CENDI
Federal STI Managers Group
&
NKOS Networked Knowledge Organization Systems

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

GSA Advantage!
www.gsaAdvantage.gov

Access Innovations [GS-35F-0820N]

Data Harmony

Access Innovations

Search?

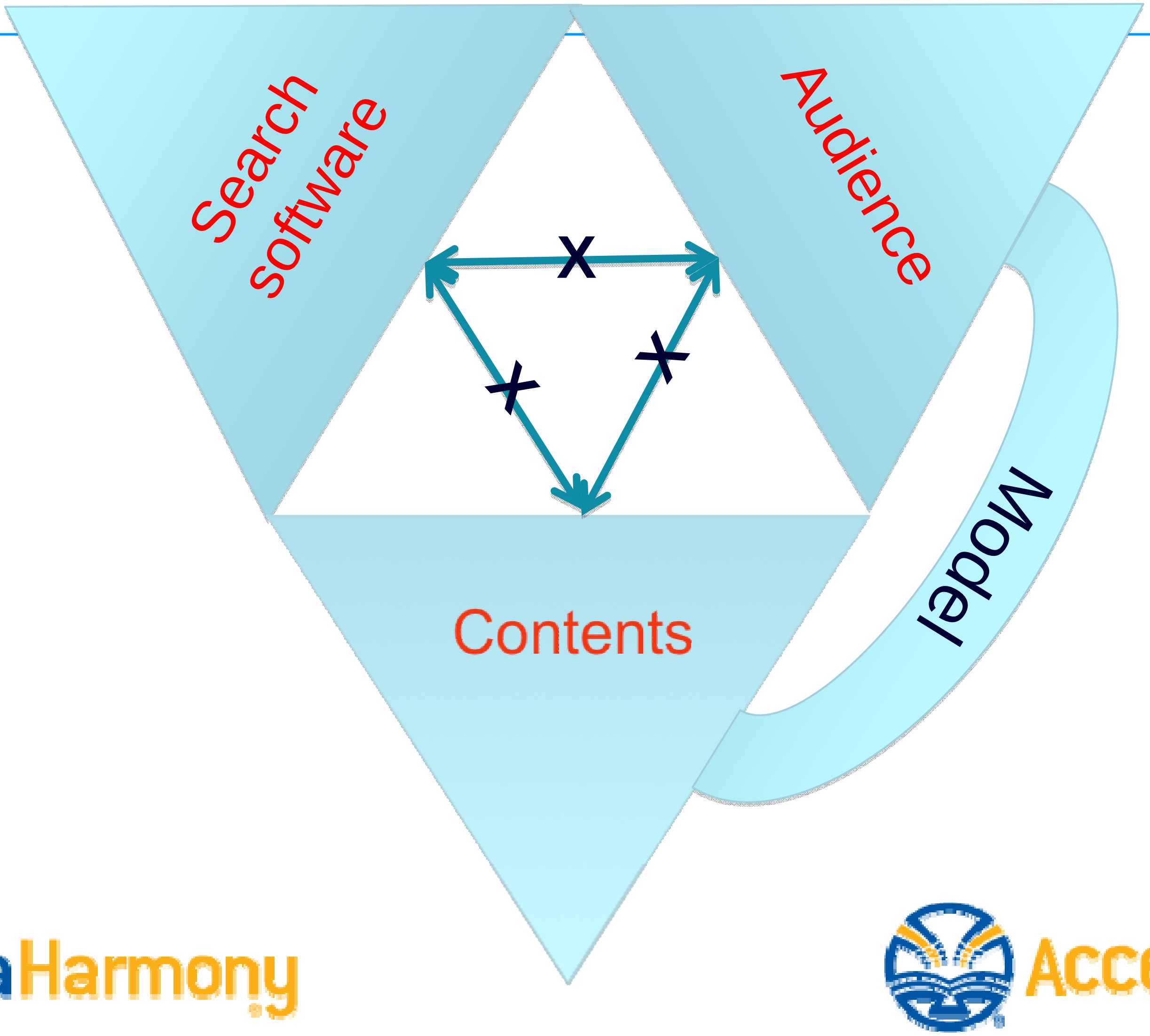
Doesn’t work!

How bad is it?

The Pain of Search

	Percent	Number of Employees	Search & Use TimeI Per Week	Time Searching Per Week	Time Analysing Per Week	Average Loaded Salary	Annual Cost of Looking	Search Time Reduction	Difference
Mission critical		1000	Hours	Hours	Hours	$ Per Hour		10%	
High	10	100	14	8.4	5.6	200	8,736,000	7,862,400	873,600
Medium	80	800	12	7.2	4.8	150	44,928,000	40,435,200	4,492,800
Low	10	100	10	6	4	100	3,120,000	2,808,000	312,000
							$56,784,000	$51,105,600	$5,678,400

Mismatch



Search software

Audience

Contents

Model

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)

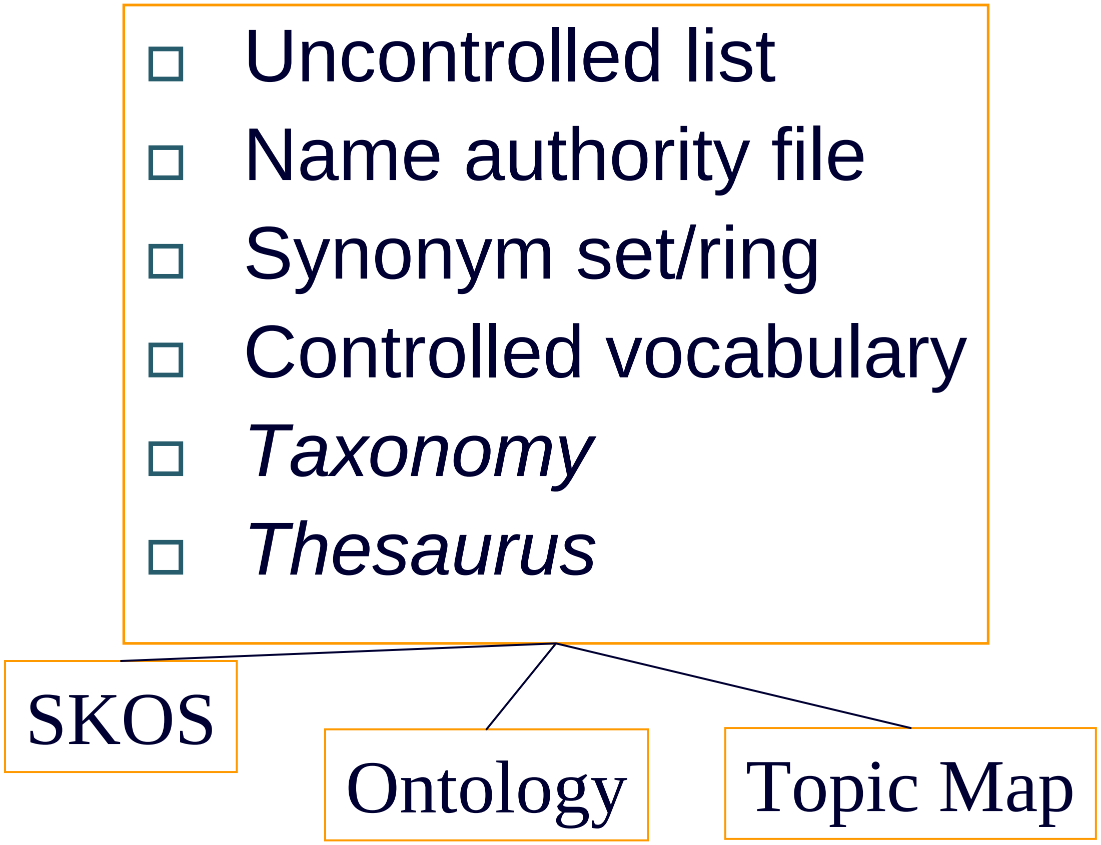
Back at the lab

- 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)

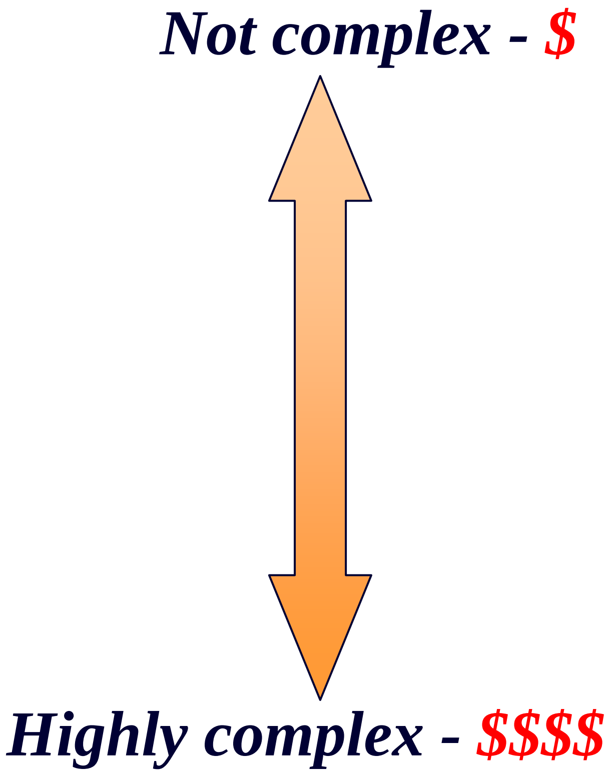


- □ Uncontrolled list
- □ Name authority file
- □ Synonym set/ring
- □ Controlled vocabulary
- □ *Taxonomy*
- □ *Thesaurus*

SKOS

Ontology

Topic Map



Not complex - $

Highly complex - $$$$

LOTS OF OVERLAP!

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.” In*stitute 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

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…

Thank you!

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

Access Innovations [GS-35F-0820N]