Representing Information Content

Objectives of this lecture
- Define concepts and terms related to subject representation
- Examine approaches and processes for subject representation
- Link lecture’s concepts and terms to IOP Section 4

What is a subject?
- When a user asks the following...
  - Do you have anything on the subject of [...]?
  - What is this book about?
- They are using a commonsense understanding of the concept of subject
- Subject: The intellectual content of an information object; can include reference to its:
  -- Aboutness -- Expressed concepts or ideas
  -- Themes -- Time periods covered
  -- Topics -- Geographic areas covered
  -- Genres -- Areas of interest or knowledge

Traditional view of subject
- Based on library and bibliographic conventions for representing textual objects through use of words and/or codes
- Distinguishes between what an object is about (intellectual content) and what an object is (physical container)
- Assumes an object has identifiable content that can be represented in words and/or code

Review of library practices
- Descriptive Cataloging
  - Represents information object as a physical entity, based on intrinsic attributes (e.g., book title)
- Subject Analysis/Cataloging
  - Represents intellectual content of information object through use of words and phrases (controlled vocabularies, natural language)
- Classification
  - Represents intellectual content of information object within a larger system of objects through use of code and notation; used to logically and/or physically group related items

Representing intellectual content
- We now move from
  - Bibliographic access (container oriented)
  - Intellectual access (content oriented)
- Distinguish between what an object is (physical container) from what an object is about (intellectual content)
- Goal: Represent adequately the intellectual content of an object for the searcher to find and select an object
- Two main approaches:
  -- Verbal subject representation: using words or phrases to represent concepts, ideas, topics
  -- Classification: represent the overall contents of an object within a larger system of objects using a notation system
Subject representations

- Subject representations are...
  - Secondary information objects that describe the intellectual content (aboutness) of primary information objects (e.g., documents and queries)
  - Words, phrases, sentences and codes that serve as surrogates for information object’s content
- Foundation of representation is **subject analysis**: determining what the object is about

How does one represent subjects

- Traditional LIS processes for representing content are:
  - Subject cataloging
  - Classification
  - Indexing
  - Abstracting
- These processes vary in:
  - Their purpose
  - Entity level described
  - Form of the representation created
  - Type of language or code used
  - Source of the language or code
- The form of representations vary in:
  - Number of terms and symbols used
  - Precision and specificity of terms and symbols

Functions of subject representation

- Inform searchers about intellectual content of documents
- Enables subject access (for searching)
- Distinguish subject representation from subject access
- When controlled vocabulary is used, increases consistency of subject representation and thus improve retrieval results

Challenges to subject representation

- Subjects are often difficult or impossible to identify due to...
  - Subjective interpretation of intellectual content
  - Domain expertise of person doing subject representation
  - Materials with meanings that are not based on language (e.g., images, music)
- “How can nonbook materials, such as visual and musical works, be subject-indexed using the medium of language?” (Svenonius, 1990)

Subject analysis: The first step

- Goal: Determine intellectual content of an object
- Process: Analysis of the object
  - By information professional (cataloger, indexer)
  - By computer (automatic indexing)
- Objectives:
  - Clarify and organize subjects of documents
  - Express subjects precisely
  - Achieve consistency between document content and likely search terms
Steps in subject analysis

1. Familiarization: Get acquainted with the general content of the object
2. Extraction: Identify and select significant concepts, topics, etc. to represent
3. Assignment of terms (see following slides)
   - The challenge is to adequately represent the content so future users can find, identify, and select a needed object!

Verbal Subject Representation

- Uses one or more words or phrases to represent the concepts, topics, etc., identified during subject analysis
- Can use **controlled vocabulary**, and if so:
  - Assign a term from vocabulary to represent the concept, topic, etc.
  - Formally enter the term into the record paying attention to rules for format, spelling, punctuation
- Can use **natural language**
  - Use the term, as found in the object, for naming the concept

Indexing languages

- Any set of terms that are available for use in representing concepts, topics, etc.
- Two primary types:
  - **Controlled** indexing languages or assigned-term systems (see discussion in previous lecture)
  - **Natural** indexing languages or derived-term systems
- Oriented toward either
  - Pre-coordination (at time of indexing)
  - Post-coordination (at time of searching)

Subject Access

```
Subject Analysis
Familiarization
Extraction
If no authority control
Assign natural language term
If authority control
Assign controlled vocabulary term (use thesaurus)

Subject Representation
```

Specificity: Measure of indexing language

- How specific are the terms available in the indexing language? How general?
- Are they specific and precise enough to adequately represent the identified concepts, topics, etc.

**Topic to represent:** Web Search Engines

**Available terms:**
- Internet
- Web Information Retrieval
- World Wide Web
- Search Engines

Which terms are more specific?

Depth of indexing: How many terms?

- The number of terms assigned (depth of indexing) to an object is a measure of indexing practice
- If a document contains many concepts, topics, etc.:
  - How many should be represented?
  - Should specific terms or broader terms be used?
- Depth of indexing: To what extent is the content of an object represented?
  - Summarization level: summarize content in few terms (typical of library cataloging)
  - Exhaustive level: use many terms to represent content (typical of indexing services)
- Across a collection in an information organization system, consistent depth of indexing is desirable.
Representing Information Content

The semantic challenge

- With controlled vocabulary . . .
  - The record creator (indexer, cataloger) must enter only authorized terms in database records
  - The searcher must enter terms that match those in the record in order to retrieve the record

- The problem:
  - How do record creator and searcher come up with the same terms?
  - How does controlled vocabulary deal with ambiguity and variability in language?

Solution: identify semantic relationships

- Equivalent
  - Same or similar meaning
  - Directs one to authorized term (Dog not Pooch)

- Hierarchical
  - Broader and narrower meanings
  - Lets one decide which is appropriate (Breed vs. Poodle)

- Associative
  - Similar meanings but not interchangeable
  - Lets one discover related sources (Showing and Training)

Indexing language, indexing, searching

- Particularly with controlled vocabularies, information organizers and searchers share the indexing language
- Organizers assign terms from the language to represent selected concepts, topics, etc.
- Searchers choose terms from the language to represent the concepts, topics, etc., for which they are searching

What is classification

- Arrangement of entities or concepts in logical groups according to their similarities
- Classification is used in many disciplines (e.g., biology, botany, etc.)
- Bibliographic classification
  - Used in libraries or information centers
  - Many bibliographic classifications systems exist
  - Most, if not all, based on the notion of subject
  - Libraries base physical organization on the classification system used

Bibliographic Classification

- Based on subject analysis; provides intellectual access
- Represents the object in a larger system of objects via a number/code
- Classification can be used for physical arrangement, but this is a secondary use
- Classification as representation:
  - Uses a controlled vocabulary in the form of a classification scheme
  - Form of the representation is a number/code from the scheme
  - Assigns a number/code to object
  - Number/code attempts to represent overall content (summarization)

Classifying an object

- To classify means to assign an object to a particular category or group (i.e., class)
- Assignment is based on how similar an object is to others in the group
- We represent the assignment by applying a notation or code to the object to identify it as belonging to a specific class
- When using the classification system as basis for physical organization, we extend the class number/code with a unique identifier
Representing Information Content

Notation

- Classification number (or code): notation to represent a subject
- Notation: set of codes representing classes and subclasses in classification system
- Notation styles:
  - Pure: Only numerals or only letters
  - Mixed: Letters plus numerals; alphanumeric
  - Both may include punctuation: decimals, commas, slashes, spaces, etc.
- Object is assigned to a class
- Notation labels the object

Classification vs. call number

- Classification number (or code): assigned to an object to represent what it is about
- Call number: Classification number plus unique code/number → creates unique shelf address

Major approaches: Hierarchical

- Subjects and their subordinate subjects are prearranged in classes and subclasses in a formal classification system
- Usually enumerative, attempting to include every possible concept in the system
- Notation for each class is predetermined
- Examples:
  - Dewey Decimal Classification (DDC): truly hierarchical and somewhat enumerative
  - Library of Congress Classification System (LCC): strongly enumerative and somewhat hierarchical

Hierarchy

- Trees
  - Maples
    - Red maples
    - Silver maples
  - Oaks
    - Live oaks
    - Pin oaks

Hierarchical notation

1 Trees
   1.1 Maples
     1.1.1 Red maples
     1.1.2 Silver maples
   1.2 Oaks
     1.2.1 Live oaks
     1.2.2 Pin oaks
2 Bushes

Major approaches: Faceted

- Facets: General attributes or characteristics; types of classes. May include:
  - subject description, such as topic, geographic coverage, language, literary form
  - physical description, such as format, date
- Facets will likely be predetermined
- Foci or classes: Concepts within facets possessing common attributes
- Foci or classes may or may not be prearranged
- Notation is synthesized by drawing together notation from different facets
Representing Information Content

### Facets, foci/classes, and notation

<table>
<thead>
<tr>
<th>Foci</th>
<th>Language</th>
<th>Form</th>
<th>Period</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>American</td>
<td>Poetry</td>
<td>16th C</td>
<td>Nature</td>
</tr>
<tr>
<td>2</td>
<td>English</td>
<td>Drama</td>
<td>17th C</td>
<td>Sports</td>
</tr>
<tr>
<td>3</td>
<td>German</td>
<td>Essays</td>
<td>18th C</td>
<td>Politics</td>
</tr>
<tr>
<td>4</td>
<td>French</td>
<td>Fiction</td>
<td>19th C</td>
<td>People</td>
</tr>
<tr>
<td>5</td>
<td>Spanish</td>
<td>Humor</td>
<td>20th C</td>
<td>Finance</td>
</tr>
</tbody>
</table>

**Notation**
Collection of 19th c. French poetry = A4C1E4

### Unique identifier

- The goal of classification is to "lump" similar objects together.
- How do we "split" out individual objects?
  —by adding a unique identifier.
- This can be . . .
  - the record ID number
  - a unique code already in/on the object
  - a derived code (e.g., Cutter number)

### Exploring the challenges

- The IOP presents the opportunity to:
  - Engage in subject analysis and assignment of subject terms
  - Provide subject representation of your objects
- In IOP section 4 you will:
  - Do subject analysis
  - Create a thesaurus
  - Assign controlled vocabulary terms to objects
  - Create and apply a classification system

### Concepts and terms

- **Abstracting**
- **Access points**
- **Automatic indexing**
- **Bibliographic access**
- **Bibliographic classification**
- **Call number**
- **Class**
- **Classification**
- **Classification system**
- **Colocation**
- **Controlled vocabulary**
- **DDC**
- **Depth of indexing**
- **Descriptive cataloging**
- **Descriptor**
- **Document analysis**
- **Enumerative**
- **Exhaustive**
- **Facet**
- **Faceted classification**
- **Hierarchical classification**
- **Hierarchy**
- **Indexing**
- **Indexing language**
- **Intellectual access**
- **LCCC**
- **LCSH**
- **Natural language**

- **Notation**
- **Physical arrangement**
- **Pre-coordination**
- **Post-coordination**
- **Specificity**
- **Subject**
- **Subject analysis**
- **Subject cataloging**
- **Subject representation**
- **Summarization**
- **Synopsis**
- **Thesaurus**
- **Unique code/number**
- **Verbal subject representation**