Cost-Benefits of Applying Controlled Vocabularies

NKOS – Saturday - 15 October 2016

Jay Ven Eman, Ph.D., CEO
Access Innovations, Inc. / Data Harmony
+1.505.998.0800 / www.accessinn.com / www.dataharmony.com
j_ven_eman@accessinn.com
Search…

…doesn’t work!
Indexing Helps

- 5% to 10% “improvement” as reported yesterday:

- Ying-Hsang Liu, “University Metadata and Retrieval: The Death of the Library Catalog?” DC-2016, Copenhagen, Denmark

- 14 October 2016
Indexing your content with controlled vocabularies

2016-10-13
“Solving the Indexing Challenge”
By Jay Ven Eman

The process of indexing a content object begins with...

Unstructured

Data Harmony
MAstro Server

Categorization System

Content Repository
e.g. Database

Structured

<Category>Indexing</Category>
<Category>Thesauri</Category>
<Category>Standards</Category>
<Category>Classification</Category>
How much could 10% help?

The Pain of Search

<table>
<thead>
<tr>
<th>Mission critical</th>
<th>Percent</th>
<th>Number of Employees</th>
<th>Search &amp; Use Time Per Week</th>
<th>Time Searching Per Week</th>
<th>Time Analysing Per Week</th>
<th>Average Loaded Salary $ Per Hour</th>
<th>Annual Cost of Looking</th>
<th>Search Time Reduction</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>10</td>
<td>100</td>
<td>14</td>
<td>8.4</td>
<td>5.6</td>
<td>200</td>
<td>8,736,000</td>
<td>7,862,400</td>
<td>873,600</td>
</tr>
<tr>
<td>Medium</td>
<td>80</td>
<td>800</td>
<td>12</td>
<td>7.2</td>
<td>4.8</td>
<td>150</td>
<td>44,928,000</td>
<td>40,435,200</td>
<td>4,492,800</td>
</tr>
<tr>
<td>Low</td>
<td>10</td>
<td>100</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>100</td>
<td>3,120,000</td>
<td>2,808,000</td>
<td>312,000</td>
</tr>
</tbody>
</table>

$56,784,000 $51,105,600 $5,678,400
The Problem - Mismatch
Some perspective

- Online from the 70’s
  - Dialog
  - Data Star
  - Many others
- Secondary publishers
  - Mead – LexisNexis®
  - CAS
  - NASA & DOE & many others
Online search services

- Worked very well
  - Focused
  - Controlled
  - Specialized

- Content analysis
  - Database design - context
  - Extensive markup
  - Proprietary formats (Dialog format b)
The one goal, the one ring, the holy grail

- Computer science
  - Understanding human language
- Physics
  - Unified field theory
So back at the computer lab

- Computer science
  - Full text
  - Isolated
  - Content without context
- Developing shortcuts became critical
  - Relevance
  - Weighting
  - Probabilities
Many approaches

- Bayesian
- Inference
- Vector
- Natural language
- Neural linguistic
- Computational linguistics
- Statistical
- Co-occurrence
- Morphological
- Grammatical
- Lemmatization
- Semantic
- Syntactic
- Phraseological
- Clustering
- Full text
Search in the real world

- Structured – (multiple meanings)
- Unstructured
- Applications environment
- Turf wars
- Language wars
  - Ownership
  - Role-based language
“Meaning” starts with a knowledge organization system (KOS)

- Uncontrolled list
- Name authority file
- Synonym set/ring
- Controlled vocabulary
- Taxonomy
- Thesaurus

SKOS

Not complex - $

Highly complex - $$$$
Taxonomic strategy

- Can save search
  - Taxonomy like an ordinance map
    - Latitude, longitude
    - Rosetta Stone
  - Search like a treasure map
    - Fun – clustering is likable, but lacks consistency and reproducibility
    - Dangerous, time consuming, fraught with hazards like searching for the ‘Black Pearl’
- But not by itself
Comprehensive Information Strategy

- User needs – who, what, when, where
- Business drivers
- Information flow(s)
  - Origin
  - Production
  - Destination
  - Delivery
  - Disposition
  - Storage/Retrieval
  - Reuse
Information strategy

- Meta-data strategy
  - Taxonomy – Got one? Get one!
  - Indexing process – How to!
  - Structural elements (e.g. Dublin Core, JATS)
    - DTD, schema
    - Markup
- Promotion, advertising, training
- Maintenance, upkeep
Information strategy

- Metadata strategy - limitations
- Results only as good as the metadata
- Indexing quality
  - Subjective
  - Can get expensive
  - Automation essential
More benefits than just improving retrieval

- Image indexing
- Mining for grants & funding
- Patent mining & analysis
- Hot-topic page production
- Ad serving
- Analytics for product development, competitive analysis, metrics
More than improving retrieval

- Website navigation
  - Broaden / narrow
  - Related concepts
- Better recommendations
- Improve e-commerce
  - The correct items
  - Related items
- Affinity
- Peer review management
Cart then horse

- Information strategy must be done first!
- Then shop for search software
- Select search software with the features & functions that will drive your content.
- Or else…
Cost-Benefits of Applying Controlled Vocabularies

Jay Ven Eman, Ph.D., CEO
Access Innovations, Inc. / Data Harmony
+1.505.998.0800 / www.accessinn.com / www.dataharmony.com
j_ven_eman@accessinn.com