Knowledge Organization Systems and Search

Thursday, 11 September 2008
10:45 – 11:15 AM EDT
Presented by Jay Ven Eman, Ph.D., CEO
Access Innovations, Inc. / Data Harmony – woman-owned, small business
505.998.0800 / www.accessinn.com / www.dataharmony.com
j_ven_eman@accessinn.com

CENDI & NKOS
Networked Knowledge Organization Systems
Search?

Doesn’t work!
## How bad is it?

### The Pain of Search

<table>
<thead>
<tr>
<th>Mission critical</th>
<th>Percent</th>
<th>Number of Employees</th>
<th>Search &amp; Use Timel Use 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>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>10</td>
<td>1000</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>
<td></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>
<td></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>
<td></td>
</tr>
</tbody>
</table>

$56,784,000 $51,105,600 $5,678,400
Mismatch

Search software

Audience

Contents

Model

DataHarmony

Access Innovations
Many approaches

A
- Bayesian
- Inference
- Vector
- Natural language
- Neural linguistic
- Computational linguistics
- Statistical
- Clustering

B
- Morphological
- Grammatical
- Lemmatization
- Semantic
- Syntactic
- Phraseological
- Clustering
- Co-occurrence
The one goal – the holy grail

- Computer science
  - Understanding human language
- Physics
  - Unified field theory
In the meantime

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

- Worked very well
  - Focused
  - Controlled
  - Specialized

- Content analysis
  - Database design - context
  - Extensive markup
  - Proprietary formats (Dialog format b)
Back at the lab

- Computer science
  - Full text
  - Isolated
  - Content without context
- Developing shortcuts became critical
  - Relevance
  - Weighting
  - Probabilities
Search in the real world

- Structured
- 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 - $

Ontology
Topic Map

LOTS OF OVERLAP!

Highly complex - $$$$$
Taxonomic strategy

- Can save search
  - Taxonomy like a USGS map
    - Latitude, longitude
    - Rosetta Stone
  - Search like a treasure map
    - Fun – clustering is likable, but lacks consistency
    - Dangerous, time consuming, fraught with hazards like searching for the ‘Black Pearl’
Access customers say:

- “There is now a 92% accuracy rating accuracy on accounting and regulatory document search based on hit, miss and noise or relevance, precision and recall statistics…Access Innovations.” USGAO
- “IEEE had their system up and running in three days, in full production in less than two weeks.” Institute of Electrical and Electronics Engineers (IEEE)
- “The American Economic Association said its editors think using it is fun and makes time fly!” American Economic Association (AEA)
- “ProQuest CSA have achieved a 7 fold increase in productivity – thus they have four licenses.” ProQuest CSA
- “Weather Channel finds things 50% faster using Data Harmony. A significant saving in time.” The Weather Channel

DataHarmony

Access Innovations®
Taxonomies in action

- www.mediasleuth.com
- www.ask.com
- www.revolutionhealth.com
Go – No Go – What is good enough?

- Reach 85% precision to launch for productivity - assisted
- Reach 85% for filtering or categorization
  - Sorting for production
- Level of effort to get to 85%
- Integration into the workflow is efficient
Hit, Miss, Noise

- **Hit** – exactly what a human indexer would use
- **Miss** – human indexer would use but system did not assign
- **Noise** – system assigned but human did not
  - Relevant noise – could have been assigned
  - Irrelevant noise – just plain wrong
Subjective

- Relevance
  - Reflects how akin it is to the users request

- Aboutness
  - Reflects the topical match between the document content and the term
  - How well the topic describes what the document is about

- Varies with level of conceptual terms vs. factual terms in the thesaurus
Statistics

- **Precision**
  - Correct retrieval / Total retrieval
  - Hits / hits + noise

- **Recall**
  - Correct retrieval / Total correct in system
  - Hits / Hits + misses

- **Level of effort**
  - Hits / Hits + misses + noise
Benchmarks

- 15 – 20% irrelevant returns / noise
- Amount of work needed to achieve 85% level
- How good is good enough?
  - Satisfice = satisfaction + suffice
  - How good is good enough?
  - How much error can you put up with?
Information strategy

- User needs
- Business drivers
- Information flow(s)
  - Origin
  - Production
  - Destination
  - Delivery
  - Disposition
  - Storage/Retrieval
  - Reuse

Copyright © 2007 Access Innovations, Inc.
Information strategy

- Meta-data strategy
  - Taxonomy
  - Indexing
  - Structural elements (e.g. Dublin Core)
    - DTD
    - Markup
- Promotion, advertising, training
- Maintenance, upkeep
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…
Knowledge Organization Systems and Search

Thursday, 11 September 2008
10:45 – 11:15 AM EDT
Presented by Jay Ven Eman, Ph.D., CEO
Access Innovations, Inc. / Data Harmony – woman-owned, small business
505.998.0800 / www.accessinn.com / www.dataharmony.com
j_ven_reman@accessinn.com

Thank you!