing Basics
- Inverted Index
- Query Processing I
- Query Processing II
- Index Construction
- Index Compression
- Size Estimation

Indexing Basics

Definition 1 (Index [ANSI/NISO 1997]**)**

An index is a systematic guide designed to indicate topics or features of documentary units as index terms in order to facilitate their retrieval.

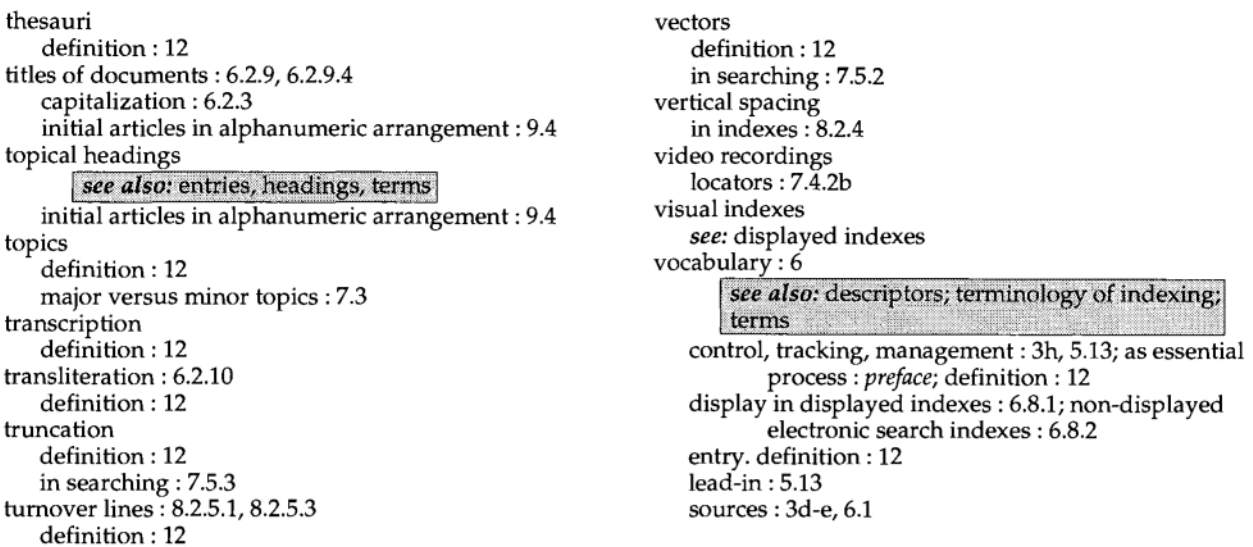
The function of an index is to provide users with an effective means for locating documentary units relevant to their information needs in answer to queries.

Indexing Basics

Definition 1 (Index [ANSI/NISO 1997])

An index is a systematic guide designed to indicate topics or features of documentary units as index terms in order to facilitate their retrieval.

The function of an index is to provide users with an effective means for locating documentary units relevant to their information needs in answer to queries.



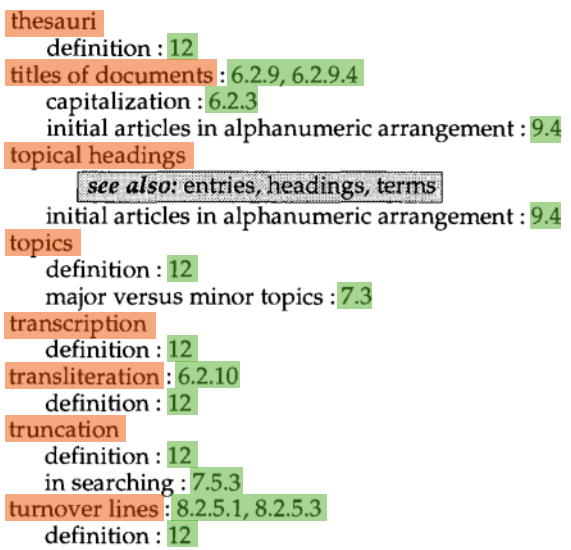
thesauri
  definition : 12
titles of documents : 6.2.9, 6.2.9.4
  capitalization : 6.2.3
  initial articles in alphanumeric arrangement : 9.4
topical headings
  see also: entries, headings, terms
  initial articles in alphanumeric arrangement : 9.4
topics
  definition : 12
  major versus minor topics : 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

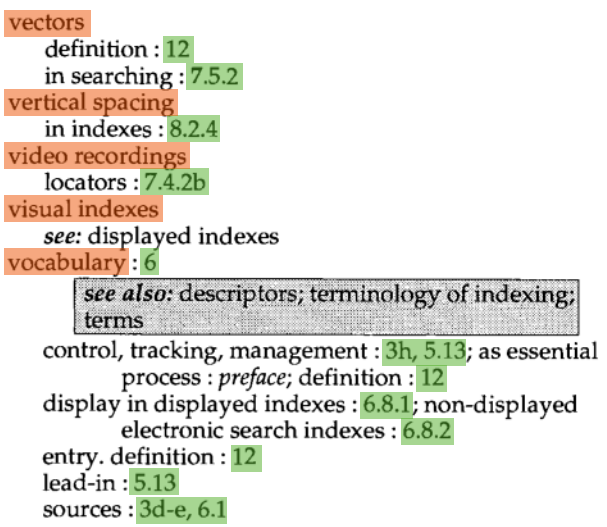
Definition 1 (Index [ANSI/NISO 1997])

An index is a systematic guide designed to indicate topics or features of documentary units as index terms in order to facilitate their retrieval.

The function of an index is to provide users with an effective means for locating documentary units relevant to their information needs in answer to queries.



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. Identification of documentary units that treat particular topics or possess particular features.

2. Indication of all important topics or features of documentary units in accordance with the level of exhaustivity appropriate for the index.

3. Discrimination between major and minor treatments of particular topics or manifestations of particular features.

4. Provision of access to topics or features using the vocabulary of prospective users.

5. Provision of access to topics or features using the vocabulary of verbal texts being indexed whenever possible.

6. Use of terminology that is as specific as documentary units warrant and the query language of users permits.

Indexing Basics

Requirements [ANSI/NISO 1997]

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

Indexing Basics

Requirements [ANSI/NISO 1997]

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

Indexing Basics

Requirements [ANSI/NISO 1997]

1. Identification of documentary units that treat particular topics or possess particular features.

2. Indication of all important topics or features of documentary units in accordance with the level of exhaustivity appropriate for the index.

3. Discrimination between major and minor treatments of particular topics or manifestations of particular features.

4. Provision of access to topics or features using the vocabulary of prospective users.

5. Provision of access to topics or features using the vocabulary of verbal texts being indexed whenever possible.

6. Use of terminology that is as specific as documentary units warrant and the query language of users permits.

Indexing Basics

Requirements (continued) [ANSI/NISO 1997]

7. Provision of access through synonymous and equivalent terms.

8. Guidance of users to terms representing related concepts (narrower terms, other related terms, and if possible, broader terms).

9. Provision 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.

10. Provision of 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 queries and for reviewing retrieved documentary units.

Indexing Basics

Requirements (continued) [ANSI/NISO 1997]

7. Provision of access through synonymous and equivalent terms.

8. Guidance of users to terms representing related concepts (narrower terms, other related terms, and if possible, broader terms).

9. Provision 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.

10. Provision of 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 queries and for reviewing retrieved documentary units.

Remarks:

- Several standards worldwide govern the (manual) construction 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.

For simplicity, henceforth, when we speak of a “document” a “documentary unit” is implied.

Indexing Basics

Definition 3 (Vocabulary, Terminology, Index 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 document in an index.

Indexing Basics

Definition 3 (Vocabulary, Terminology, Index 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 document 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

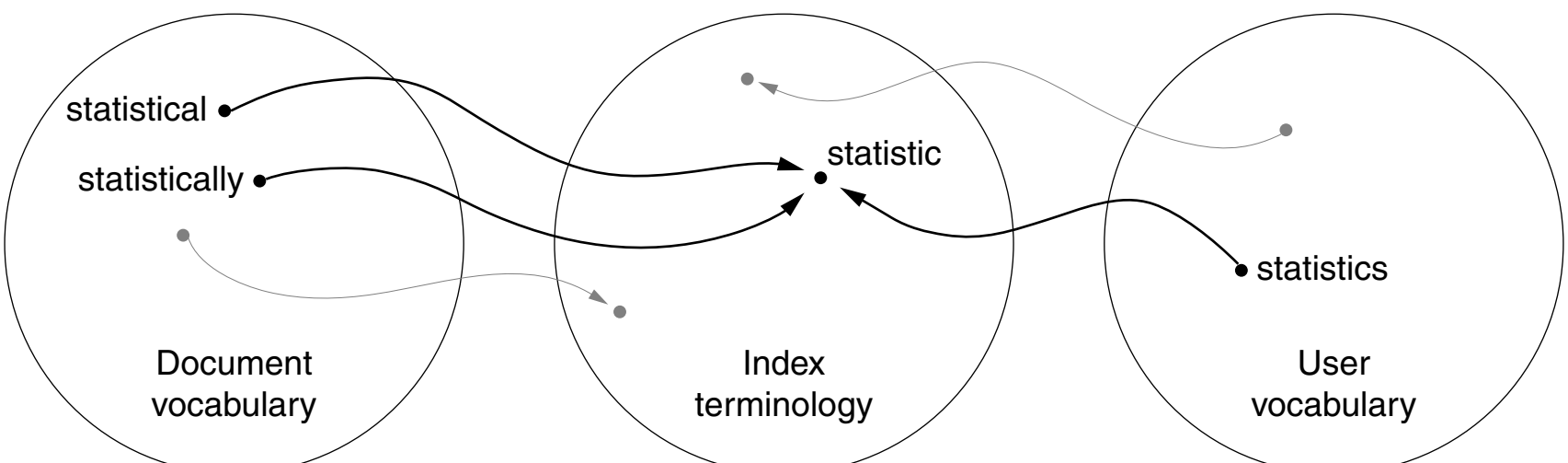
Definition 3 (Vocabulary, Terminology, Index Term [ANSI/NISO 1997]**)**

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 document in an index.

The vocabularies of documents and users are mapped to a common terminology:



Indexing Basics

Definition 3 (Vocabulary, Terminology, Index 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 document in an index.

Linguistic phenomena of words:

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

Indexing Basics

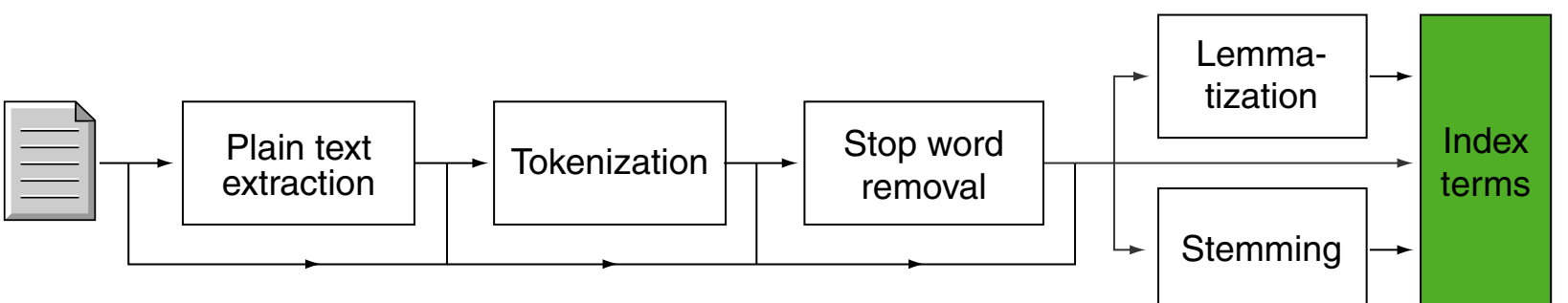
Definition 3 (Vocabulary, Terminology, Index Term [ANSI/NISO 1997]**)**

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 document in an index.

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