Leveraging KOS-Fueled Semantic Technologies to Generate Tags and Tag Clouds

Denise A. D. Bedford, Ph.d.
Goodyear Professor of Knowledge Management

College of Communication and Information
Kent State University

Overview

- Research Context - Knowledge Architecture of the Future
- Role of Social Tags and Tagging in Knowledge Architecture
- Architecture Challenges and Opportunities for Social Tags
- Exploratory Research Proposal and In Progress Results

RESEARCH CONTENT - KNOWLEDGE ARCHITECTURE FOR THE FUTURE

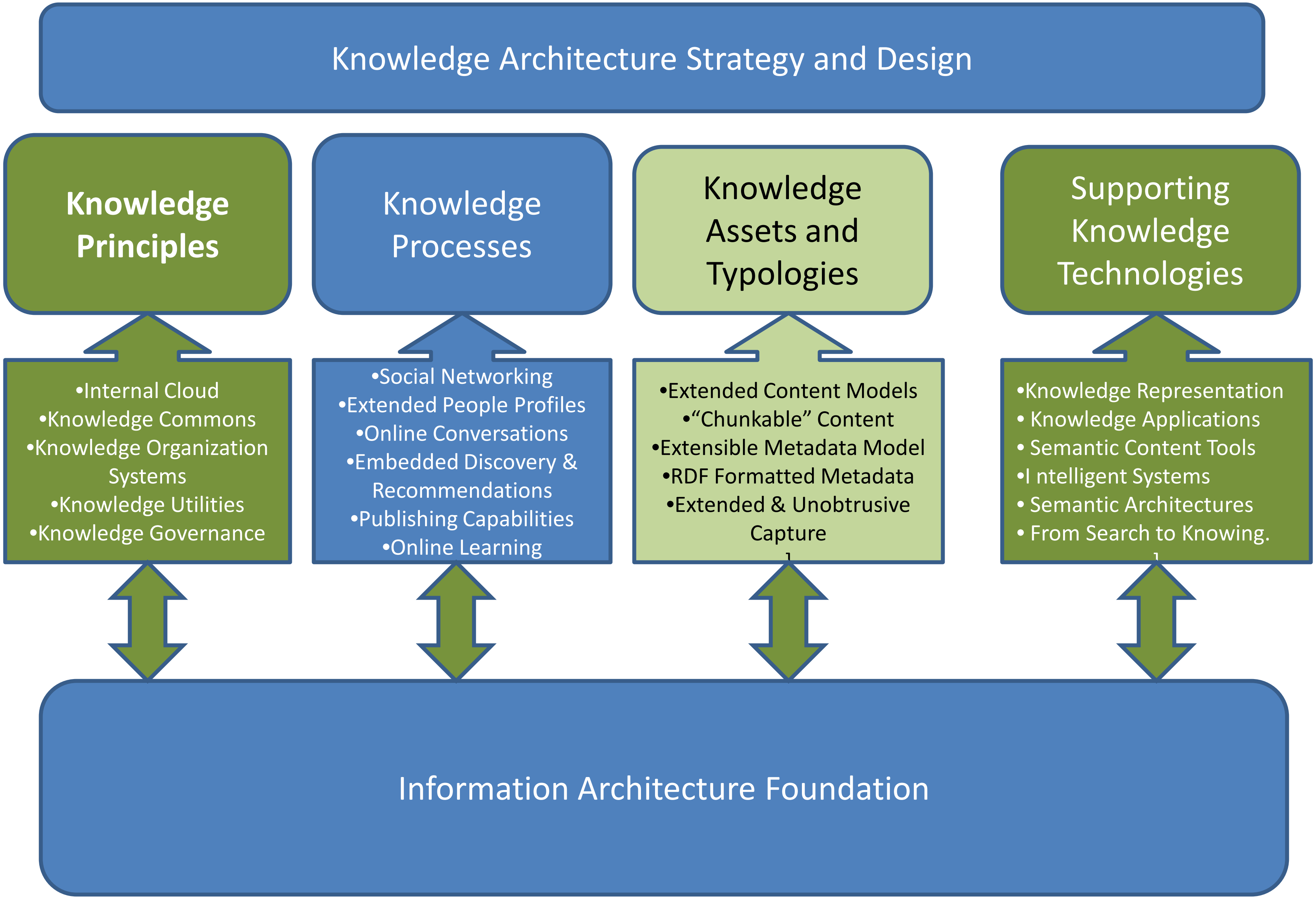
Architecture and Design

- Architecting a digital environment is not too different from architecting a house

- We consider who will live there, what they will do there, how they expect to work and interact in the environment

- We produce a series of blueprints that address different layers of functionality – business, information, knowledge, applications/software and technology infrastructure

- We produce blueprints by looking at principles, assets, pratices and technologies



Knowledge Architecture Strategy and Design

Knowledge Principles

- Internal Cloud
- Knowledge Commons
- Knowledge Organization Systems
- Knowledge Utilities
- Knowledge Governance

Knowledge Processes

- Social Networking
- Extended People Profiles
- Online Conversations
- Embedded Discovery & Recommendations
- Publishing Capabilities
- Online Learning

Knowledge Assets and Typologies

- Extended Content Models
- “Chunkable” Content
- Extensible Metadata Model
- RDF Formatted Metadata
- Extended & Unobtrusive Capture

Supporting Knowledge Technologies

- Knowledge Representation
- Knowledge Applications
- Semantic Content Tools
- I ntelligent Systems
- Semantic Architectures
- From Search to Knowing.

Information Architecture Foundation

Knowledge vs. Information Principles

Knowledge Principles

- Open
- Collaborative
- Transparent
- Interactive
- Perishable
- Embedded
- Extensible

Information Principles

- Common Vocabulary and Definition
- Accessible
- Meets End User Purpose
- Everyone’s Business
- Reused and Reusable
  Has Stewards
- Is Secure

Information Architecture versus Knowledge Architecture

Knowledge Life Cycle

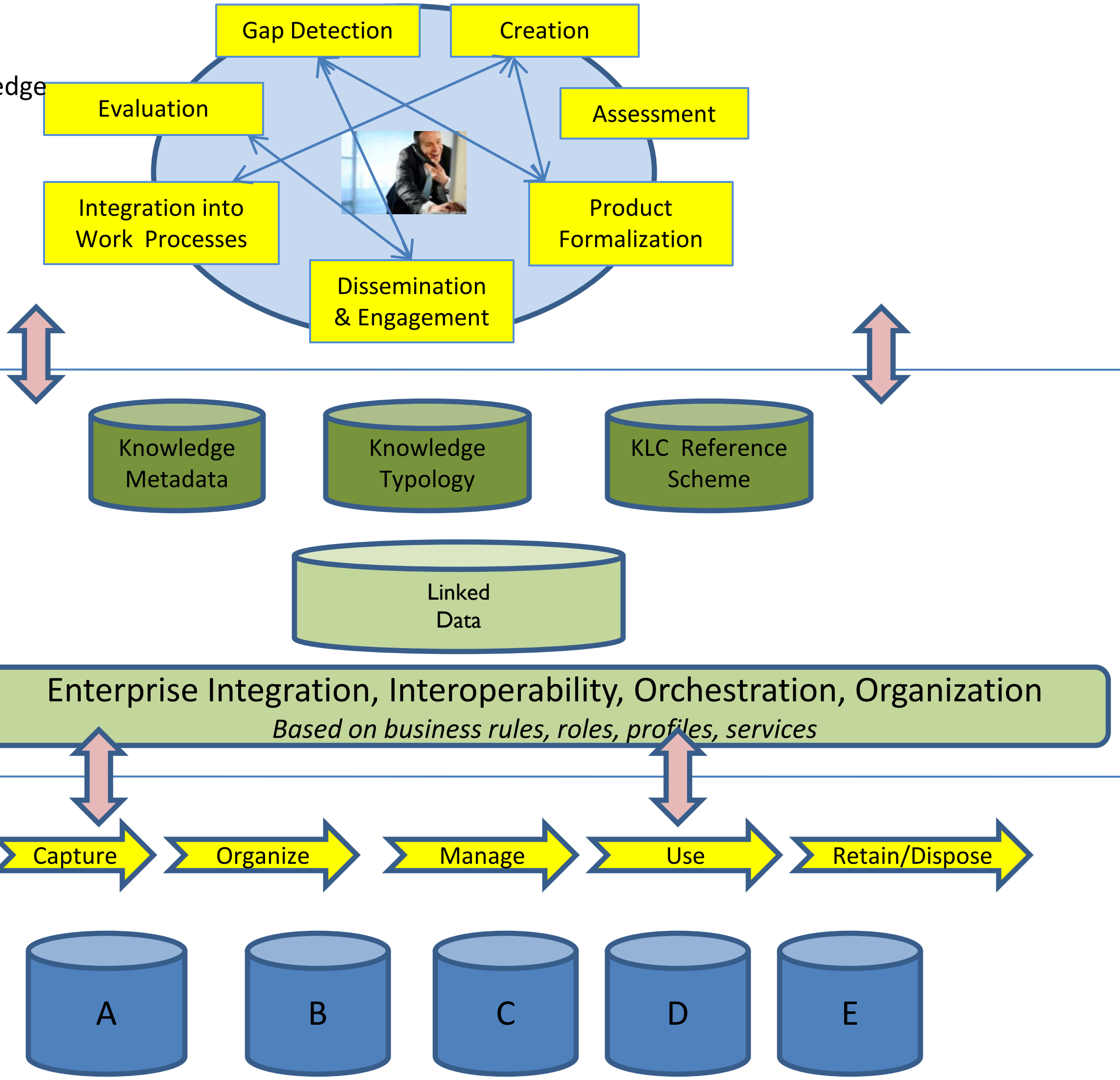
- Focus is people, connections and knowledge
- Manages expressed & tacit knowledge
- Collaboration, social networking are dynamic processes
- Knowledge is embedded in process
- Rapidly evolving technology markets
- Dynamic processes
- Plug and play components

Knowledge Architecture

- Rides on top of existing enterprise services and information architecture
- Comprised of interoperable tools
- Common K.A. components – typology, life cycle processes, knowledge structures,
- Seamless access to knowledge, info and data application across applications
- Implemented as 'linked data'

Information Architecture

- Linear process
- Steady State
- 3 – 5 yr. investment in enterprise applications
- Static, stable applications
- Backbone foundation



Gap Detection

Creation

Evaluation

Assessment

Integration into Work Processes

Product Formalization

Dissemination & Engagement

Knowledge Metadata

Knowledge Typology

KLC Reference Scheme

Linked Data

Enterprise Integration, Interoperability, Orchestration, Organization

Based on business rules, roles, profiles, services

Capture

Organize

Manage

Use

Retain/Dispose

A

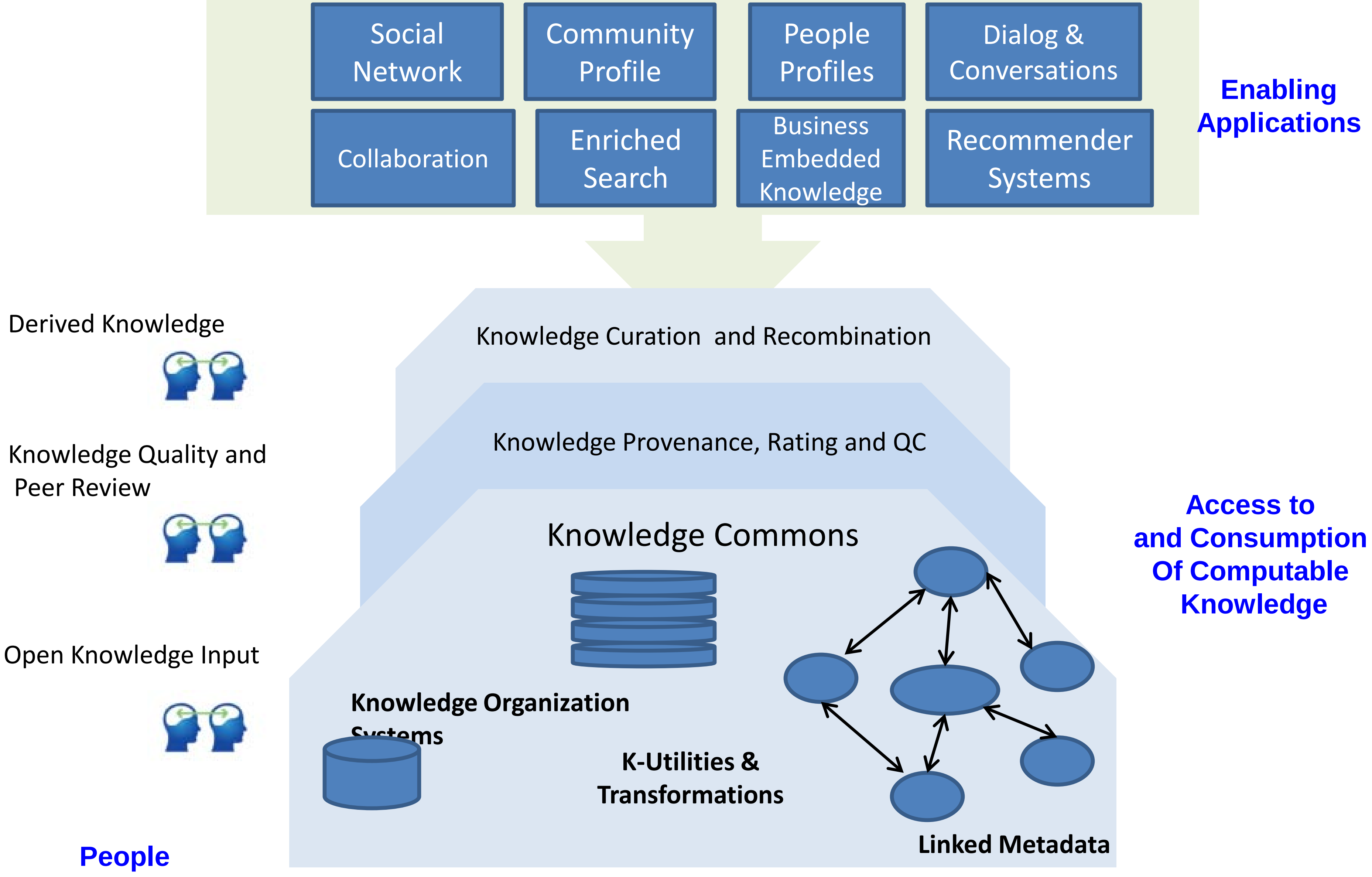
B

C

D

E

Vision of Future Knowledge Environment



Social Network

Community Profile

People Profiles

Dialog & Conversations

Collaboration

Enriched Search

Business Embedded Knowledge

Recommender Systems

Enabling Applications

Derived Knowledge

Knowledge Quality and Peer Review

Open Knowledge Input

People

Knowledge Curation and Recombination

Knowledge Provenance, Rating and QC

Knowledge Commons

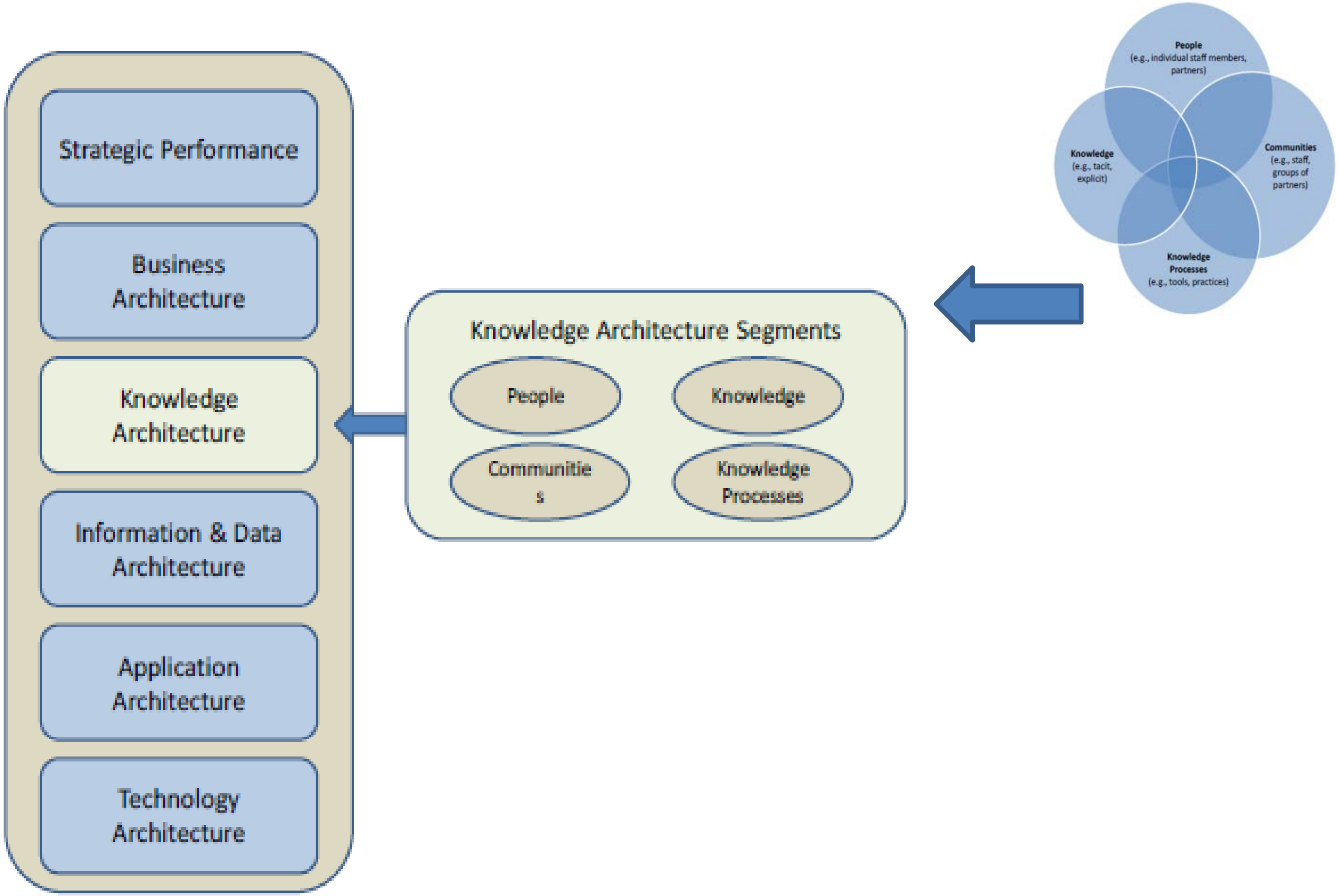
Knowledge Organization Systems

K-Utilities & Transformations

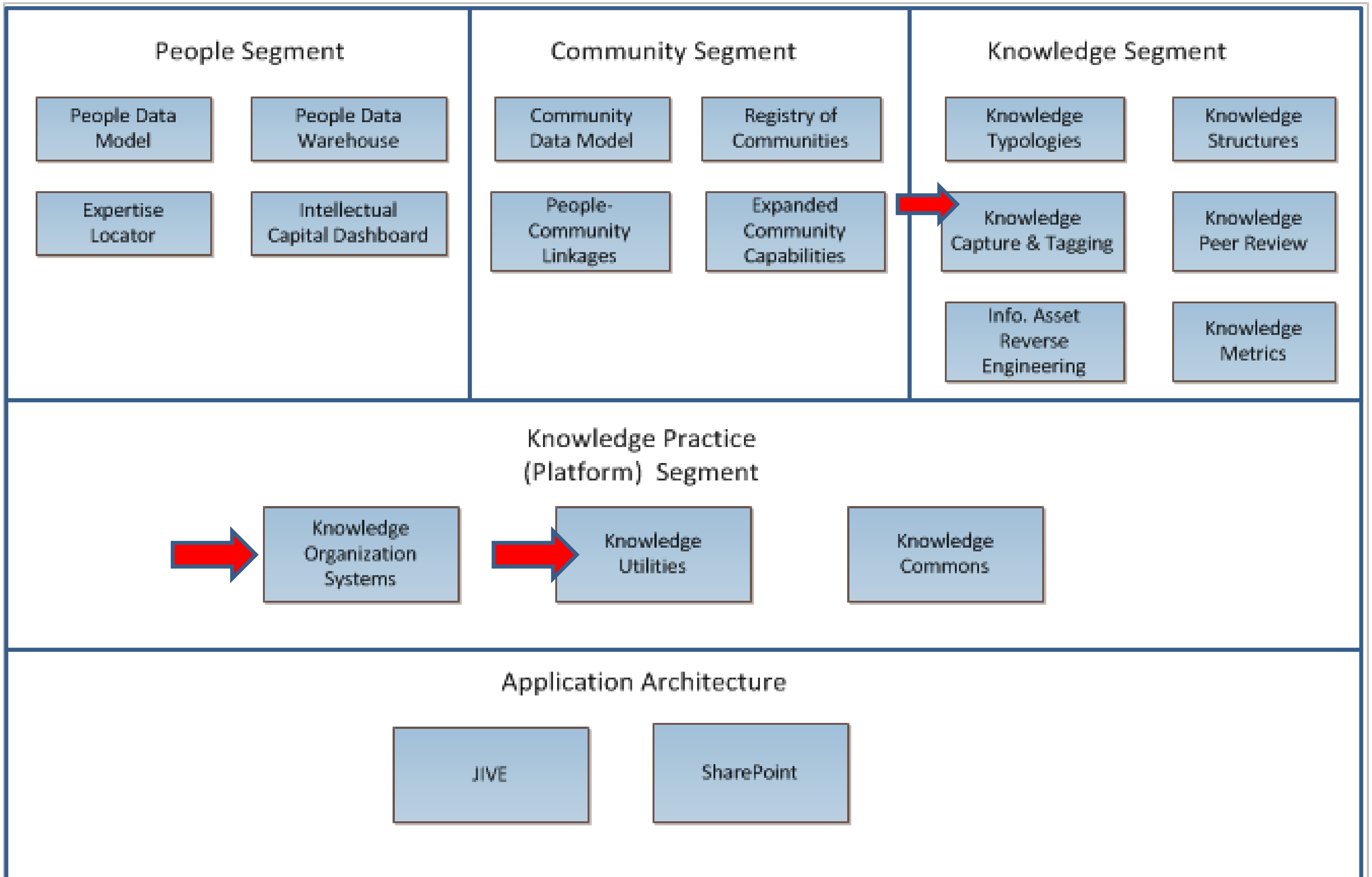
Linked Metadata

Access to and Consumption Of Computable Knowledge

Knowledge Architecture in an Enterprise Context


Strategic Performance
Business Architecture
Knowledge Architecture
Information & Data Architecture
Application Architecture
Technology Architecture
Knowledge Architecture Segments
People
Knowledge
Communitie s
Knowledge Processes
People (e.g., individual staff members, partners)
Knowledge (e.g., tacit, explicit)
Communities (e.g., staff, groups of partners)
Knowledge Processes (e.g., tools, practices)

Organizations need to start planning for the enterprise integration of KOS components



People Segment

- People Data Model
- People Data Warehouse
- Expertise Locator
- Intellectual Capital Dashboard

Community Segment

- Community Data Model
- Registry of Communities
- People-Community Linkages
- Expanded Community Capabilities

Knowledge Segment

- Knowledge Typologies
- Knowledge Structures
- Knowledge Capture & Tagging
- Knowledge Peer Review
- Info. Asset Reverse Engineering
- Knowledge Metrics

Knowledge Practice (Platform) Segment

- Knowledge Organization Systems
- Knowledge Utilities
- Knowledge Commons

Application Architecture

- JIVE
- SharePoint

Knowledge Architecture Blueprint

ROLE OF SOCIAL TAGS AND TAGGING IN KNOWLEDGE ARCHITECTURE

Social Tagging – Goals and Behaviors

- Users currently tag content for a variety of reasons using a variety of existing applications – tagging always takes place within an application

- Goulder and Huberman have identified several functions performed by tags
  - Identifying what or who the content is about
  - Identifying what it is
  - Identifying who owns it
  - Categorizing it or refining categories
  - Identifying qualities or characteristics
  - Providing self references
  - Aligning with a task or a business function

Role of Tagging in Knowledge Architecture

- We have seen that there is a clear role for tagging in the knowledge architecture of the future

- Tagging practices align with general knowledge organization and knowledge management functions

- Tagging can both augment access to knowledge and add value to KOS but this is not a trivial task or effort

- The question is how to accomplish this integration – and how to do it in the most effective and efficient way

KNOWLEDGE ARCHITECTURE CHALLENGES AND OPPORTUNITIES

Levels of Knowledge Architecture Functional Integration

- Application – Across applications where tags are treated as annotations that may be distinct knowledge objects
- User – Across applications, across all objects – to see a user's tags
- Knowledge Object – As extended metadata for a single object in any context where it may be used
- Knowledge Organization Systems – As extended values in KOS
- Tag – As faceted values and as distinct semantic units
- Integration has to begin at the tag level before we can move up the scale

Tag Level Integration Challenges

- Many tags appear to represent multiple facets, i.e. country + topic [Guy & Tomkin, Hammond et al, Pond]

- Tag values may be synonymous, i.e., mitigation, eradication, elimination [Golder & Huberman, Guy & Tonkin, Kroski, Mathes, Merholz, Powers, others]

- Tag values may be polysemous, i.e., contagion [Yi, Fernandez-Tobias, others]

- May represent different levels of specificity, i.e. “the basic problem” [Golder and Huberman, Kroski, others]

- Simple redundancy across tags - observed in some contexts but not yet documented in a controlled research environment

Tag Level Integration - Opportunities

- Tag values can be objective and aligned with KOS/LC [Lawson]

- Tags can be enhanced by knowledge organization systems [Matthews et al]

- An ontology of tag facets, with actual identified classes, is feasible but has not been to date – research has instead focused on:
  - UTO focused on clustered top concepts only [Ding, Jacob et al]
  - User vs. expert-assigned subject tags and LC Subdivisions [Lu, Park and Hu]

- Tag values can be recommended for users to select based on user tags identified through clustering/semantic similarity measures [Shiri, Razikin et al, Fu et al]

Tag Level Integration - Opportunities

- Within the knowledge architecture context, tags must be managed along three dimensions:
  - Metadata issues - user proposed values, professionally generated, semantically generated
  - Kinds of metadata -- full range of metadata facets appears to be represented in tags
  - Tag sets and tags as knowledge organization systems - ingest and reconciliation

- Challenges present a significant amount of labor intensive manipulation of tags and tag values before integration is possible

EXPLORATORY RESEARCH PROPOSAL AND IN PROGRESS RESULTS

Exploratory Research Proposal

- Is it possible to use KOS and semantic engines to semantically generate tags for user selection and promotion? Can semantic generation address all of the current functions supported by end user tagging?

- This is a significant research effort focused on three primary research questions:
  - Question 1: Can we use semantic engines to generate social tags that align with tags currently created by end users?
  - Question 2: Is it possible to more effectively manage tags when a KOS is embedded in the semantic engine?
  - Question 3: If we can generate tags semantically, will users select them?

Research Data and Context

- **Focus Area** – Topical information which is tagged to five areas: agriculture, environment, transport, health, education
- **Data Set** – Goal is to collect 300 examples in each topical area
- **Data Sources** – Open Web, CiteULike,
- **Data Capture** – Manual capture of tags, citations and full text for full testing
- **Research Methodology** – Semantic generation of tags using SAS/Teragram semantic engine with embedded Topic Knowledge Organization System
- **Review and Validation** – Manual review and comparison

Semantic Generation of Tags

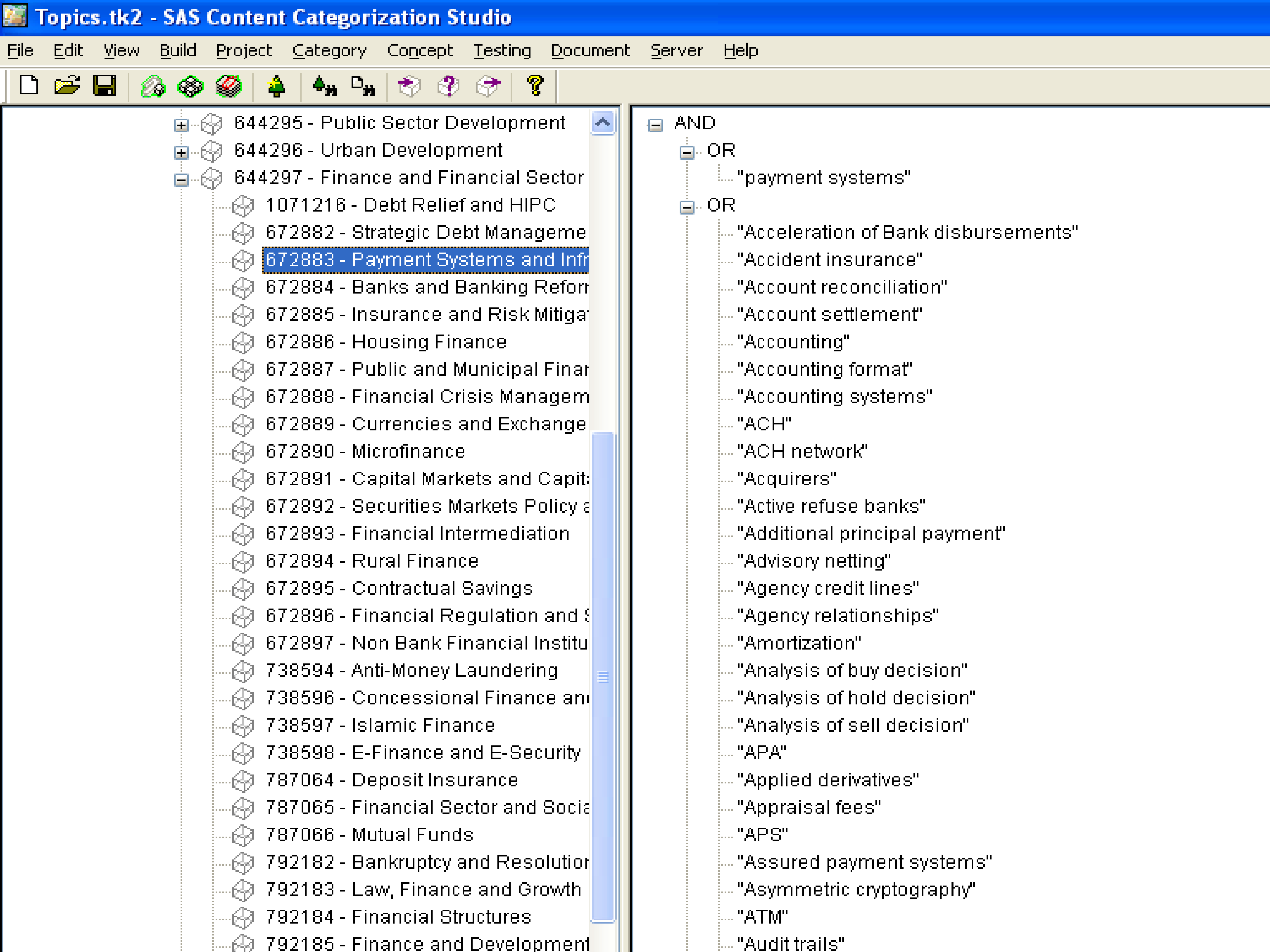
- Use a semantic engine with a strong NLP foundation to support categorization and conceptual indexing

- Semantic engine enables integration of KOS – we are leveraging the World Bank’s original topic classification scheme as a cross-topic deep conceptual thesaurus

- Semantic engine semantically indexes the content, applies the topic profile and then generates concepts (i.e., tags)

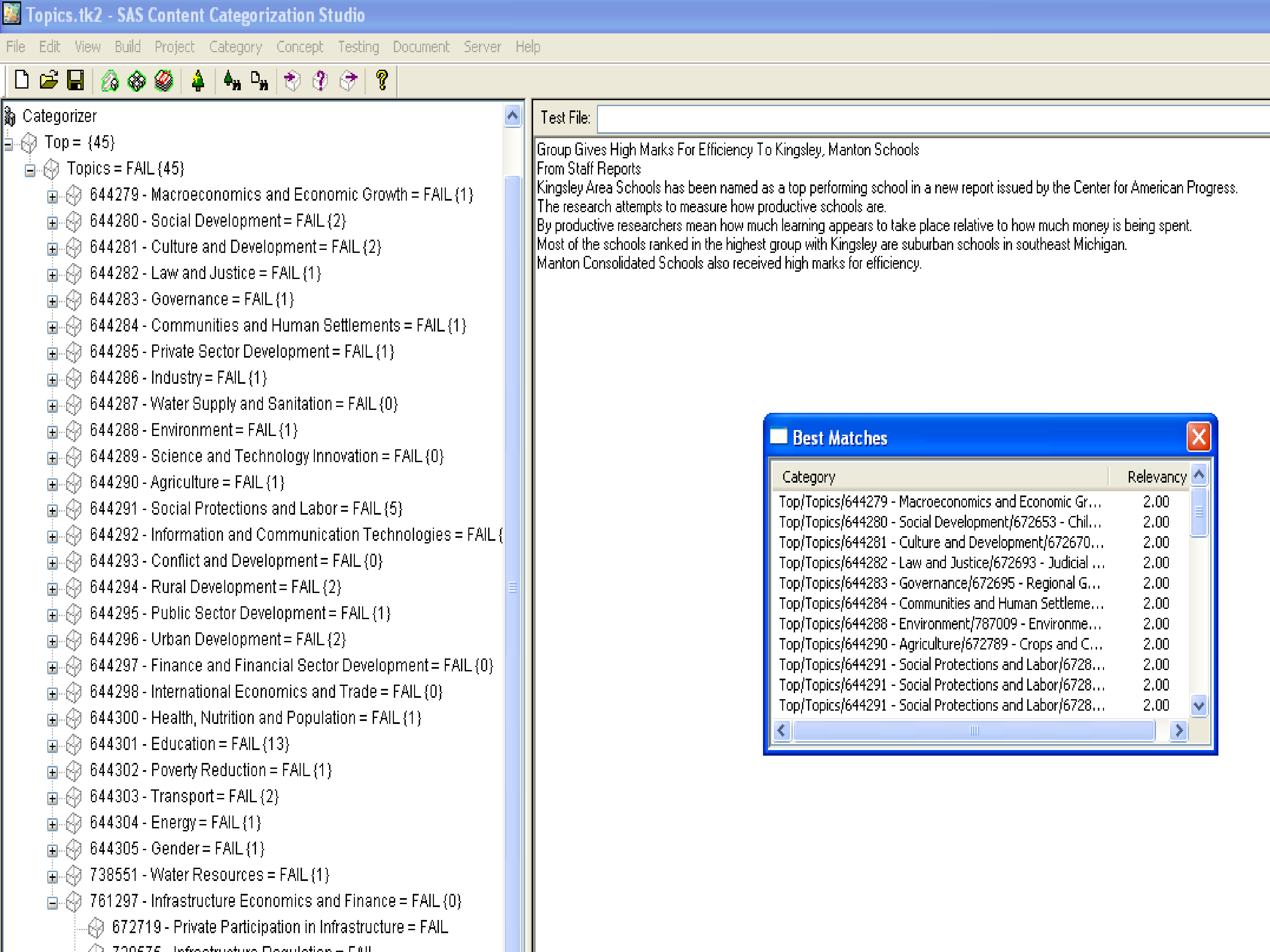
- Need to have the full text object in order to generate tags

- Following screen captures illustrate how this is accomplished



```
⊞ 644295 - Public Sector Development
⊞ 644296 - Urban Development
⊟ 644297 - Finance and Financial Sector
    1071216 - Debt Relief and HIPC
    672882 - Strategic Debt Manageme
    672883 - Payment Systems and Infr
    672884 - Banks and Banking Refor
    672885 - Insurance and Risk Mitiga
    672886 - Housing Finance
    672887 - Public and Municipal Finar
    672888 - Financial Crisis Managem
    672889 - Currencies and Exchange
    672890 - Microfinance
    672891 - Capital Markets and Capit
    672892 - Securities Markets Policy a
    672893 - Financial Intermediation
    672894 - Rural Finance
    672895 - Contractual Savings
    672896 - Financial Regulation and S
    672897 - Non Bank Financial Institu
    738594 - Anti-Money Laundering
    738596 - Concessional Finance and
    738597 - Islamic Finance
    738598 - E-Finance and E-Security
    787064 - Deposit Insurance
    787065 - Financial Sector and Socia
    787066 - Mutual Funds
    792182 - Bankruptcy and Resolution
    792183 - Law, Finance and Growth
    792184 - Financial Structures
    792185 - Finance and Development
```

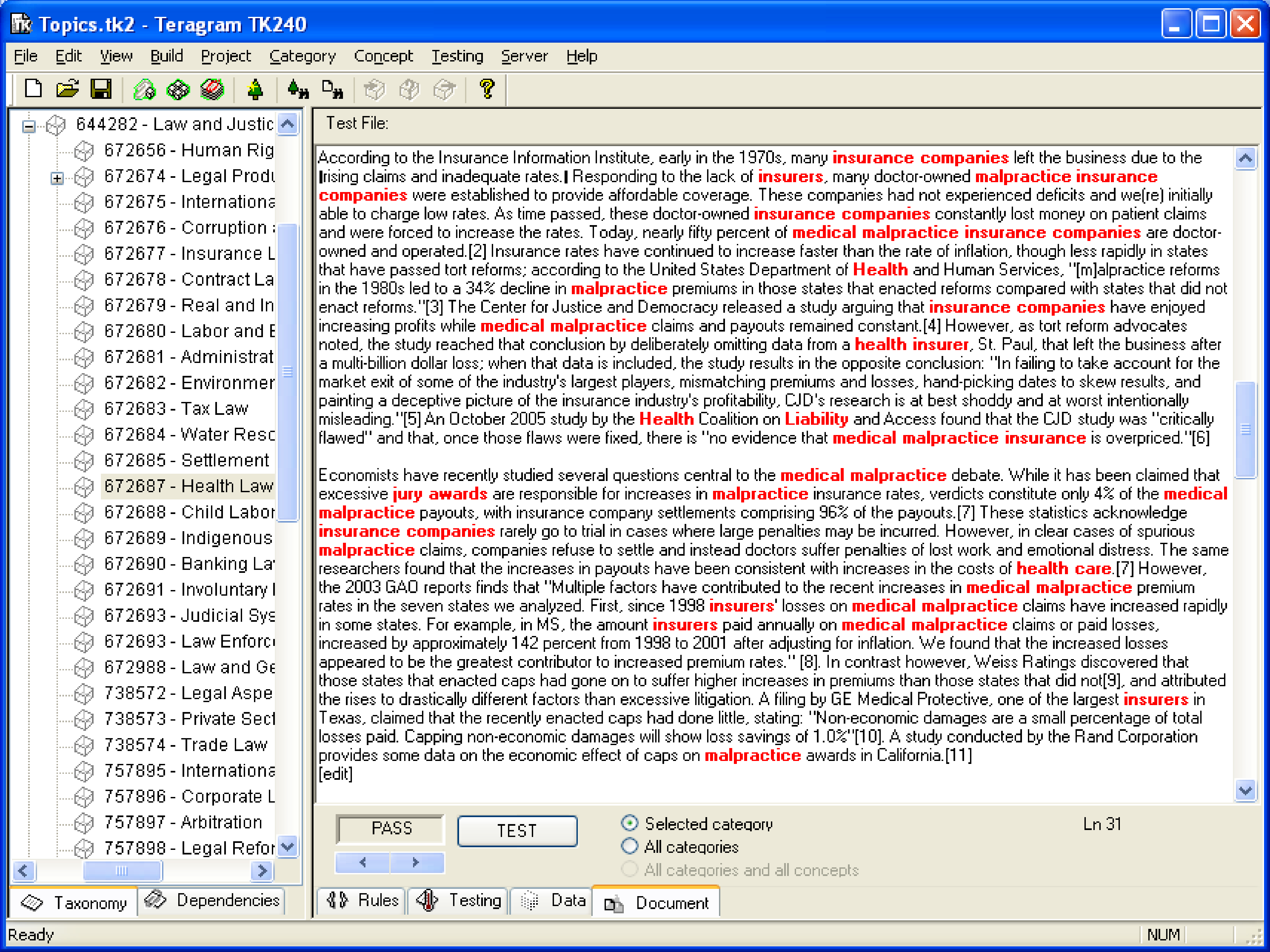
```
⊟ AND
    ⊟ OR
        "payment systems"
    ⊟ OR
        "Acceleration of Bank disbursements"
        "Accident insurance"
        "Account reconciliation"
        "Account settlement"
        "Accounting"
        "Accounting format"
        "Accounting systems"
        "ACH"
        "ACH network"
        "Acquirers"
        "Active refuse banks"
        "Additional principal payment"
        "Advisory netting"
        "Agency credit lines"
        "Agency relationships"
        "Amortization"
        "Analysis of buy decision"
        "Analysis of hold decision"
        "Analysis of sell decision"
        "APA"
        "Applied derivatives"
        "Appraisal fees"
        "APS"
        "Assured payment systems"
        "Asymmetric cryptography"
        "ATM"
        "Audit trails"
```


Topics.tk2 - SAS Content Categorization Studio
File Edit View Build Project Category Concept Testing Document Server Help
Test File:
Best Matches

Categorizer
Top = {45}
Topics = FAIL {45}
644279 - Macroeconomics and Economic Growth = FAIL {1}
644280 - Social Development = FAIL {2}
644281 - Culture and Development = FAIL {2}
644282 - Law and Justice = FAIL {1}
644283 - Governance = FAIL {1}
644284 - Communities and Human Settlements = FAIL {1}
644285 - Private Sector Development = FAIL {1}
644286 - Industry = FAIL {1}
644287 - Water Supply and Sanitation = FAIL {0}
644288 - Environment = FAIL {1}
644289 - Science and Technology Innovation = FAIL {0}
644290 - Agriculture = FAIL {1}
644291 - Social Protections and Labor = FAIL {5}
644292 - Information and Communication Technologies = FAIL {
644293 - Conflict and Development = FAIL {0}
644294 - Rural Development = FAIL {2}
644295 - Public Sector Development = FAIL {1}
644296 - Urban Development = FAIL {2}
644297 - Finance and Financial Sector Development = FAIL {0}
644298 - International Economics and Trade = FAIL {0}
644300 - Health, Nutrition and Population = FAIL {1}
644301 - Education = FAIL {13}
644302 - Poverty Reduction = FAIL {1}
644303 - Transport = FAIL {2}
644304 - Energy = FAIL {1}
644305 - Gender = FAIL {1}
738551 - Water Resources = FAIL {1}
761297 - Infrastructure Economics and Finance = FAIL {0}
672719 - Private Participation in Infrastructure = FAIL
720575 - Infrastructure Regulation = FAIL

Group Gives High Marks For Efficiency To Kingsley, Manton Schools
From Staff Reports
Kingsley Area Schools has been named as a top performing school in a new report issued by the Center for American Progress.
The research attempts to measure how productive schools are.
By productive researchers mean how much learning appears to take place relative to how much money is being spent.
Most of the schools ranked in the highest group with Kingsley are suburban schools in southeast Michigan.
Manton Consolidated Schools also received high marks for efficiency.

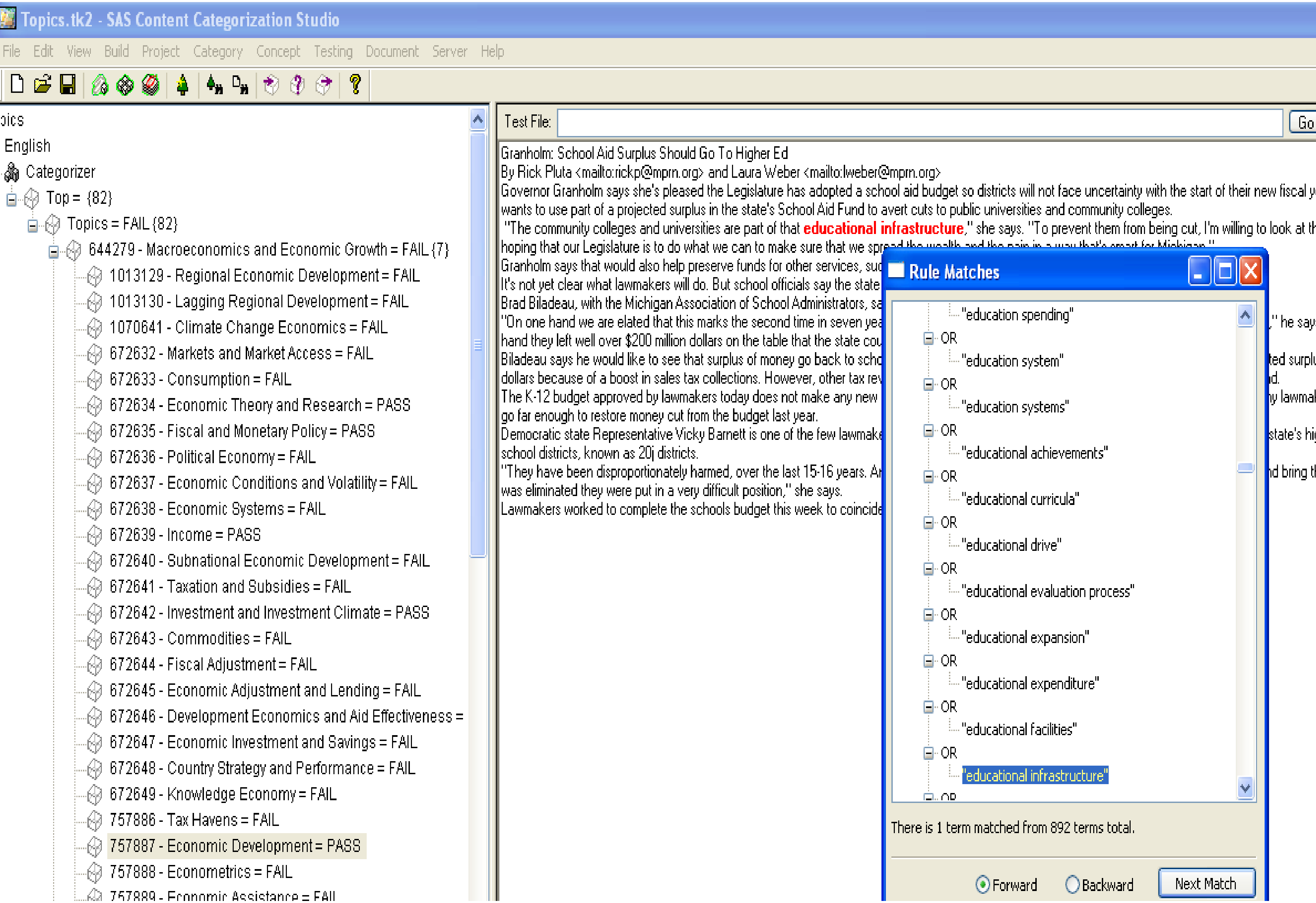
Category	Relevancy
Top/Topics/644279 - Macroeconomics and Economic Gr...	2.00
Top/Topics/644280 - Social Development/672653 - Chil...	2.00
Top/Topics/644281 - Culture and Development/672670...	2.00
Top/Topics/644282 - Law and Justice/672693 - Judicial ...	2.00
Top/Topics/644283 - Governance/672695 - Regional G...	2.00
Top/Topics/644284 - Communities and Human Settleme...	2.00
Top/Topics/644288 - Environment/787009 - Environme...	2.00
Top/Topics/644290 - Agriculture/672789 - Crops and C...	2.00
Top/Topics/644291 - Social Protections and Labor/6728...	2.00
Top/Topics/644291 - Social Protections and Labor/6728...	2.00
Top/Topics/644291 - Social Protections and Labor/6728...	2.00



According to the Insurance Information Institute, early in the 1970s, many **insurance companies** left the business due to the rising claims and inadequate rates. Responding to the lack of **insurers**, many doctor-owned **malpractice insurance companies** were established to provide affordable coverage. These companies had not experienced deficits and we(re) initially able to charge low rates. As time passed, these doctor-owned **insurance companies** constantly lost money on patient claims and were forced to increase the rates. Today, nearly fifty percent of **medical malpractice insurance companies** are doctor-owned and operated.[2] Insurance rates have continued to increase faster than the rate of inflation, though less rapidly in states that have passed tort reforms; according to the United States Department of **Health** and Human Services, "[m]alpractice reforms in the 1980s led to a 34% decline in **malpractice** premiums in those states that enacted reforms compared with states that did not enact reforms."[3] The Center for Justice and Democracy released a study arguing that **insurance companies** have enjoyed increasing profits while **medical malpractice** claims and payouts remained constant.[4] However, as tort reform advocates noted, the study reached that conclusion by deliberately omitting data from a **health insurer**, St. Paul, that left the business after a multi-billion dollar loss; when that data is included, the study results in the opposite conclusion: "In failing to take account for the market exit of some of the industry's largest players, mismatching premiums and losses, hand-picking dates to skew results, and painting a deceptive picture of the insurance industry's profitability, CJD's research is at best shoddy and at worst intentionally misleading."[5] An October 2005 study by the **Health** Coalition on **Liability** and Access found that the CJD study was "critically flawed" and that, once those flaws were fixed, there is "no evidence that **medical malpractice insurance** is overpriced."[6]

Economists have recently studied several questions central to the **medical malpractice** debate. While it has been claimed that excessive **jury awards** are responsible for increases in **malpractice** insurance rates, verdicts constitute only 4% of the **medical malpractice** payouts, with insurance company settlements comprising 96% of the payouts.[7] These statistics acknowledge **insurance companies** rarely go to trial in cases where large penalties may be incurred. However, in clear cases of spurious **malpractice** claims, companies refuse to settle and instead doctors suffer penalties of lost work and emotional distress. The same researchers found that the increases in payouts have been consistent with increases in the costs of **health care**.[7] However, the 2003 GAO reports finds that "Multiple factors have contributed to the recent increases in **medical malpractice** premium rates in the seven states we analyzed. First, since 1998 **insurers**' losses on **medical malpractice** claims have increased rapidly in some states. For example, in MS, the amount **insurers** paid annually on **medical malpractice** claims or paid losses, increased by approximately 142 percent from 1998 to 2001 after adjusting for inflation. We found that the increased losses appeared to be the greatest contributor to increased premium rates." [8]. In contrast however, Weiss Ratings discovered that those states that enacted caps had gone on to suffer higher increases in premiums than those states that did not[9], and attributed the rises to drastically different factors than excessive litigation. A filing by GE Medical Protective, one of the largest **insurers** in Texas, claimed that the recently enacted caps had done little, stating: "Non-economic damages are a small percentage of total losses paid. Capping non-economic damages will show loss savings of 1.0%"[10]. A study conducted by the Rand Corporation provides some data on the economic effect of caps on **malpractice** awards in California.[11]
[edit]

● Selected category
○ All categories
○ All categories and all concepts
Ln 31


Topics.tk2 - SAS Content Categorization Studio
File Edit View Build Project Category Concept Testing Document Server Help
Test File:
Go
Rule Matches
Forward
Backward
Next Match

pics
English
Categorizer
Top = {82}
Topics = FAIL {82}
644279 - Macroeconomics and Economic Growth = FAIL {7}
1013129 - Regional Economic Development = FAIL
1013130 - Lagging Regional Development = FAIL
1070641 - Climate Change Economics = FAIL
672632 - Markets and Market Access = FAIL
672633 - Consumption = FAIL
672634 - Economic Theory and Research = PASS
672635 - Fiscal and Monetary Policy = PASS
672636 - Political Economy = FAIL
672637 - Economic Conditions and Volatility = FAIL
672638 - Economic Systems = FAIL
672639 - Income = PASS
672640 - Subnational Economic Development = FAIL
672641 - Taxation and Subsidies = FAIL
672642 - Investment and Investment Climate = PASS
672643 - Commodities = FAIL
672644 - Fiscal Adjustment = FAIL
672645 - Economic Adjustment and Lending = FAIL
672646 - Development Economics and Aid Effectiveness =
672647 - Economic Investment and Savings = FAIL
672648 - Country Strategy and Performance = FAIL
672649 - Knowledge Economy = FAIL
757886 - Tax Havens = FAIL
757887 - Economic Development = PASS
757888 - Econometrics = FAIL
757889 - Economic Assistance = FAIL

Granholm: School Aid Surplus Should Go To Higher Ed
By Rick Pluta <mailto:rickp@mprn.org> and Laura Weber <mailto:lweber@mprn.org>
Governor Granholm says she's pleased the Legislature has adopted a school aid budget so districts will not face uncertainty with the start of their new fiscal y
wants to use part of a projected surplus in the state's School Aid Fund to avert cuts to public universities and community colleges.
"The community colleges and universities are part of that **educational infrastructure**," she says. "To prevent them from being cut, I'm willing to look at th
hoping that our Legislature is to do what we can to make sure that we spread the wealth and the pain in a way that's smart for Michigan."
Granholm says that would also help preserve funds for other services, suc
It's not yet clear what lawmakers will do. But school officials say the state
Brad Biladeau, with the Michigan Association of School Administrators, sa
"On one hand we are elated that this marks the second time in seven yea ," he say
hand they left well over $200 million dollars on the table that the state cou
Biladeau says he would like to see that surplus of money go back to scho ted surplu
dollars because of a boost in sales tax collections. However, other tax rev d.
The K-12 budget approved by lawmakers today does not make any new ny lawmal
go far enough to restore money cut from the budget last year.
Democratic state Representative Vicky Barnett is one of the few lawmake state's hig
school districts, known as 20j districts.
"They have been disproportionately harmed, over the last 15-16 years. Ar nd bring th
was eliminated they were put in a very difficult position," she says.
Lawmakers worked to complete the schools budget this week to coincide

"education spending"
OR
"education system"
OR
"education systems"
OR
"educational achievements"
OR
"educational curricula"
OR
"educational drive"
OR
"educational evaluation process"
OR
"educational expansion"
OR
"educational expenditure"
OR
"educational facilities"
OR
"educational infrastructure"
OR

There is 1 term matched from 892 terms total.

○ Forward ○ Backward

Two Research Challenges

- Two challenges have slowed the pace of our research:
  - In scholarly context, users often use tags to mark pointers to references or bibliographic records, rather than the content itself – a separate search has to be conducted about 40% of the time to find the original content
  - At this time, we have a few hundred examples – the labor intensive nature of retrieving the content and running each example takes more time than we anticipated

Preliminary Results

- Research is still in progress due to the challenge of collecting both original source materials and tagged metadata

1. 90% of the time, the semantic engine when powered by a KOS will promote the core topical term

2. The semantic engine, when powered only by a topic-focused KOS, will promote 45% of all the terms suggested by end users.
   - *The remaining percentages largely derived from other types of KOS which were not initially included in the research. We are updating the methodology. This rate can be improved by leveraging other types of KOS.*

3. Semantic engine will generate anywhere between 1.5 to 10 times as many topical tags as are suggested by single users – possibility of generating a tag cloud

Observations and Lessons Learned

- Semantic Density of the Content
  - Number of tags semantically generated varies with the density of the content tagged – sparse content likely to generate fewer tags, dense content generates more tags.
  - User tagging does not seem to vary with the density of the content but with the popularity or difficulty of finding the content.

- Nature of the Vocabulary
  - Number of tags generated also varies with the nature of the topical vocabulary – where the vocabulary is weak or thin, few tags may be semantically generated
  - Where the vocabulary is rich and stable (e.g., the subject domain is stable) more tags are likely to be semantically generated
  - Where the vocabulary is dynamic and broad (e.g., the subject domain is emerging or fragmented) the number of tags semantically generated is expected to be a bit more unpredictable (dependent upon the currency and coverage of the KOS)

Observations and Lessons Learned

- There is a strong mix of descriptors and identifiers in the tags – we need multiple KOS and different semantic profile types to increase our coverage rates

- Faceting of tags appears to have some relevance to the locality, familiarity and popularity of the content.
  - Content with a local flavor is more likely to have tags with names of people, organizations, geographical entities, etc.
  - Content which is current or more popular culture in nature appears to have more faceting, and is also more prone to redundant values

- Tagging of content with an academic topic focus appears to be quite different in behavior from tagging of popular culture or news media content suggesting a common mental model of indexers and users

Observations and Lessons Learned

- User tagging appears to serve different purposes across subject domains – these differences may reflect the nature of the literature
  - Agriculture is tagged to provide more granular access
  - Transportation is tagged to “locate” scarce resources (a challenging information domain)
  - Education and Health seem to be tagged for personal collection building
  - Environment content appears to follow no single pattern at this time

Work in Progress

- Complete the testing of the full data set of 1,500 content objects

- Each sample will be sufficiently rigorous to draw reliable conclusions

- Complete a second pass of the data set with additional semantic profiles – People KOS, Geographical KOS, Organizations KOS, Event KOS

- Undertake and complete the end user review and selection testing

Questions and Discussions….

THANK YOU!