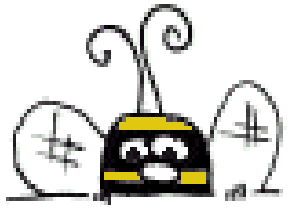


NewsIndexer:
Taxonomy-Based News Categorization

Alice Redmond-Neal

Access Innovations, Inc.

SLA Annual Meeting

June 12, 2006

About Access Innovations, Inc.

- Established 1978
- KM World's 100 companies that matter, 2005
- Services
  - Database design, construction, maintenance
  - Taxonomy construction
  - Categorization rule development
  - XML, SGML markup
  - Abstracting and indexing
- Data Harmony software
  - Thesaurus Master™
  - M.A.I.™ (Machine Aided Indexer)
  - XIS™ (XML Intranet System)

NewsIndexer = one implementation of Data Harmony

Access Innovations, Inc.

DATA HARMONY

Access Innovations' focus

- ***Access*** to information through effective storage and retrieval

HOW?

- Organize topics within domain
- Apply meta-data systematically
  - Subject meta-data from controlled vocabulary, i.e. *TAXONOMIES*
  - Other meta-data specific to project
- Use indexing rules to govern categorization

Lessons we've learned

Well-established principles and practices in information storage and retrieve are still current, valid, reliable.

- To retrieve by topic, use controlled vocabularies (ideally thesauri).
- To retrieve by other document features, structure content using logical fields.
  - Document type, delivery channel, author…
  - Faceted taxonomy is another approach
- Rely on standards, e.g. ANSI/NISO Z39.19-2005
- Human editors are still necessary for best taxonomy development and precise indexing.

Taxonomy construction strategy at Access Innovations

- Interview client for needs and perspective
- Establish scope – primary and peripheral subject areas
- Build or buy decision
- Build conceptual skeleton of categories
  - Authoritative overview of domain – top down approach
  - Review actual content – bottom up approach
- Expand to reflect granularity as needed
- Identify synonyms, duplications; disambiguate homographs
- Link related terms
- Gap analysis, fill in terms for missing concepts

Access Innovations, Inc.

DATA HA

Test, edit, test, implement, maintain

Indexing strategy at Access Innovations

- Combine best features of
  - Human knowledge, interpretation, analysis
    - Focus on indexing specificity, precision, exhaustivity
  - Machine-enabled efficiency
    - Speed
    - Consistency
    - Breadth
    - Specificity
- Apply indexing terms as permanent meta-data
- Result – Relevance and Precision *with* Speed and Consistency

Taxonomies and meta-data

- Taxonomies

  describe what the content is about – the basis for subject meta-data

- Meta-data

  provide a handle to capture that description for a website or sorted in a database

Why use taxonomies?

- Organization
  - Bring order to chaos
  - Navigable hierarchy structure
- Vocabulary control
  - Link synonyms
  - Disambiguate homographs
  - No need to guess writer's word
- Added information through thesaurus relationships
  - Provide context for a term/concept
  - Suggest related topics

Value of taxonomy terms as meta-data

- Bring precision to retrieval phase
- Searchable as meta-data
  - Gives more *precise* results than full text search – for known topic
  - Prevents hit on random occurrence of your query word – noise/false drops
  - Increases productivity
- Boost findability for a positive user experience

Taxonomy descriptors become subject meta-data

- Descriptors are selected and XML meta-tagged
- Subject meta-data are stored with document
  - Meta-data are persistent, reliable
  - Concept boundaries don't shift with new data
- Descriptors available as database or webpage meta-data
- XML metatags are portable, not captive in one software program

Ahh, the challenge of managing text...

- Avalanche of data, needle in haystack ...
- Challenges for *NEWS*
  - Current content
    - Huge volume, fast flow
    - Fast semi-obsolescence, quick to archive
    - Largely region-specific
    - Familiar current vocabulary
  - Archived content
    - Super-huge volume
    - Largely region-specific
    - Unfamiliar and variable vocabulary over years

News searchers' needs

- Internal – writers, editors, news librarians
  - Retrieve specific item, i.e. known content
  - Troll for general information, background on specific topic
  - Time limits – on deadline!
- External – public users
  - Browse and explore general topics
  - Recover specific fact or information
  - Satisfice = satisfy + suffice = good enough
- No specialized, accepted terminology

Best practices for search in Information Architecture

Provide multiple retrieval methods

- Organized and browsable taxonomy categories
- Facilitated search – recognize concept of taxonomy term even if query didn't use the term
- Full text search for unique known queries

Taxonomies support better search

- Organized and browsable categories
- Stable meta-data, persistent categorization
- Targeted search with priority on meta-data
- Enhanced term/concept information (broader/narrower terms, synonyms, related concepts, etc.)

- Not limited to precise taxonomy term for query
- Faceted taxonomies ~ field formatted data

(Additional ways to describe an item for better identification and retrieval)

Value of a browsable taxonomy

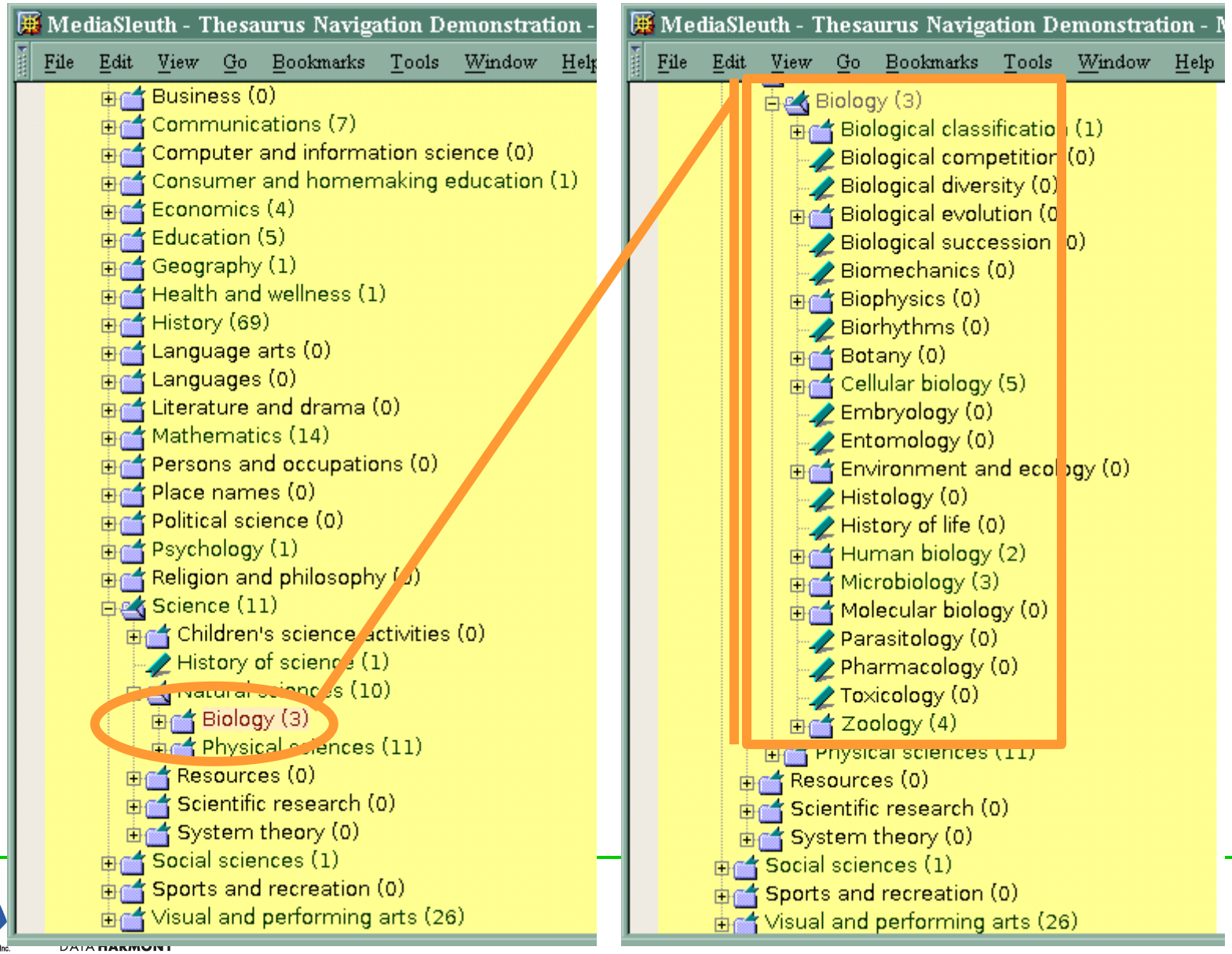
- Searchers find info 50% faster using browsable categories than using list returned from free text search
  - Results even stronger when results are not in top 20 returns
- Searchers prefer browsable category search

Chen, H., Dumais, S., *Bringing order to the web: automatically categorizing search results*. Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'00), ACM (2000) 145-152.

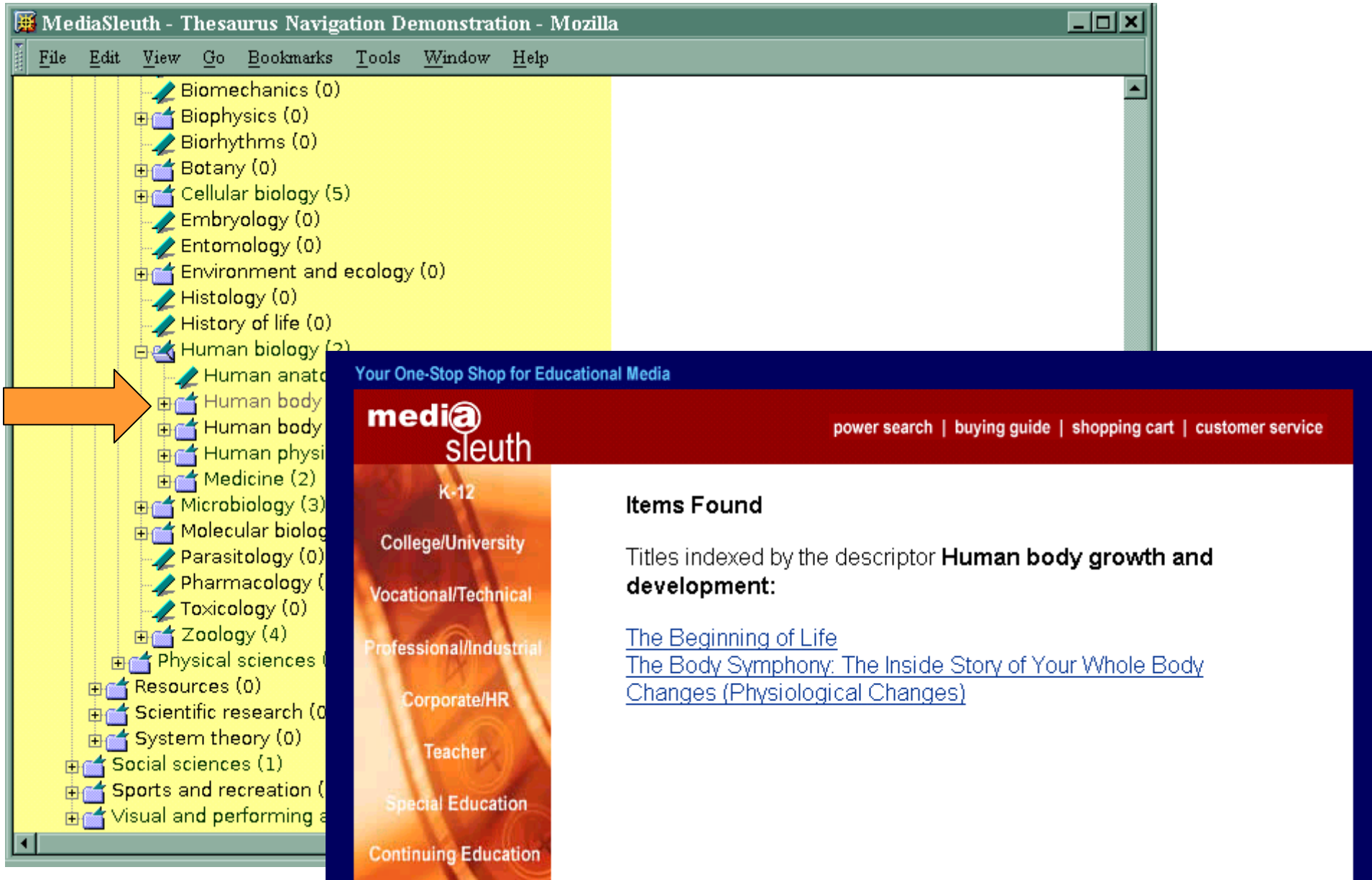
Browsable taxonomy view

- Direct route to precise content topics
- NavTree™ (part of NewsIndexer's underlying software) connects to your CMS

Browsable NavTree displays taxo categories



Select taxonomy category to see associated content

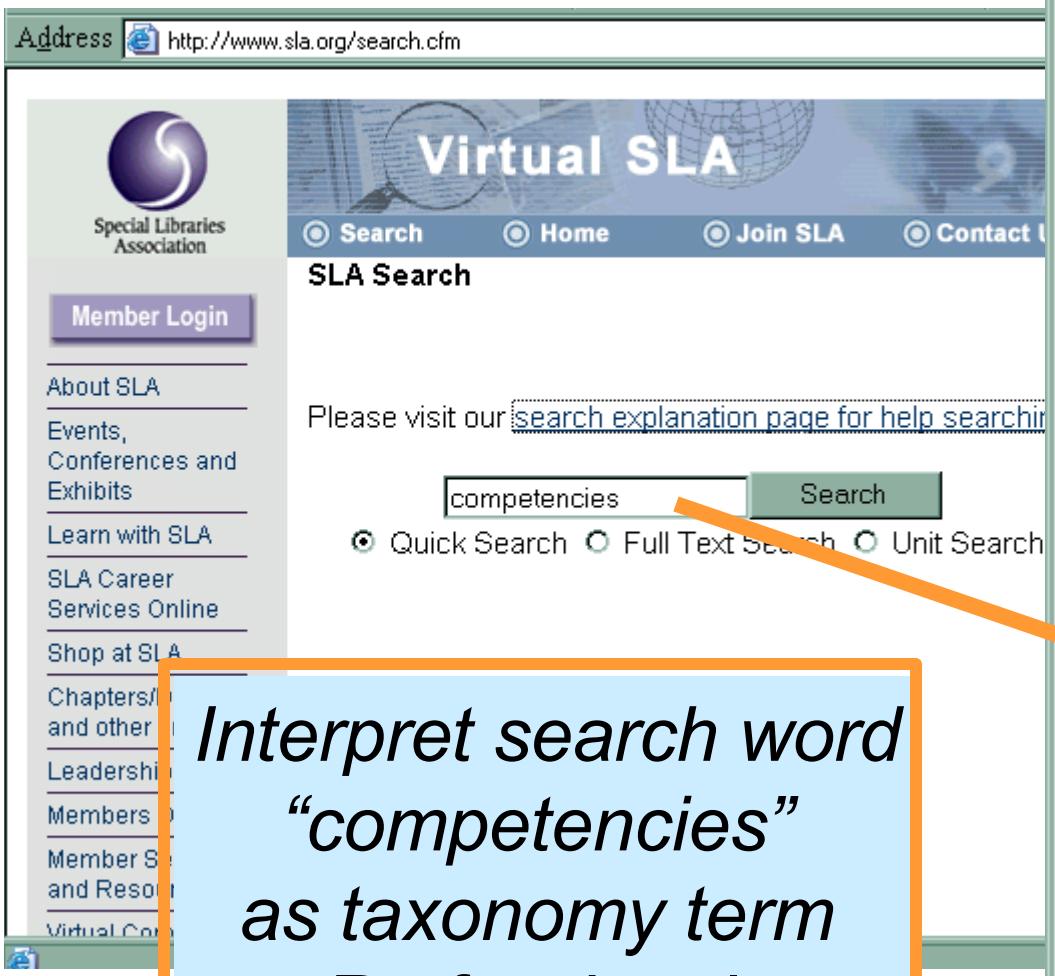


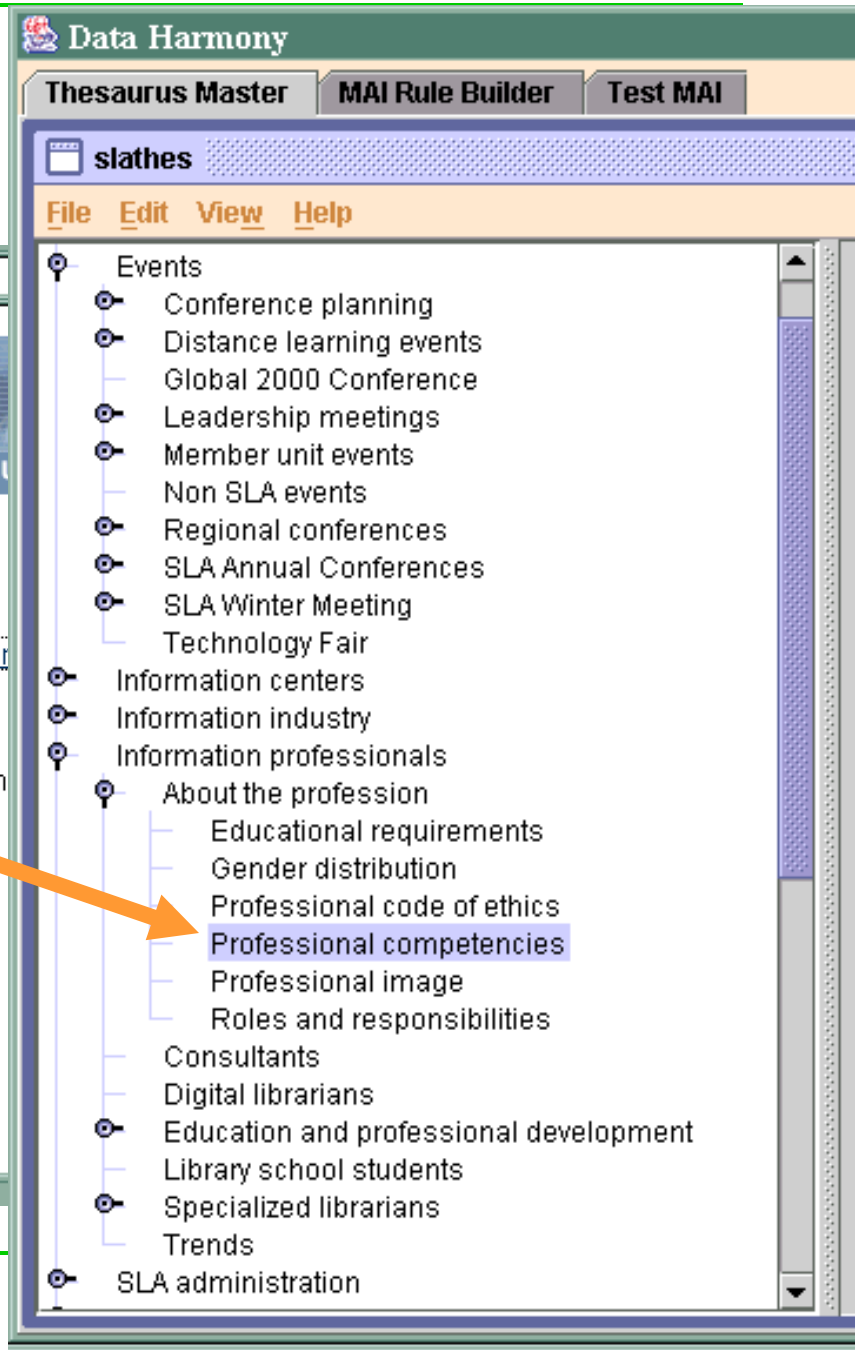
Taxonomy terms as meta-data facilitate full text search

- Link query words with taxonomy term or synonym
  - germs → Microbes
  - vaccin* → Pharmaceutical drugs
- Search can check meta-data first, then full text
- Retrieves content meta-tagged with taxonomy terms

Recognizing term equivalents enhances search

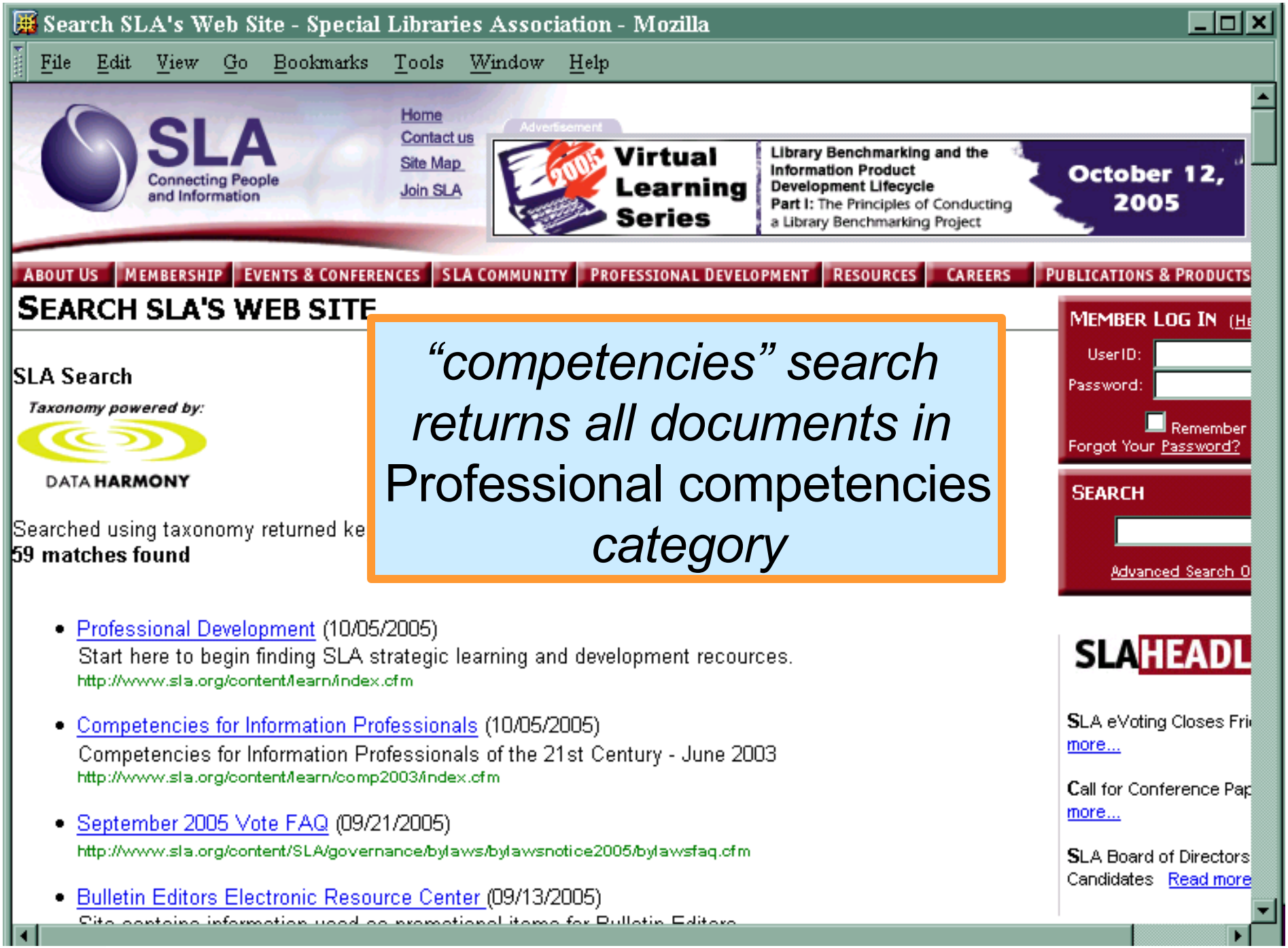
Taxonomy terms support query expansion

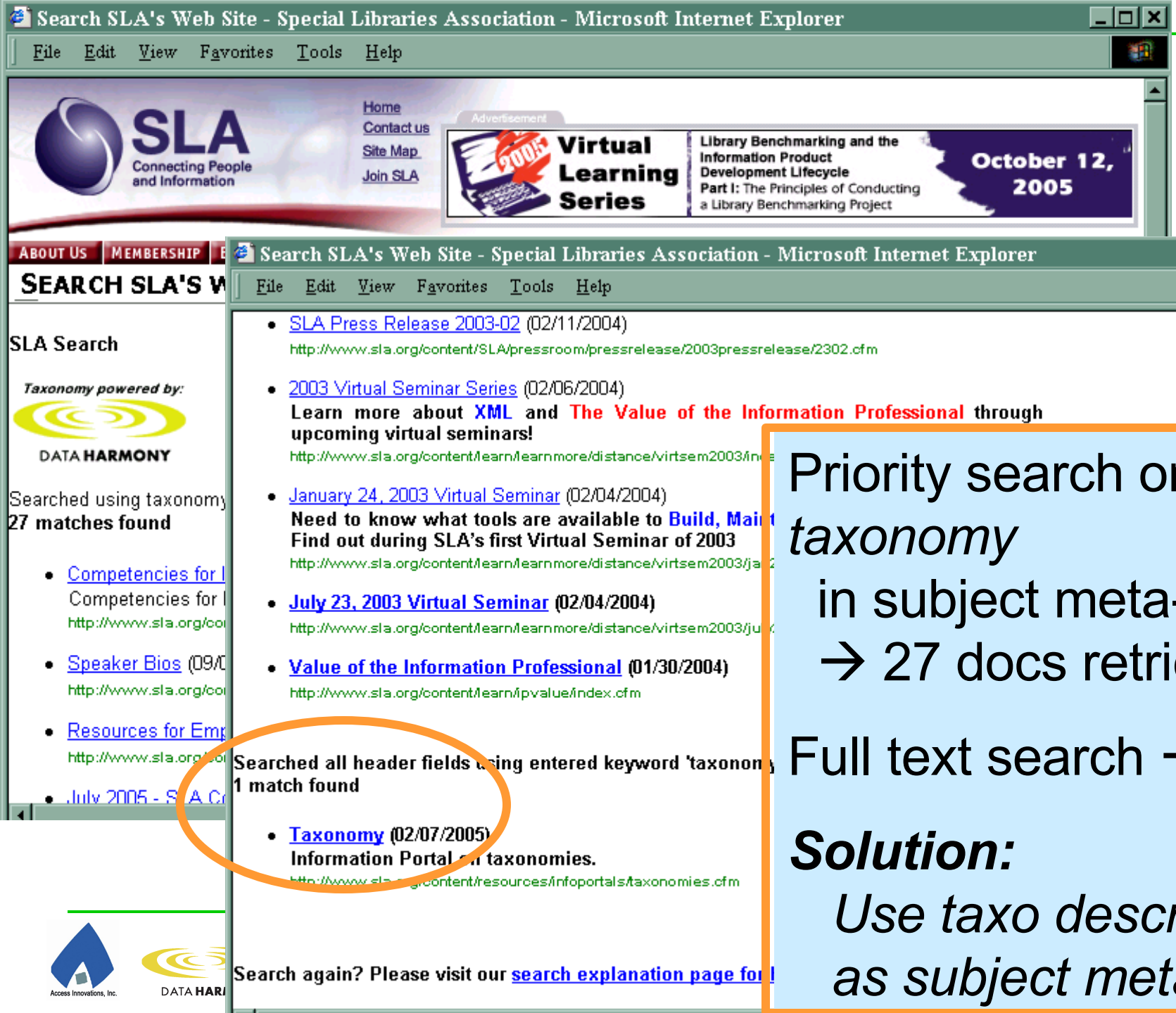




Interpret search word "competencies" as taxonomy term Professional competencies

Access Innovations, Inc.


Search SLA's Web Site - Special Libraries Association - Mozilla
File Edit View Go Bookmarks Tools Window Help
SLA
Connecting People and Information
Home
Contact us
Site Map
Join SLA
Advertisement
2005 Virtual Learning Series
Library Benchmarking and the Information Product Development Lifecycle
Part I: The Principles of Conducting a Library Benchmarking Project
October 12, 2005
ABOUT US MEMBERSHIP EVENTS & CONFERENCES SLA COMMUNITY PROFESSIONAL DEVELOPMENT RESOURCES CAREERS PUBLICATIONS & PRODUCTS
SEARCH SLA'S WEB SITE
SLA Search
Taxonomy powered by:
DATA HARMONY
Searched using taxonomy returned ke
59 matches found
"competencies" search returns all documents in Professional competencies category
MEMBER LOG IN
UserID:
Password:
Remember
Forgot Your Password?
SEARCH
Advanced Search
Professional Development (10/05/2005)
Start here to begin finding SLA strategic learning and development recources.
http://www.sla.org/content/learn/index.cfm
Competencies for Information Professionals (10/05/2005)
Competencies for Information Professionals of the 21st Century - June 2003
http://www.sla.org/content/learn/comp2003/index.cfm
September 2005 Vote FAQ (09/21/2005)
http://www.sla.org/content/SLA/governance/bylaws/bylawsnotice2005/bylawsfaq.cfm
Bulletin Editors Electronic Resource Center (09/13/2005)
SLAHEADL
SLA eVoting Closes Fri
more...
Call for Conference Pap
more...
SLA Board of Directors
Candidates Read more



Priority search on *taxonomy*
in subject meta-data
→ 27 docs retrieved

Full text search → 1

Solution:
Use taxo descriptors as subject meta-data

Search using query expansion

- Connects search query to possible taxonomy categories

  "bug" → Insects? Microbes? Software app's?

- Searcher sees results for each category, focuses on interest area
- MAI Query™ – an extension of M.A.I. underlying NewsIndexer

Leverage categorization rules to improve search results

Unconditional full text search … or The Joy of Goooooogling

- Good for new words, vernacular, names
- De facto standard
- Meets minimum user expectations
  - *Satisficing*
- Better than nothing, good when all else fails

BUT

- Searchers increasingly savvy, demanding
- Better results elsewhere raise expectations for precision
- Power searchers don't have time to review mixed results

Use information in taxonomy and indexing rules to support search

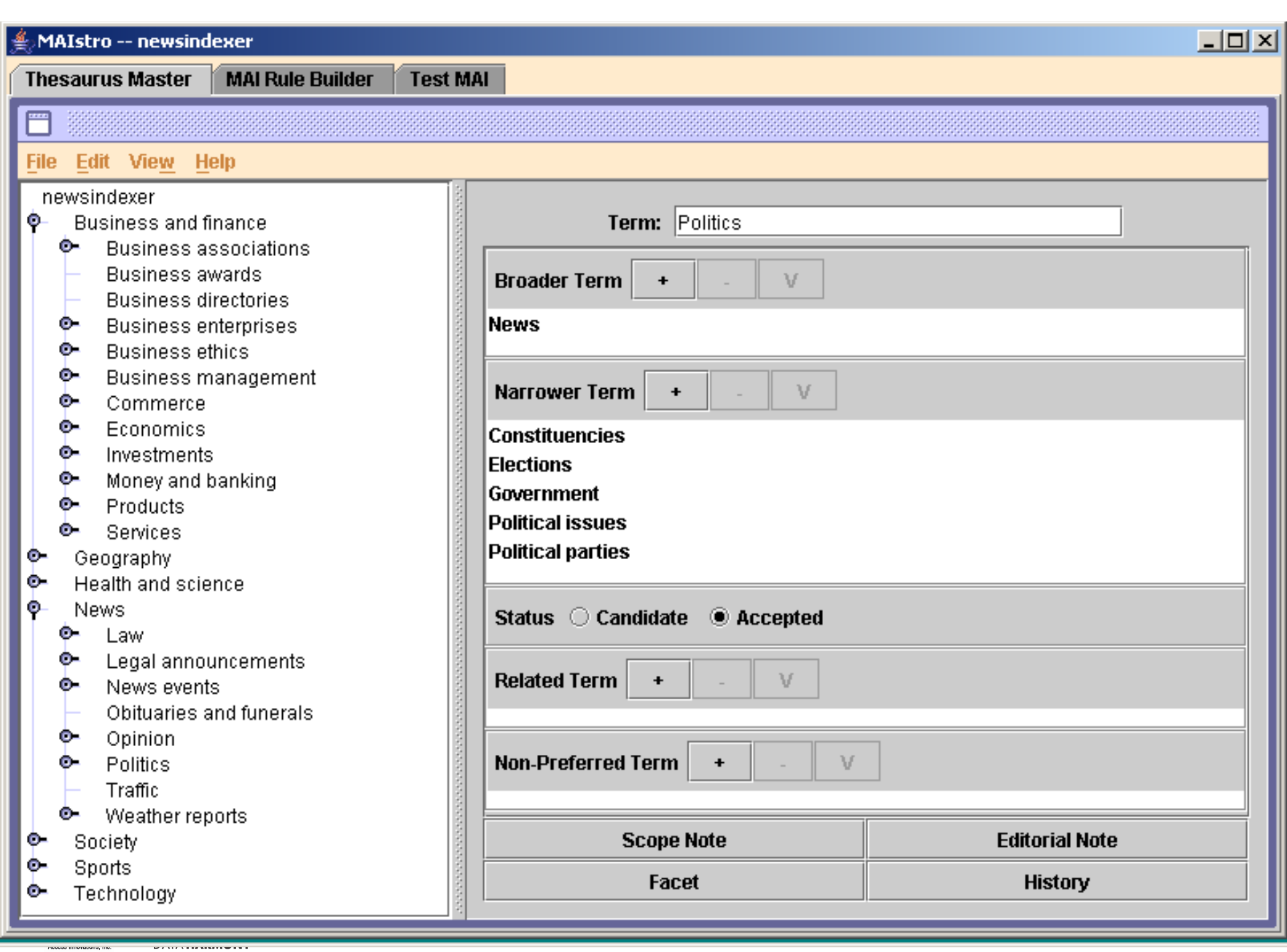
- Thesaurus relationships expand searcher's awareness (see *Ask.com)*
  - Broader/narrower concepts
  - Related concepts
  - Synonyms
- Categorization rules capture indexer's analysis, knowledge, interpretation of text
  - Identify context cues that affect interpretation
  - Provide default indexing if one option does not qualify

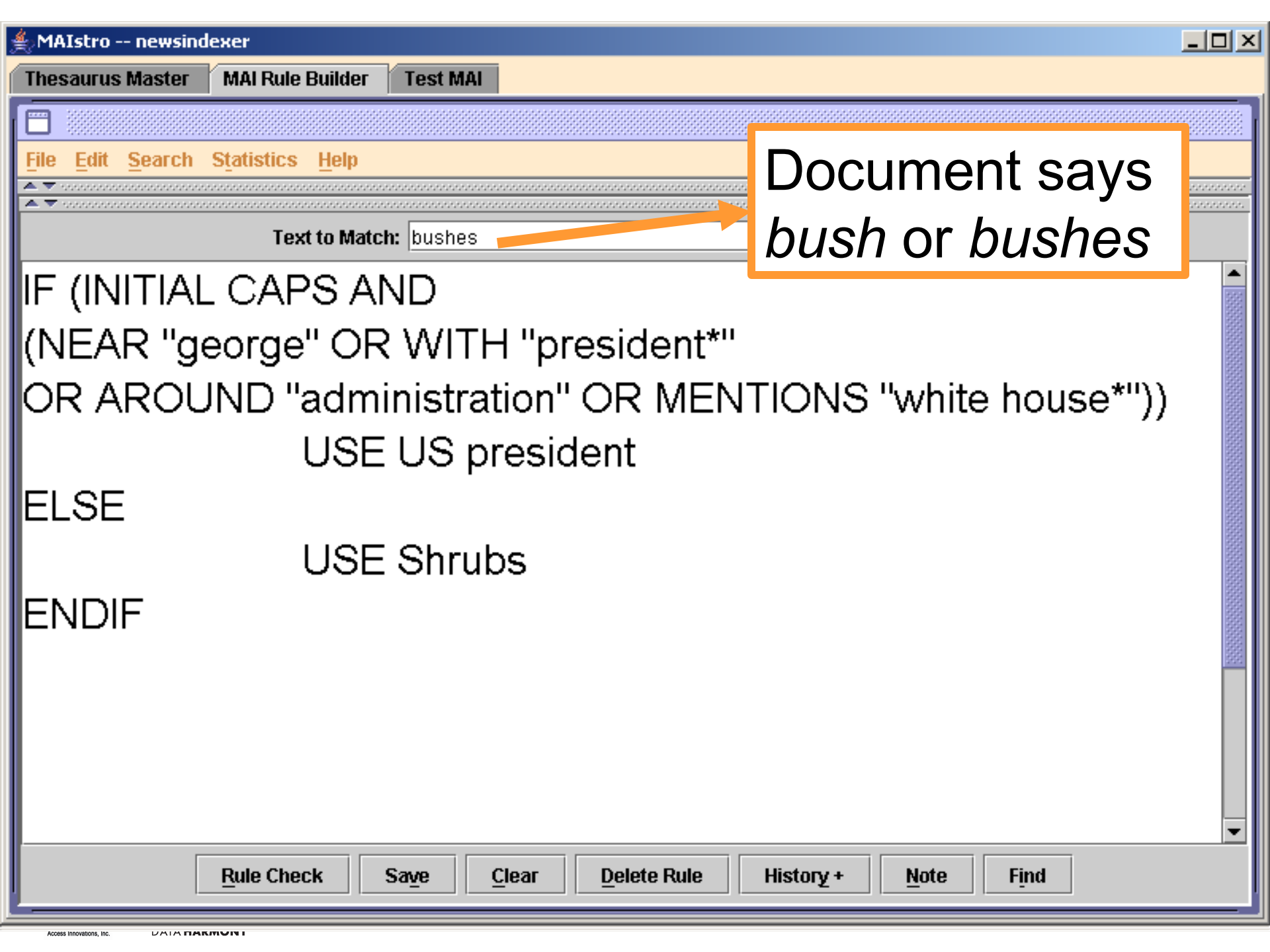
About NewsIndexer

- Taxonomy-based news categorization system
- Special implementation of Thesaurus Master and M.A.I. (taxonomy and indexing tools)
- Used with *your* information content

NewsIndexer's core

- Taxonomy built for the news industry
  - 5,000 concept terms
  - 34,000 geographic terms
  - 1,000,000 organization terms (optional)
  - Flexible, customizable by editors
- Rules for using the taxonomy terms to categorize or filter content
  - Suggest terms based on context
  - Transparent – not a black box
  - Easily finetuned by editors to increase precision


MAIstro -- newsindexer
Thesaurus Master
MAI Rule Builder
Test MAI
File
Edit
View
Help
newsindexer
Business and finance
Business associations
Business awards
Business directories
Business enterprises
Business ethics
Business management
Commerce
Economics
Investments
Money and banking
Products
Services
Geography
Health and science
News
Law
Legal announcements
News events
Obituaries and funerals
Opinion
Politics
Traffic
Weather reports
Society
Sports
Technology
Term: Politics
Broader Term
+
-
V
News
Narrower Term
+
-
V
Constituencies
Elections
Government
Political issues
Political parties
Status
Candidate
Accepted
Related Term
+
-
V
Non-Preferred Term
+
-
V
Scope Note
Editorial Note
Facet
History


MAIstro -- newsindexer
Thesaurus Master
MAI Rule Builder
Test MAI
File Edit Search Statistics Help
Text to Match: bushes
Document says *bush* or *bushes*
IF (INITIAL CAPS AND
(NEAR "george" OR WITH "president*"
OR AROUND "administration" OR MENTIONS "white house*"))
USE US president
ELSE
USE Shrubs
ENDIF
Rule Check
Save
Clear
Delete Rule
History +
Note
Find

NewsIndexer's taxonomy *is:*

- Geared to the news industry
- Easily customized
- Simple to add or import terms
  - Basic categorization rules added automatically
- Unlimited in size
- Unlimited in number of hierarchy levels
- Unlimited in full thesaurus relationships
- Viewable in multiple display formats
- Stored in XML

NewsIndexer *is not:*

- *Not* a static vocabulary
- *Not* opaque to taxonomists and indexers
- *Not* difficult to enhance precision
- *Not* time-consuming to improve categorization accuracy
- *Not* dependent on document training sets

How does NewsIndexer work?

- Scans article
- Identifies text words that trigger categorization rules
- Suggests taxonomy terms that meet conditions set in rule

- Editor review mode – editor reviews term suggestions, selects/rejects, adds other terms from the taxo as needed
- Auto-categorization mode – accept term suggestions, all or top n suggestions

NewsIndexer Demonstration

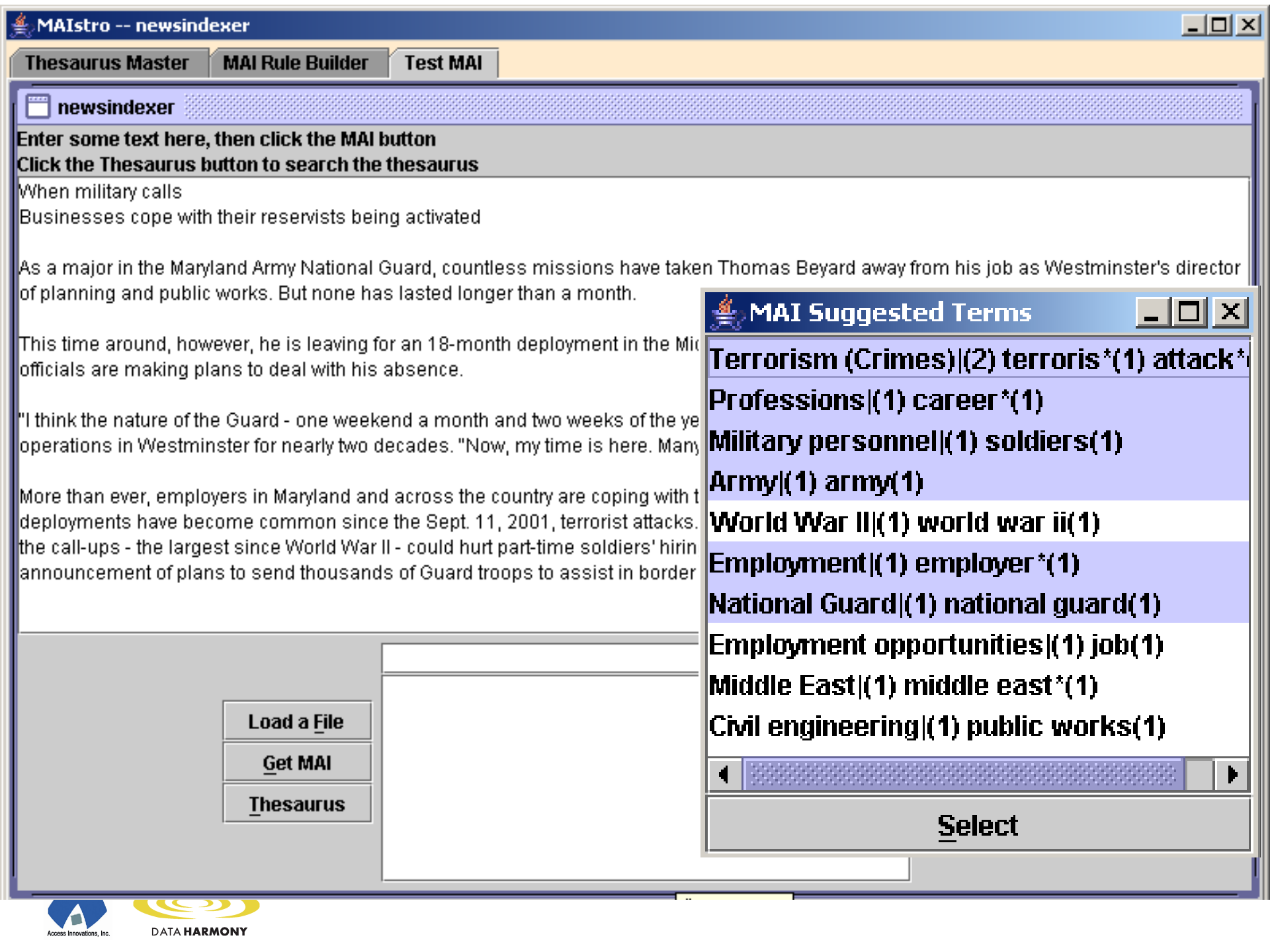
The index terms have been sorted by frequency (how often your text invoked the term).

To see what a rule looks like, click on the link.

When military calls Businesses cope with their reservists being activated As a major in the Maryland Army National Guard, countless missions have taken Thomas Beyard away from his job as Westminster's director of planning and public works. But none has lasted longer than a month. This time around, however, he is leaving for an 18-month deployment in the Middle East. While Beyard prepares to leave next month, city officials are making plans to deal with his absence. I think the nature of the Guard - one weekend a month and two weeks of the year - it's changing, said Beyard, who has been part of the daily operations in Westminster for nearly two decades. Now, my time is here. Many other people have had their time, second and third. More than ever, employers in Maryland and across the country are coping with their workers' extended military leave as longer and multiple deployments have become common since the Sept. 11, 2001, terrorist attacks. Now, some Guard advocates are worried that the frequency of the call-ups - the largest since World War II - could hurt part-time soldiers' hiring chances or their careers. And President Bush's recent announcement of plans to send

- ☐ Terrorism (Crimes) (2)
- ☐ Professions (1)
- ☐ Military personnel (1)
- ☐ Army (1)
- ☐ World War II (1)
- ☐ Employment (1)
- ☐ National Guard (1)
- ☐ Employment opportunities (1)
- ☐ Middle East (1)
- ☐ Civil engineering (1)

Let us know what you think. Check the terms above that you think are good, and enter below what you think we missed.
Need another index term? Click here to search for more.


MAIstro -- newsindexer
Thesaurus Master
MAI Rule Builder
Test MAI
newsindexer
Enter some text here, then click the MAI button
Click the Thesaurus button to search the thesaurus
When military calls
Businesses cope with their reservists being activated
As a major in the Maryland Army National Guard, countless missions have taken Thomas Beyard away from his job as Westminster's director of planning and public works. But none has lasted longer than a month.
This time around, however, he is leaving for an 18-month deployment in the Mi
officials are making plans to deal with his absence.
"I think the nature of the Guard - one weekend a month and two weeks of the ye
operations in Westminster for nearly two decades. "Now, my time is here. Many
More than ever, employers in Maryland and across the country are coping with t
deployments have become common since the Sept. 11, 2001, terrorist attacks.
the call-ups - the largest since World War II - could hurt part-time soldiers' hirin
announcement of plans to send thousands of Guard troops to assist in border
Load a File
Get MAI
Thesaurus
MAI Suggested Terms
Terrorism (Crimes)|(2) terroris*(1) attack*
Professions|(1) career*(1)
Military personnel|(1) soldiers(1)
Army|(1) army(1)
World War II|(1) world war ii(1)
Employment|(1) employer*(1)
National Guard|(1) national guard(1)
Employment opportunities|(1) job(1)
Middle East|(1) middle east*(1)
Civil engineering|(1) public works(1)
Select

NewsIndexer results

Content is categorized

- Accurately
- Quickly
- Consistently
- Permanently

What about IPTC NewsCodes?

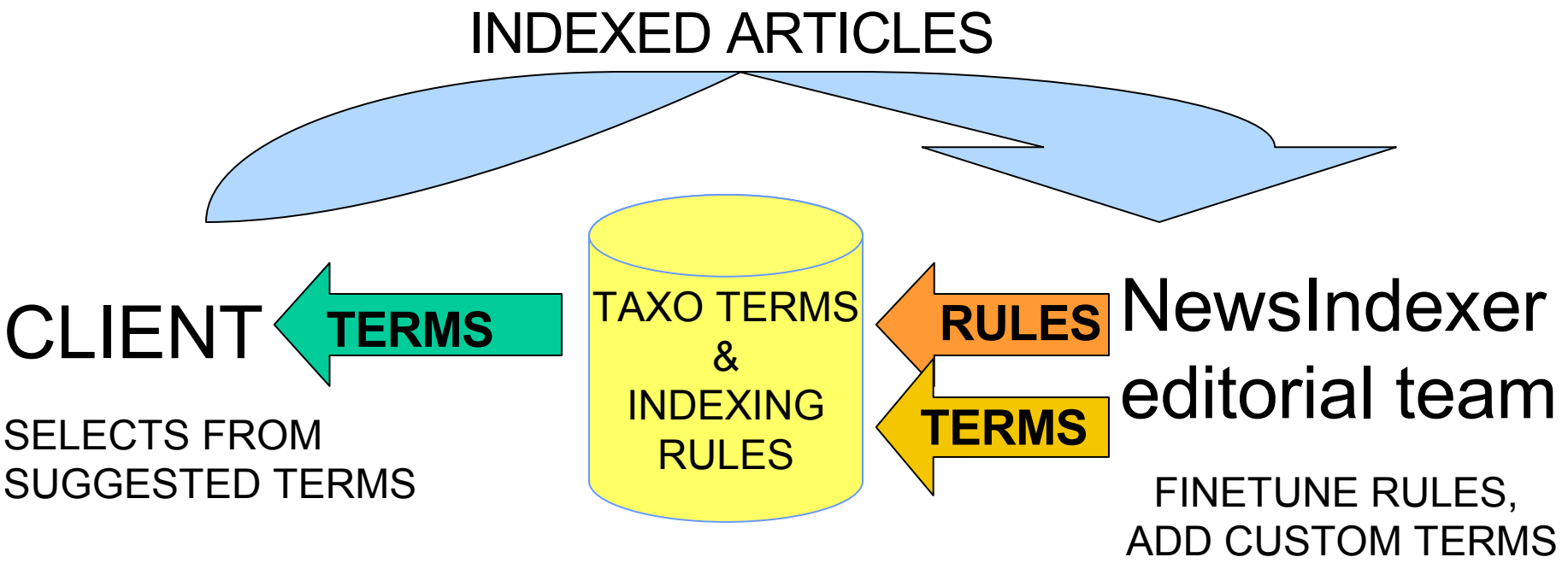
- International Press Telecommunications Council
- ~1300 subject codes, heavy on sports
- Primarily for coding photos
- Often used with NewsML
- Promotes repurposing content
- Used by 55 organizations

- Compatible with NewsIndexer taxonomy and rules
- NewsIndexer taxonomy 4x more detailed, more distributed across topics
- Same text prompts NewsIndexer terms and IPTC terms

NewsIndexer – ASP model

- Client accesses M.A.I. for intelligent indexing suggestions, ~1 second/article or batch mode
  - Indexer has choice and control
    - add terms from taxonomy, reject suggestions
  - Retains maximum precision
  - Four-fold productivity increase
- Client choices conveyed daily to NewsIndexer editorial team at Access Innovations
  - NewsIndexer team monitors client choices
  - Editorial team updates rulebase
  - Future indexing suggestions increasingly accurate

NewsIndexer – ASP model



Feedback loop permits constant monitoring to update terms and rules.

NewsIndexer – Client-owner model

- Client licenses taxonomy and rulebase
  - Customize as needed
- Client purchases software
  - Thesaurus Master for taxonomy management
  - M.A.I. for rule-based categorization
- Client maintains taxonomy and rulebase

NewsIndexer recap

- Taxonomy-based news categorization
- Rules govern suggestion/application of terms
- Editor-review or automatic mode
- Terms applied as permanent meta-data
- Taxo-metatags stored in portable XML format
- ASP model or client-owned software
- Employs Thesaurus Master and M.A.I.
  - Flexible, Customizable, User-friendly interface
  - No limits on taxonomy terms or indexing rules
  - Supports additional taxonomy projects

Closing thoughts on taxonomies

- Taxonomy is the foundation of content organization
- Plan how you'll apply and integrate in workflow
- Plan for growth, avoid any system limitations
- Plan to invest intellectual capital, time, money
  - "Automatic" taxonomy generation *delays* but does not avoid need for *human* analysis, editing, maintenance
- Start somewhere, start small, but start!
  - Many "how to" resources available
  - Follow Z39.19-2005 taxo construction standard
- If not building your own, borrow or license an existing taxonomy and customize it

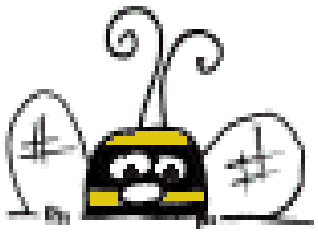
Thanks for your interest in **NewsIndexer**!

For more information:

Visit www.NewsIndexer.com

Contact me:

Alice Redmond-Neal
Access Innovations, Inc.
ared@accessinn.com
(505) 998-0800



Access Innovations, Inc.

DATA HARMONY