An Overview of Indexing Approaches

Joseph A Busch, Principal
Today’s talk … more or less

- Approaches to indexing
  - Natural language vs. controlled vocabulary
  - Human computer-assisted indexing vs. automatic indexing
  - Faceted indexing and classification systems
  - Image indexing and retrieval
  - Indexing of the web

- Evaluating indexing
  - Indexing system performance
  - Indexer consistency
An Overview of Indexing Approaches

APPROACHES TO INDEXING
Approaches to indexing: Overview

- Myth #1: Taxonomies are monolithic hierarchies
- Myth #2: People retrieve content by topical subjects
- Myth #3: Nobody else can index content
- Myth #4: All a search engine can retrieve is a list
Natural language vs. controlled vocabulary?

- 19% Character errors. (Young, et al)
- 40% Vocabulary errors. (Seaman)
- 20% Index confusion.
- 21% Successful (Nielsen)

... It’s all about Search Failure
Problems with metadata

- Inconsistent category assignments
  - CA vs. California
  - RiM vs. Research in Motion

- Changes to classification systems over time
  - ICD-9 vs. ICD-10
  - SIC vs. NAICS

- Use of multiple overlapping or different categorization schemes
  - States vs. SMSA’s
  - ICD-9 vs. CDC Diseases and Conditions
  - NASA Taxonomy vs. NASA Thesaurus
Imagine a world where the same plant is called different names in every town in every valley in every country. It existed before 1753, when Linnaeus brought order to the chaos with his famous work *Systema Naturae*.
Search solution

- Generate more consistent content to search on.
- Correct user errors.
- Map the language of users to the language of the target content.
- Augment search results with linked data.
## What does controlled vocabulary do for search?

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Related search</td>
<td>Query corrections … did you mean?</td>
</tr>
<tr>
<td>Concept search</td>
<td>Query expansion with synonyms, abbreviations, acronyms, etc. … do you also want?</td>
</tr>
<tr>
<td>Ontology-based search</td>
<td>Query expansion with narrower or broader terms; scoping exhaustive search results</td>
</tr>
<tr>
<td>Faceted search</td>
<td>Dynamic filtering of search results; online shopping</td>
</tr>
<tr>
<td>Clustering</td>
<td>Dynamically bucketing search results into pre-defined categories</td>
</tr>
<tr>
<td>Subscriptions</td>
<td>RSS feeds, alerts, SDI (selective dissemination of information), etc.</td>
</tr>
<tr>
<td>Personalization</td>
<td>Weighting search results based on explicit profiles and implicit data (where you’ve been and what you’ve done)</td>
</tr>
</tbody>
</table>
Problem:

- Need to promote agency *behavioral health* program to heterogeneous audiences:
  - Human services professionals
  - Concerned family
  - Policy makers

- Merge heterogeneous information sources:
  - Alcohol and drug information
  - Mental health information
  - Other agency and inter-agency resources
    - Drug Abuse Warning Network (DAWN)
    - Treatment Episode Data Set (TEDS)
    - Uniform Reporting System (URS)
Case study: Multiple categorization schemes (2)

Solution:

- Faceted content tagging and navigation taxonomy
  - Powers the [SAMHSA Store](#) as illustrated in a [YouTube video](#)
  - The framework for agency key performance indicators.
  - Increases the availability and visibility of SAMHSA information.
  - Offers [tools](#) for analysis, visualization and mash ups with other sources.
Case study: Multiple categorization schemes (3)

SAMHSA Store Taxonomy facets
Case study: Multiple categorization schemes (4)
Case study: Multiple categorization schemes (5)

SAMHSA Info Tools

Taxonomy Strategies The business of organized information
Solution for search alternatives

- Predictable standardized structures, and
- Consistent semantics to work on

... so machines can understand it.
Augmented search results with linked data

Google's new right rail
Managed vocabularies

- Names of people, organizations, products, events, locations, etc.
  + Alternate labels
    - Synonyms
    - Abbreviations
    - Acronyms
    - etc.
  + Additional information
    - Unique identifiers
    - Coverage dates
    - Descriptions
    - etc.

- A set of concepts, optionally including statements about semantic relationships between those concepts.
Semantic schemes: Simple to complex

A system for identifying and naming things, and arranging them into a classification according to a set of rules.

An arrangement of knowledge usually enumerated, that does not follow taxonomy rules. E.g., Dewey Decimal Classification.

A set of words/phrases that can be used interchangeably for searching. E.g., Hypertension, High blood pressure.

A list of preferred and variant terms.

A facet taxonomies but uses richer semantic relationships among terms and attributes and strict specification rules.

A tool that controls synonyms and identifies the semantic relationships among terms.

Equivalence

Hierarchy

Associative

Synonym Ring

Controlled Vocabulary

Taxonomy

Classification Scheme

Thesaurus

Ontology

After: Amy Warner. *Metadata and Taxonomies for a More Flexible Information Architecture*
Faceted indexing and classification systems

- Searching for content should be as easy as shopping for shoes on Zappo’s!
What is a taxonomy?

- A categorization framework agreed upon by business and content owners (with the help of subject matter experts) that will be used to tag content.
  - 6 broad, discrete divisions (called facets)
  - 2-3 levels deep.
  - Up to 15 terms at each level.
  - 1200 terms total.
  - With some logic—hierarchical, equivalent and associative relationships between terms.
Effectiveness of taxonomies

- Categorize in multiple, independent, categories.
- Allow combinations of categories to narrow the choice of items.
- 4 independent categories of 10 nodes each have the same discriminatory power as one hierarchy of 10,000 nodes ($10^4$)
  - Easier to maintain.
  - Easier to reuse existing material.
  - Can be easier to navigate, if software supports it.

```plaintext
Main Ingredients
- Chocolate
- Dairy
- Fruits
- Grains
- Meat & Seafood
- Nuts
- Olives
- Pasta
- Spices & Seasonings
- Vegetables

Meal Type
- Breakfast
- Brunch
- Lunch
- Supper
- Dinner
- Snack

Cuisines
- African
- American
- Asian
- Caribbean
- Continental
- Eclectic/ Fusion/ International
- Jewish
- Latin American
- Mediterranean
- Middle Eastern
- Vegetarian

Cooking Methods
- Advanced
- Bake
- Broil
- Fry
- Grill
- Marinade
- Microwave
- No Cooking
- Poach
- Quick
- Roast
- Sauté
- Slow Cooking
- Steam
- Stir-fry
```

42 values to maintain ($10+6+11+15$)
9900 combinations ($10\times6\times11\times15$)
### 9 Common taxonomy facets

<table>
<thead>
<tr>
<th>Facet</th>
<th>Definition</th>
<th>Example Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Type</td>
<td>The various genres of content being created, managed and/or used.</td>
<td>AGLS Document Type, AAT Information Forms, Records management policy, etc.</td>
</tr>
<tr>
<td>Audience</td>
<td>Subset of constituents to whom a content item is directed or intended to be used.</td>
<td>GEM, ERIC Thesaurus, IEEE LOM, etc.</td>
</tr>
<tr>
<td>People</td>
<td>Names of important people such as authors, politicians, leaders, actors, etc.</td>
<td>LC NAF, NYTimes Topics-People</td>
</tr>
<tr>
<td>Organization</td>
<td>Names of organizations, their aliases and the relationships between them.</td>
<td>FIPS 95-2, D&amp;B, Ticker Symbols, LC NAF, NYTimes Topics-Organizations, etc.</td>
</tr>
<tr>
<td>Industry</td>
<td>Broad market categories such as industry sector codes.</td>
<td>FIPS 66, SIC, NAICS, etc.</td>
</tr>
<tr>
<td>Location</td>
<td>Names of places of operations, activities, constituencies, etc.</td>
<td>ISO 3166, FIPS 5-2, FIPS 55-3, USPS, NYTimes Topics-Places etc.</td>
</tr>
<tr>
<td>Function</td>
<td>Activities and processes performed to accomplish goals.</td>
<td>FEA Business Reference Model, AAT Functions, etc.</td>
</tr>
<tr>
<td>Product</td>
<td>Names of products and services that are produced by an organization or people.</td>
<td>Household Products Database, etc.</td>
</tr>
<tr>
<td>Topic</td>
<td>Topical subjects and themes that are not included in other facets.</td>
<td>LCSH, NYTimes Topics-Subjects, etc.</td>
</tr>
</tbody>
</table>
Image indexing and retrieval

❖ “Mute” content requires text annotations to search on
  ▪ Captions/transcriptions
  ▪ Metadata with controlled values

❖ Ways to improve retrieval
  ▪ Lexicons and other types of semantic resources to interpret and enhance user queries (same as with text searching)
  ▪ Collections of indexed content to serve as “training sets”

❖ Training sets
  ▪ WordNet database has 117,000 English synonym sets (synsets)
  ▪ ImageNet database has 14 million labeled objects based on WordNet. (NY Times, 11/20/12)
Indexing of the web

- Discover web pages
  - Support indexing, harvesting and linking to web pages
  - Enable organic (web) and website search
  - Use Dublin Core which is the ISO and W3C standard

- Linked data to combine information from more than one source on the web
  - Support mash-ups
  - Publish vocabulary namespaces (Web enabled unique ID’s)
  - Use SKOS (Simple Knowledge Organization System) which is a W3C standard
Negotiations With Iran Over Nuclear Program May Resume

By STEVEN LEW MYERS and RICK GLADSTONE
Published: February 17, 2012

WASHINGTON — The United States and the European Union signaled on Friday that negotiations with Iran over its nuclear program could soon resume for the first time in years, even as a telecommunications network vital to the global banking industry prepared to expel Iranian banks.

While senior American and European officials stopped short of declaring a diplomatic breakthrough, Iran dropped previously unacceptable preconditions for talks in a letter this week from its senior nuclear negotiator, Saeed Jalili, who declared his country’s “readiness for dialogue” at “the earliest possibility.”

After weeks of official bluster, ominous threats of military action, and the cancellation of the annual nuclear conference in the United States, some signs of a possible thaw emerged in the most disputed region of the world.

"NY Times linked data"
Oracle events mashup with Google maps

[Image of Oracle events application with Google Maps integration]
An Overview of Indexing Approaches

Evaluation indexing
## How to demonstrate indexing system performance

<table>
<thead>
<tr>
<th>Goal</th>
<th>Number of Items</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illustrate metadata schema</td>
<td>1-3</td>
<td>Random (excluding junk)</td>
</tr>
<tr>
<td>Develop training documentation</td>
<td>10-20</td>
<td>Show typical &amp; unusual cases</td>
</tr>
<tr>
<td>Qualitative test of small vocabulary (&lt;100 categories)</td>
<td>25-50</td>
<td>Random (excluding junk)</td>
</tr>
<tr>
<td>Quantitative test of vocabularies*</td>
<td>3-10X number of categories</td>
<td>Use computer-assisted methods when more than 10-20 categories. Pre-existing metadata is the most meaningful.</td>
</tr>
</tbody>
</table>

*Quantitative methods require large amounts of tagged content. This requires specialists, or software, to do tagging. Results may be very different from how “real” users would categorize content.
How evenly does it divide the content?

- Documents do not distribute uniformly across categories
- Zipf (long tail) distribution is expected behavior
- 80/20 Pareto rule in action
How does taxonomy “shape” match that of content?

<table>
<thead>
<tr>
<th>Term Group</th>
<th>% Terms</th>
<th>% Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrators</td>
<td>7.8</td>
<td>15.8</td>
</tr>
<tr>
<td>Community Groups</td>
<td>2.8</td>
<td>1.8</td>
</tr>
<tr>
<td>Counselors</td>
<td>3.4</td>
<td>1.4</td>
</tr>
<tr>
<td>Federal Funds Recipients and Applicants</td>
<td>9.5</td>
<td>34.4</td>
</tr>
<tr>
<td>Librarians</td>
<td>2.8</td>
<td>1.1</td>
</tr>
<tr>
<td>News Media</td>
<td>0.6</td>
<td>3.1</td>
</tr>
<tr>
<td>Other</td>
<td>7.3</td>
<td>2.0</td>
</tr>
<tr>
<td>Parents and Families</td>
<td>2.8</td>
<td>6.0</td>
</tr>
<tr>
<td>Policymakers</td>
<td>4.5</td>
<td>11.5</td>
</tr>
<tr>
<td>Researchers</td>
<td>2.2</td>
<td>3.6</td>
</tr>
<tr>
<td>School Support Staff</td>
<td>2.2</td>
<td>0.2</td>
</tr>
<tr>
<td>Student Financial Aid Providers</td>
<td>1.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Students</td>
<td>27.4</td>
<td>7.0</td>
</tr>
<tr>
<td>Teachers</td>
<td>25.1</td>
<td>11.4</td>
</tr>
</tbody>
</table>

**Background:**
- Hierarchical taxonomies allow comparison of “fit” between content and taxonomy areas.

**Methodology:**
- 25,380 resources tagged with taxonomy of 179 terms. (Avg. of 2 terms per resource)
- Counts of terms and documents summed within taxonomy hierarchy.

**Results:**
- Roughly Zipf distributed (top 20 terms: 79%; top 30 terms: 87%)
- Mismatches between term% and document% are flagged in red.

Source: Courtesy Keith Stubbs, US. Dept. of Ed.
Indexer consistency

- Studies have consistently shown that levels of consistency vary, and that high levels of consistency are rare for:
  - Indexing
  - Choosing keywords
  - Prioritizing index terms
  - Choosing search terms
  - Assessing relevance
  - Choosing hypertext links

- Semantic tools and automated processes can help guide users to be more consistent.
Blind sorting of popular search terms

Results: Excellent
84% of terms were correctly sorted 60-100% of the time.

Difficulties
- For Methadone, confusion when, in this case, a substance is a treatment.
- For general terms such as Smoking, Substance Abuse and Suicide, confusion about whether these are Conditions or Research topics.
Content tagging consensus

Results: Good
Test subjects tagged content consistent with the baseline 41% of the time.

Observations
- Many other tags were reasonable alternatives.
- **Correct + Alternative tags** accounted for 83% of tags.
- Over tagging is a minor problem.
QUESTIONS

Joseph A Busch
Mobile 415-377-7912
jbusch@taxonomystrategies.com
Abstract

- This session will provide an overview of the various approaches to indexing and the issues to be considered when designing a subject access system. The overview will touch on natural language vs. controlled vocabulary, human computer-assisted indexing vs. automatic indexing, faceted indexing and classification systems; image indexing and retrieval; and indexing of the web. The session will also discuss indexing system performance and evaluations of indexer consistency.