Subject Analysis

The process of ascertaining the “aboutness” of a document by describing its topic, the discipline in which the topic is treated, and the form of the document.

Functions of Bibliographic Tools

• Finding (find a work of which subject is known)
• Collocating (find what repository has on subject)
• Evaluating (assist in making informed decision)

Three Stages

1. Abstracting, Subject analysis, or the analysis of the subject content
2. Indexing, Subject cataloging, or the assignment of verbal subject terms
3. Classification, or the assignment of notational symbols to represent the subject content

Elements of Subject Analysis

1. DISCIPLINE
An area or a branch of knowledge. The discipline is distinct from the thing being studied by the discipline. A broad field of inquiry; the context in which any subject is treated. For example,
A medical dictionary of diseases
discipline = Medicine

Technical Processes in Bibliographic Control

1. Description
2. Name access
3. Subject analysis
4. Record formatting
5. Record organization
Elements of Subject Analysis

2. SUBJECT CONCEPTS, or ABOUTNESS TOPICS (PHENOMENA)
Broadly, the things studied by disciplines

The psychology of the adolescent adolescent is the phenomenon (topic) studied by the discipline psychology.

Elements of Subject Analysis

3. FORM, or what the document is:
   a. Physical form (book, pamphlet)
   b. Presentation (symbol (language), arrangement (anthology), audience)
   c. Intellectual form (history, encyclopedia, etc.)

Abstracting: Kinds of Abstracts

- Indicative (aboutness): restricts itself to indicating the content of an item and makes general statements about the document. Avoids statements such as “is discussed” or “has been surveyed”, but do not report the outcome.
- Informative (specific data): present as much of the quantitative or qualitative information contained in the document as possible. Objective is to offer a substitute for the document.
- Critical: evaluates a document and the work that it records. Includes abstractor’s opinions of the merit of the article. Critical abstracts are rare.
- Slanted: similar to critical abstracts. Written from a point of view of a specific group.

Informative Abstract Contains

- Objectives and scope (what was the author’s purpose)
- Methodology
- Results or outcomes
- Conclusions

Informative Abstract Does Not Contain

- Historical background
- Introduction
- Redundancy
- Old information
- Details of standard procedures
- Information the reader is expected to know
- Ideas for future research
- Raw data
- Abstractor’s prejudices

Example

Dykstra, Mary

LC Subject Headings Disguised as a Thesaurus

Examples: Indicative Abstract

- Discusses recent changes in the Library of Congress Subject Headings (LCSH), arguing that the adoption of the well-recognized codes for thesauri has created confusion because the LCSH is not a true thesaurus. The distinction between subject headings and terms are clarified and a possible solution to the problems with the LCSH is suggested.

Examples: Informative Abstract

- While outdated or inappropriate terminology in Library of Congress Subject Headings (LCSH) has been partially to blame for users having given up trying to use manual catalogs, it is clear that the far deeper problems lie in the way the headings themselves are constructed and related to each other. The recent developments made in adopting clear codes, used in a standard thesaurus, is not a step towards improvement. The LCSH list is not a thesaurus and any attempt to make it into one simply by changing the codes on the existing headings in the present structure is misleading and impossible to achieve. It is important to grasp the fundamental distinction between subject heading and term, as this distinction forms the basis of where LCSH and thesauri part company with each other.

Examples: Critical Abstract

- Discusses recent changes in the Library of Congress Subject Headings (LCSH), arguing that the adopting of the well-recognized codes for thesauri has created confusion since the LCSH is not a true thesaurus. The distinctions between subject headings and terms are clarified and possible solutions to the problems with the new LCSH are suggested. This article generated controversy, with a reply from LC that they never intended the new edition of LCSH to be a true thesaurus. The debate still continues, as LCSH looks like a thesaurus, but does not meet thesaurus standards.

Examples: Slanted Abstract

- Catalogers will be especially intrigued by this debate on whether or not the new edition of Library of Congress Subject Headings (LCSH) is a true thesaurus. The author argues that the way the LCSH headings are constructed and relate to each other point to deep problems in the list. The recent developments made in adopting clear codes, used in a standard thesaurus, is not a step towards improvement. The LCSH list is not a thesaurus and any attempt to make it into one simply by changing the codes on the existing headings in the present structure is misleading and impossible to achieve. It is important to grasp the fundamental distinction between subject headings and term, as this distinction forms the basis of where LCSH and the thesauri part company with each other. Indexers and catalogers do the same intellectual activities, but their tools differ in fundamental and structural ways.

Examples: Exercising objectivity

The big lie: the Pentagon plane crash that never happened / Thierry Meyssan.

Dawn; the herald of a new and better day.

The silent subject: reflections on the unborn in American culture / edited by Brad Stetson.

Examples: Exercising objectivity

- Mein Kampf / Adolf Hitler
- The Holocaust story and the lies of Ulysses / Paul Rassinier.
- Slander: liberal lies about the American right / Ann H. Coulter.
**Vocabulary control**

Recall—response to query

Relevance—materials in system relevant to query

Precision—proportion of materials retrieved that are relevant to query.

Recall = \( \frac{\text{number of relevant documents retrieved}}{\text{total number of relevant documents in the file}} \)

Precision = \( \frac{\text{number of relevant documents retrieved}}{\text{total number of documents retrieved in the file}} \)

Inverse relation between recall and precision.

**RECALL AND PRECISION EXAMPLE**

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of relevant items in the system</td>
<td>50</td>
</tr>
<tr>
<td>Total of retrieved items</td>
<td>20</td>
</tr>
<tr>
<td>Total of relevant items retrieved</td>
<td>12</td>
</tr>
<tr>
<td>Total of irrelevant items retrieved</td>
<td>8</td>
</tr>
<tr>
<td>Recall</td>
<td>12/50 = .24</td>
</tr>
<tr>
<td>Precision</td>
<td>12/20 = .60</td>
</tr>
</tbody>
</table>

**IMPROVING RECALL AND PRECISION**

- Improving recall
  - to get more relevant items
  - to relax search terms and to make query broader

- Improving precision
  - to get less irrelevant items
  - to narrow search terms and make query narrower

**Basic Principles (Shera and Egan “objectives for any form of subject cataloging”)**

1. To provide access by subject to all relevant material.
2. To provide subject access to materials through all suitable principles of subject organization, e.g., matter, process, applications, etc.
3. To bring together references to materials [that] treat substantially the same subject regardless of disparities in terminology.
   Trucks vs. Lorries
4. To show affiliations among subject fields.

**What is a Controlled Vocabulary?**

- From Wikipedia: A controlled vocabulary is a carefully selected list of words and phrases … The terms are chosen and organized by trained professionals (including librarians and information scientists) who possess expertise in the subject area. Controlled vocabulary terms can accurately describe what a given document is actually about, even if the terms themselves do not occur within the document’s text.
Controlled Vocabularies: Subject Heading lists vs. Thesauri

- Thesauri
  - Created largely in indexing communities
  - Made up of single terms and bound terms representing single concepts (usually called descriptors). Bound terms occur when some concepts can only be represented by two or more words (e.g., Type A Personality)

- Subject heading lists
  - Created largely in library communities
  - Consist of phrases and other precoordinated terms in addition to single terms

Controlled Vocabularies: Subject Heading lists vs. Thesauri

- Thesauri
  - More strictly hierarchical. Because they are made up of single terms, each term usually has only one broader term
  - Narrow in scope. Usually made up of terms from one specific subject area
  - More likely to be multilingual. Because single terms used, easier to maintain in multiple languages

- Subject heading lists
  - Not strictly hierarchical. Some headings may have no broader and/or narrower terms
  - More general in scope, covering a broad subject area, or the entire scope of knowledge
  - Usually not multilingual

Translating key words & concepts into controlled vocabulary

- Controlled vocabulary
  - Thesauri (examples)
    - Art & Architecture Thesaurus (AAT)
    - Thesaurus for Graphic Materials I: Subject Terms (TGMI)
    - Thesaurus for Graphic Materials II: Genre and Physical Characteristic Terms (TGMII)
    - Thesaurus of Geographic Names (TGN)
  - Subject heading lists (examples)
    - Library of Congress Subject Headings (LCSH)
    - Sears List of Subject Headings
    - Medical Subject Headings (MeSH)

Subject Headings / Thesauri

- Warrant:
  - Literary warrant: The subjects that have appeared in the literature or are present in a library collection used as a basis for determining the subjects to include in a classification scheme or an indexing language.
  - User warrant: Indexing language or classification scheme tailored to the needs of the user of the index or classification system.

- Uncontrolled vocabulary: natural language
  - Free-text
  - Keyword

Subject Headings / Thesauri

- Controlled vocabulary: either derived from or on dependent on natural language. Controlled vocabulary are intended to be more logical whether or not they are simpler than natural (uncontrolled) vocabulary
  - Index terms
  - Descriptors
  - Subject headings
  - Controlled keywords

Keywords vs. Controlled Terms

- System should allow for both
- Keywords give access using “non-standard” terms
- Keywords include terms not yet in vocabularies; places or names not indexed
Drawbacks to Controlled Vocabulary

- Time to assign = $
- Need for trained catalogers = $
- Time lag to add relevant terms
- Time lag to delete outdated terms
  – … so use both keywords and controlled terms

Examples:

- keyword searching challenges
  
  Above all, don’t flush! : adventures in valorous living.
  
  Let’s rejoin the human race!
  
  Dawn; the herald of a new and better day.
  
  Phantom limb.
  
  Who moved my cheese?

Subject Headings / Thesauri

- Choosing one among many
- Standardizing the form of terms
  – Noun vs. verb
  – Singular vs. plural
  – Term manipulation

Why Vocabulary Control?

- Systems of controlled vocabularies coordinate (i.e., map or connect) the various words that describe concepts.
- These concepts are maintained in a thesaurus or headings list, where each concept is located in what we’ll call a “semantic neighborhood.”
- Problems addressed by controlled vocabularies:
  – Lexical and orthographic variation
  – Word sequences
  – Abbreviations
  – Technical language
  – Synonymy
  – Homonymy

What Vocabulary Control Controls

- Problem of natural language variation
  – Variation over time in the way concepts are “labeled” (i.e., referred to)
  – Geographical variation in the way people ask for information packages when using a conceptual (i.e., subject) criterion – The Great Pop vs. Soda Controversy
- Problem centers on the fact that the same concept (e.g., fizzy water) can be represented with different strings of characters (i.e., words).
- Cataloger’s decision concerning subjects: “I know what the subject is (from an analysis of the information package in hand), but what subject data do I enter into the surrogate record?”

Lexical and Orthographic Variation

- Same concept, different spellings:
  – Varying word forms
  – Differences across cultures and countries
  – Changes over time
- Examples:
  – Varying word forms:
    – Clothes ➔ clothing
  – Differences across cultures and countries:
    – Orthopedics ➔ orthopaedics
    – Catalog ➔ catalogue
  – Changes over time:
    – On line ➔ on-line ➔ online
Word Sequences

- Same multiword concept, different word order:
  - *Inversion* is the process of reversing word order
  - Used to collocated like with like
- Examples:
  - Educational psychology ➔ psychology, educational
  - Right of asylum ➔ asylum, right of

Singular vs. Plural

- Prefer plural, but make certain singular and plural mean same thing
  - Apple, Apples
  - Art, Arts
  - Sonata, Sonatas

Abbreviations

- Same concept, different word forms:
  - Depends on the intended use of the controlled vocabulary
  - Abbreviations are normally spelled out
  - Exceptions for those concepts with worldwide recognition
- Examples:
  - Acquired Immunodeficiency Syndrome ➔ AIDS

Technical Languages

- Same concept, different words:
  - Depends on the intended use of the controlled vocabulary
  - Use words that matches that of the intended user group
- Examples:
  - Neoplasms (MeSH) ➔ cancer (LCSH)

Synonymy

- Same concept, different words.
- Examples:
  - Cats ➔ felines
  - Attire ➔ dress ➔ clothing

Homonymy

- Same word and/or pronunciation, different concepts
  - Homographs – same word, different concepts
  - Homophones – same pronunciation, different concepts
- Examples:
  - Homographs:
    - Mercury ➔ god
    - Mercury ➔ planet
  - Homophones:
    - Foul ➔ fowl
Vocabulary Control Techniques

- Use of entry terms:
  - Lexical and orthographic variation
  - Word sequences
  - Abbreviations
  - Technical language
  - Synonymy

- Use of various disambiguation techniques:
  - Homonymy

Vocabulary Control – Entry Terms

- All words or word forms are mapped to the authorized word form for that concept.
- Quality of controlled vocabulary is directly related to the number of entry terms, which themselves must be maintained over time.

Vocabulary Control – Disambiguation

- Qualification:
  - Parenthetical qualifier to disambiguate meaning
    - Mercury (God)
    - Mercury (Planet)
  - Scope notes
- Domain specification:
  - Limiting controlled vocabulary to a subject or domain
    - Mercury (God) not needed in medical vocabulary because it is out of scope
- Hierarchy
  - Embed concept in hierarchical context
    - Metals ➔ Mercury

The “Semantic Neighborhood” Concept

- The concepts represented in controlled vocabularies can be thought of as residing in a semantic neighborhood.
- Each concept has relationships to other concepts in four directions:
  - “North/south” for broader and narrower concepts
  - “East/west” for related concepts
- These relationships provide the intellectual infrastructure of a domain to the indexer and the searcher (“riding the rails”).

Broader and Narrower Concepts

- Hierarchical relationships – “riding the rails” in northerly and southerly directions:
  - In classical terms: parent/child relationships – mammal ➔ dog
  - In ontological terms: “is/a relationships” – dog is a mammal
- Broader term (BT) relationships (northerly):
  - Also termed “superordinate” relationship
  - Gives the context of the concept for disambiguation purposes
- Narrower term (NT) relationships (southerly):
  - Also termed “subordinate” relationship
  - Indexers ALWAYS use the most specific term (specific entry)

Related Concepts – East/West

- Equivalence and associative relationships – “riding the rails” in east/west directions.
- Equivalence relationships – “use” and “use for” (UF):
  - Relates authorized terms and entry terms (i.e., synonyms)
    - Maintenance
    - UF: Upkeep
    - Upkeep: use Maintenance
- Associative relationships – “related terms” (RT):
  - Also called “see also” terms
  - Terms have non-hierarchical relationship (i.e., not synonymous)
    - Birds
    - RT: Ornithology
Basic Concepts of Subject Analysis

• The reader is the focus – how do users ask for information packages when using retrieval tools?
• Collocation using subject criterion.
• Terms must represent common usage among users of the retrieval tool.
• Specific Entry – professional indexers ALWAYS index to the most specific relevant term.

Performing Subject Analysis – Two Steps

– 1. Read for subject analysis to generate concepts that describe the “aboutness” of the information package:
  • DO NOT read entire work
  • DO read:
    – Preface
    – Introduction
    – Table of contents
    – Some text, especially noting those words that are bold, italicized, part of a caption, etc.
    – Bibliographies
– 2. Translate concepts into the controlled vocabulary of the information retrieval tool.

Indexing Exhaustivity

• Addresses the question: How many subject terms per information package?
  – Important to indexers because of the need to maintain consistency across information packages indexed in a retrieval tool
  – Important for searchers to know in advance of a successful search
• Continuum: Summarization ➔ ➔ ➔ Depth Indexing
• Summarization:
  – Purpose is to summarize subject content of information package
  – Used in library cataloging (books contain secondary indexing structures)
• Depth Indexing:
  – Purpose is to exhaustively describe subject content of information package
  – Used in article indexing (and web page indexing) for information packages that often do not have secondary indexing structures

Example

**Soft drinks** (May Subd Geog)
UF Drinks, Soft
Flavored carbonated beverages
Pop (Beverage)
Soda pop
Sodas Beverages)
BT Beverages
NT Cola drinks
Irn-Bru (Trademark)

Subject Analysis in Context

• Obtain information package.
• Describe information package in surrogate record.
• Subject analyze information package in surrogate record:
  • Verbal
  • Classification

Reading for Subject Analysis Purposes

• Take notes while reading.
• Write down common themes.
• Use dictionaries, encyclopedias, etc. to clarify terminologies.
• Get a clear picture in mind about primary and secondary topics (the “aboutness”) of the work.
Subject Analysis: Effects of 3 x 5 Card

- Economize on number of assigned headings
- References not always kept up
- Difficult to update terms

Problems With Subject Heading Schemes

- Outdated terms
- Inconsistent, unpredictable format
- Too few per record

Subject Analysis: Effects of Online Environment

- Keyword searching
  - Best for distinctive or new topics
  - May use uncontrolled natural language
  - Large number of false hits
  - Alternate spellings, words with several meanings, synonyms
- Boolean operators (and/or/not)
- Automatic switching (synonyms)
- Rotation of elements of headings
- Limiting function: Refine original search
  - By date/language of publication/author/type of publication
- Subject browsing
  - Browse through the subject scheme itself and choose terms to search in the library's catalog; very few catalogs do this yet