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

Access Innovations.

DataHarmony

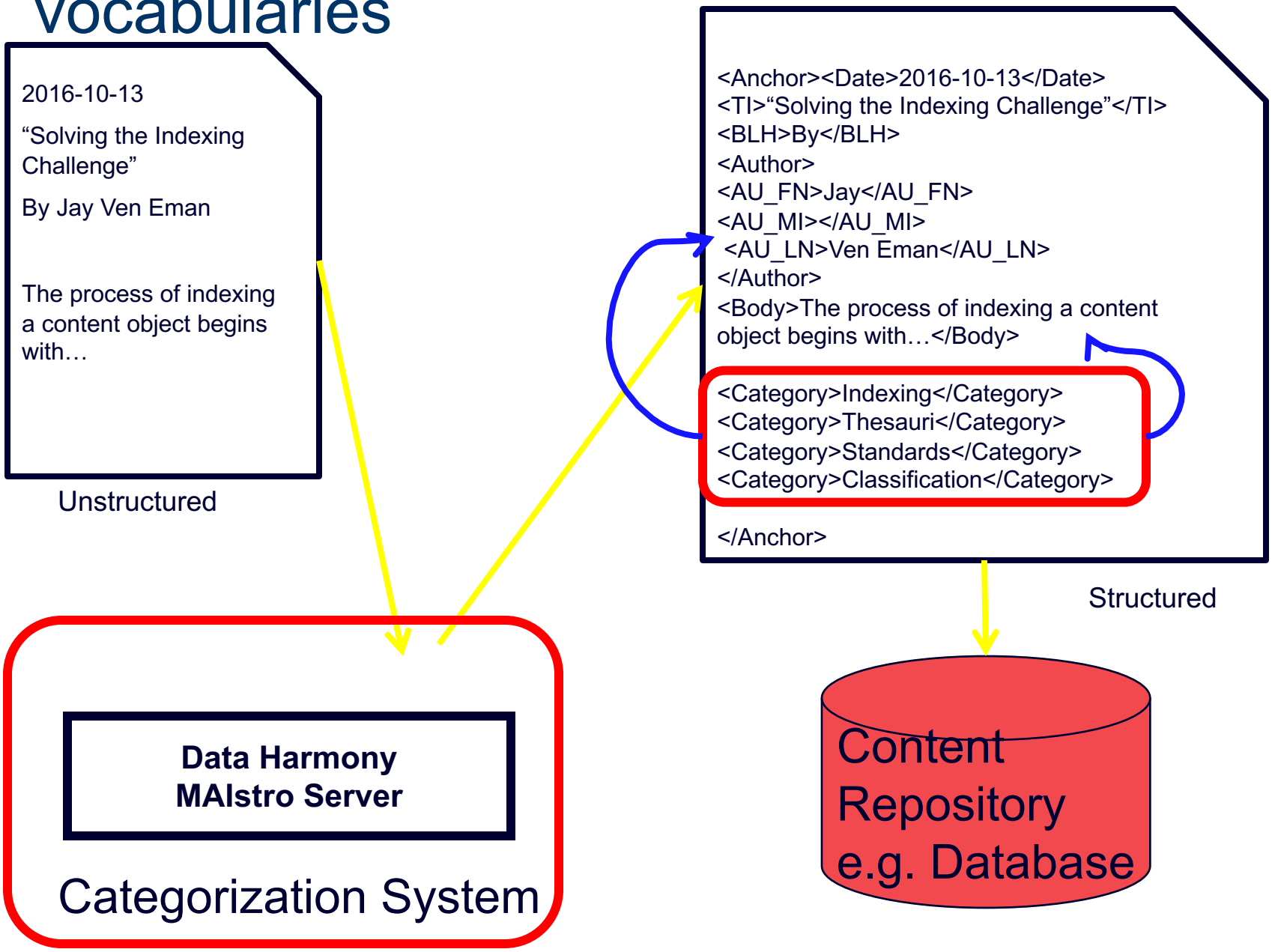
Search…

…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



How much could 10% help?

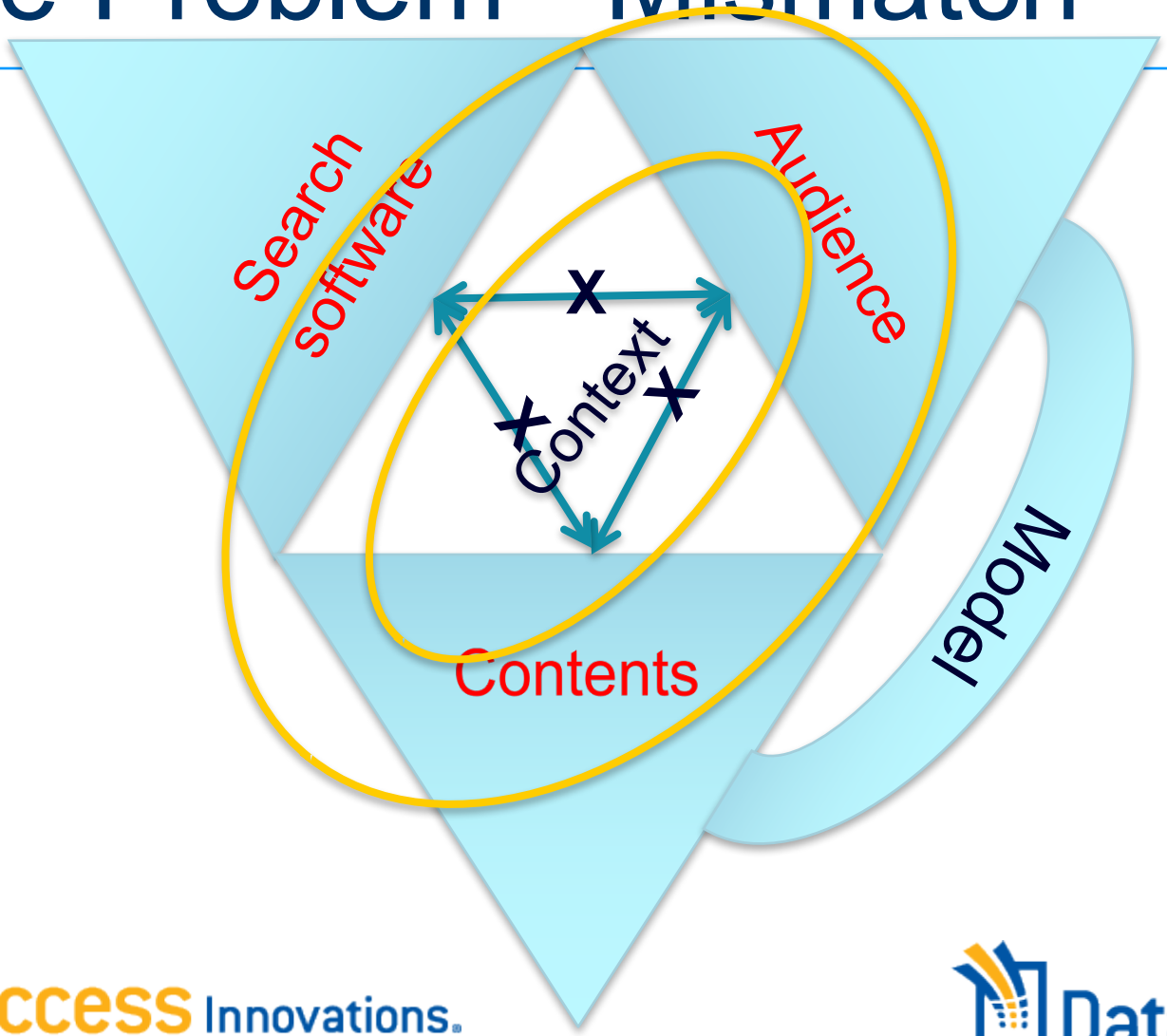
The Pain of Search

Mission critical	Percent	Number of Employees 1000	Search & Use Time Per Week Hours	Time Searching Per Week Hours	Time Analysing Per Week Hours	Average Loaded Salary $ Per Hour	Annual Cost of Looking	Search Time Reduction 10%	Difference
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

Access Innovations.

DataHarmony

The Problem - Mismatch



Access Innovations.

DataHarmony

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

A

- Bayesian
- Inference
- Vector
- Natural language
- Neural linguistic
- Computational linguistics
- Statistical
- Co-occurrence

B

- 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)

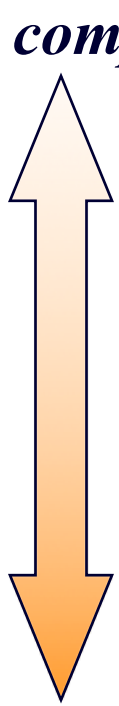
- Uncontrolled list
- Name authority file
- Synonym set/ring
- Controlled vocabulary
- *Taxonomy*
- *Thesaurus*

SKOS

Ontology

Topic Map

LOTS OF OVERLAP!



Access Innovations.

DataHarmony

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…

Thank you!

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

Access Innovations.

DataHarmony