Taxonomies for Auto-Tagging Unstructured Content

Heather Hedden
Hedden Information Management

Text Analytics World, Boston, MA

October 1, 2013

About Heather Hedden

- Independent taxonomy consultant, Hedden Information Management
- Continuing education online workshop instructor, Simmons College Graduate School of Library and Information Science
- Author of *The Accidental Taxonomist* (Information Today, Inc., 2010)
- Previously
  - taxonomy consultant employed by a consulting firm
  - taxonomy manager
  - publishing company controlled vocabulary editor
  - taxonomist for enterprise search tool vendor
  - indexer

Outline

- Introduction
- Auto-Tagging and Auto-Categorization Methods
- Taxonomy Basics

Outline

- Introduction
- Auto-Tagging and Auto-Categorization Methods
- Taxonomy Basics

Background

- "Structured data" vs. "unstructured data"
  - Data in a database or not

- "Structured content" vs. "unstructured content"

  Less formally defined
  - Structured Content
    - information or content that has been broken down and classified using metadata
  - Unstructured Content
    - content lacking most metadata
  - But there is also a lot of content with just *some* metadata.

- Metadata fields (title, author, document type, source, location, topic, etc.) are often populated with terms picked from a controlled vocabulary/ taxonomy.
- Taxonomy terms can be tagged directly to unstructured content (or its URI), not necessarily as metadata values.
- Tagging can be manual or automated.

Indexing/Categorization/Tagging Definitions

- Indexing – prominent terms extracted and listed in an index
  - Manual or automated (auto-indexing)
  - Could be for just words or could be for "concepts"
  - May or may not use a taxonomy/controlled vocabulary
- Categorization/classification – documents assigned to categories based on what they are *about*
  - Manual or automated (auto-categorization)
  - Requires a taxonomy of categories
- Tagging – terms assigned to documents for prominence *or* what the documents are about.
  - Manual or automated (auto-tagging)
  - Manual may or may not use a taxonomy/controlled vocabulary; automated requires a taxonomy/controlled vocabulary

Indexing/Categorization/Tagging Methods

Choosing human vs. automated indexing/classification/tagging

Human methods

- Manageable number of docs
- Higher accuracy in indexing
- May include non-text files
- Invest in people
- Low-tech: can build your own indexing tool/user interface
- Internal control

Automated methods

- Very large number of docs
- Greater speed in indexing
- Text files only
- Invest in technology
- High-tech: must purchase auto-indexing/classification software
- Software vendor relationship

Automated Methods

- Auto-Indexing – prominent terms extracted
  - Text analytics and text mining, based on NLP
  - Information extraction, especially entity extraction
- Auto-Categorization/Classification – documents assigned to categories
  - Main methods: Machine-learning or Rules-based
  - May also leverage results from text analytics, information extraction, text mining, etc.
- Auto-Tagging – terms assigned to documents
  - Not much different from auto-categorization, but implied more specific/granular

Outline

- Introduction
- Auto-Tagging and Auto-Categorization Methods
- Taxonomy Basics

Auto-Tagging and Auto-Categorization Methods

Methods:

1. Machine-learning based auto-categorization (Supervised learning; Statistical classification)
2. Rules-based auto-categorization

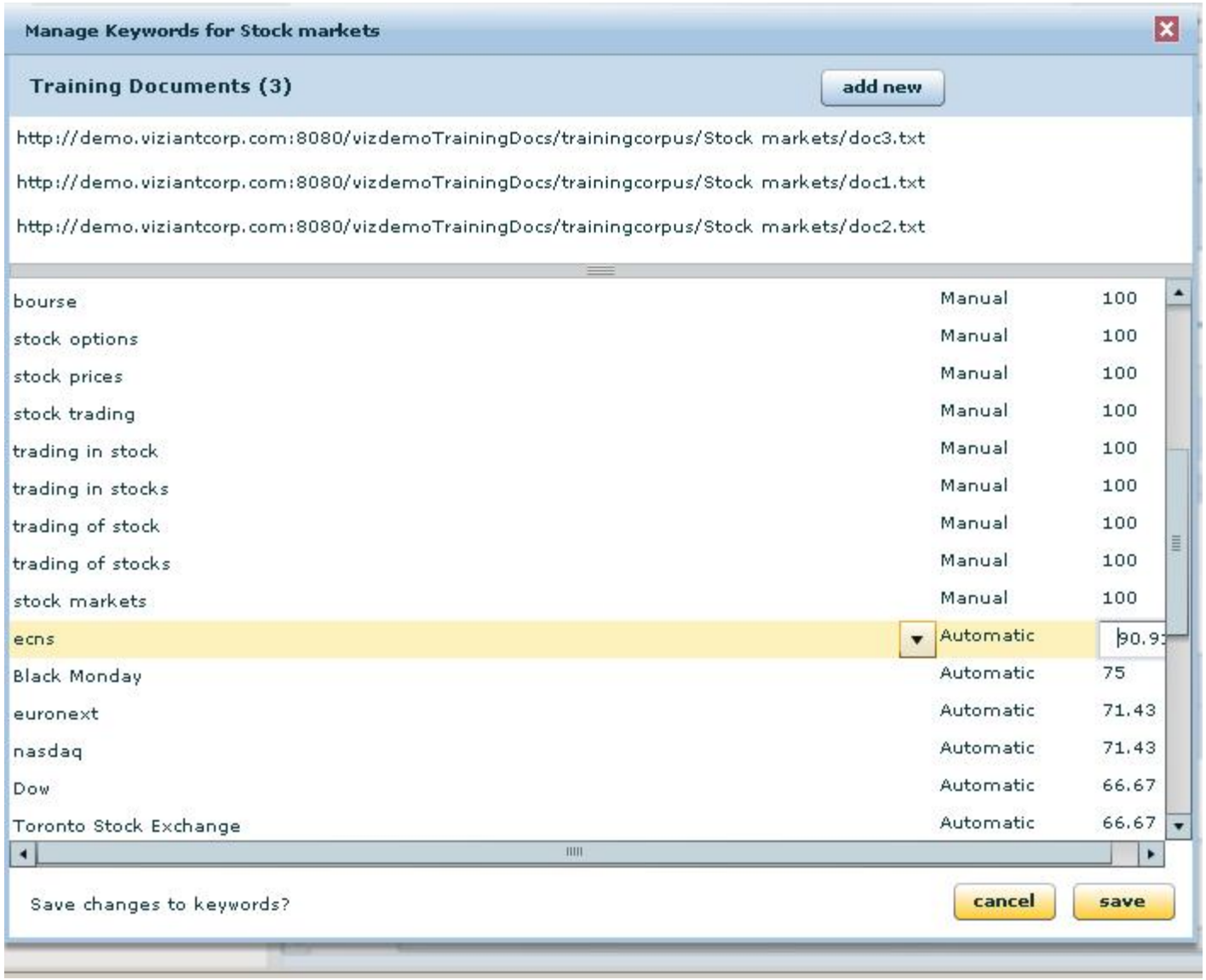
A few tools combine both methods.

Auto-Tagging and Auto-Categorization Methods

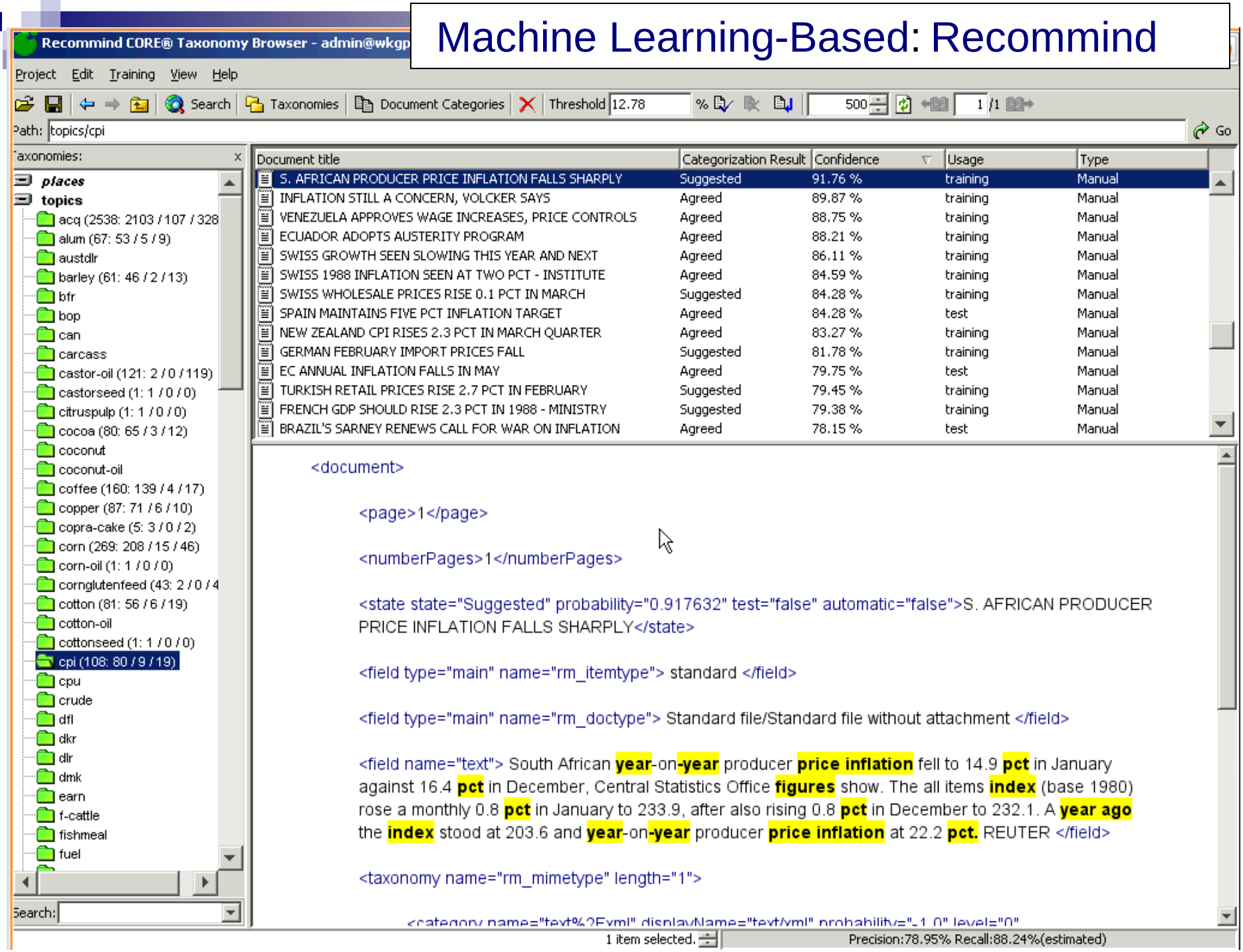
Machine-learning based auto-categorization

- Automatically categorizes/tags based on previous examples.
- System has complex mathematical algorithms.
- Content managers must provide multiple representative sample documents (50-100) for each taxonomy term to “train” the system. Results are reviewed and training sets are “tuned.”
- Matches are to terms and synonyms, which can be individually weighted.
- System may also “suggest” additional terms to add to taxonomy.
- Best if large body of pre-indexed records already exists (such as migrating from human to automated indexing)

Machine Learning-Based Auto-Categorization: Viziant

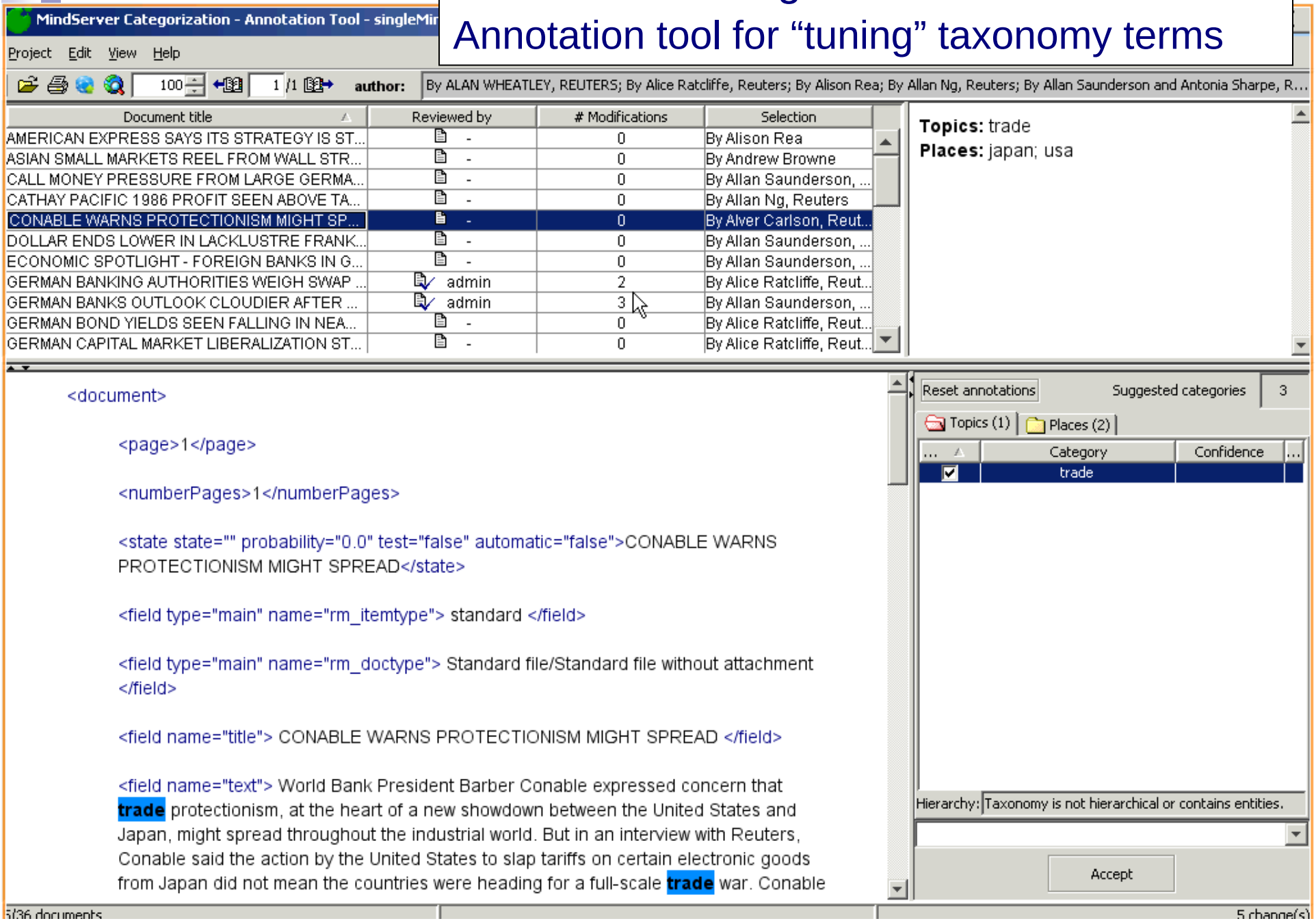


Machine Learning-Based: Recommind



Machine Learning-Based: Recommind

Annotation tool for “tuning” taxonomy terms

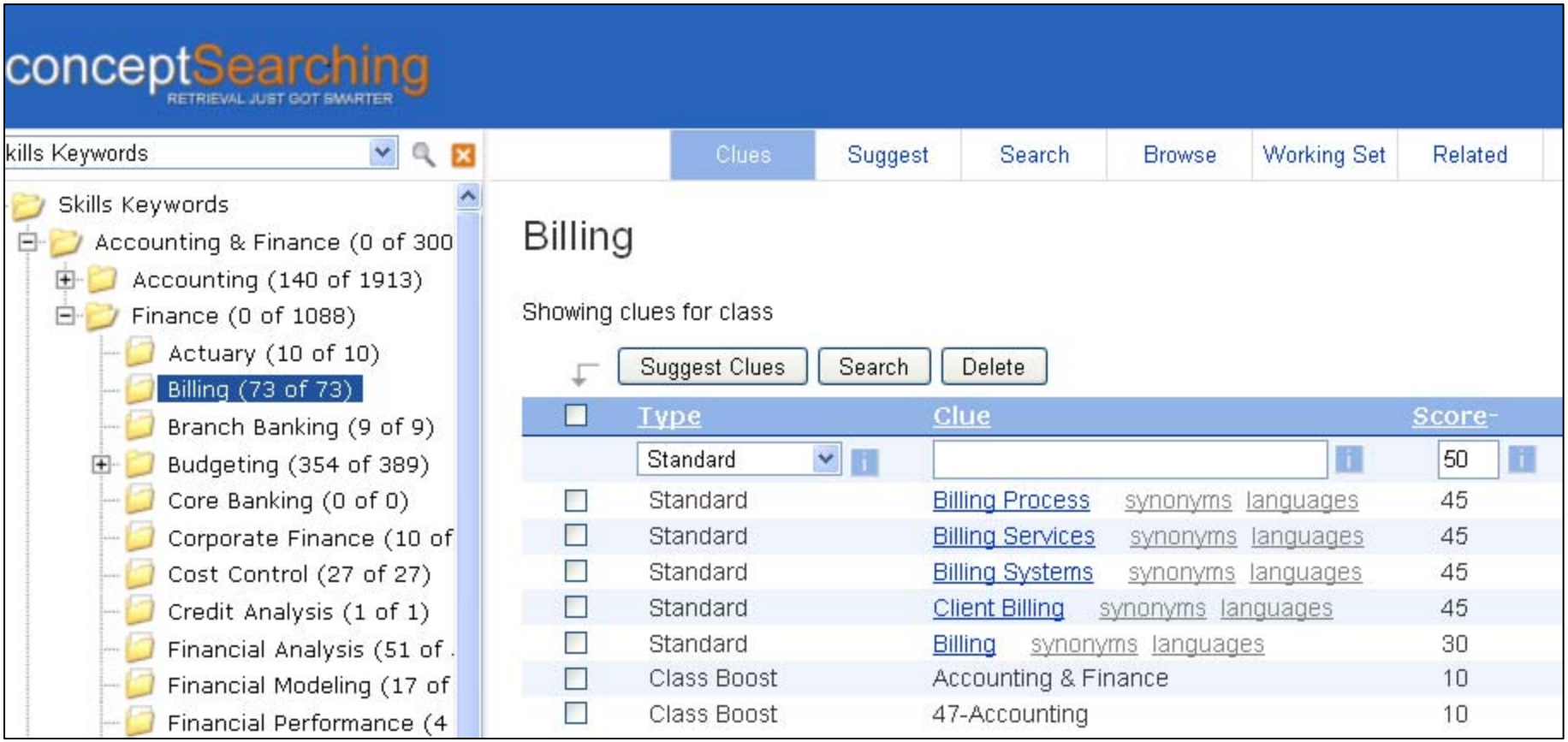


Auto-Tagging and Auto-Categorization Methods

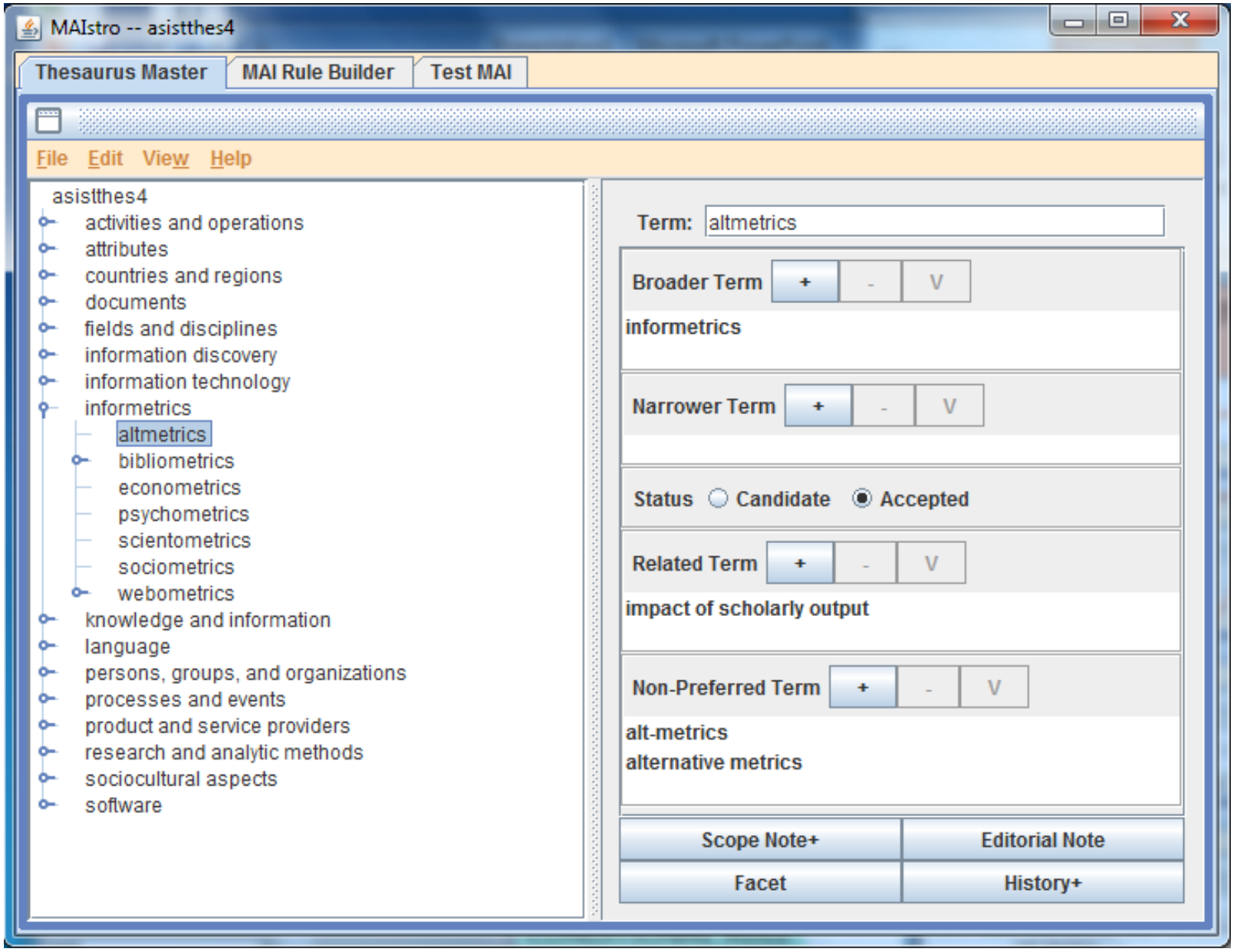
Rules-based auto-categorization

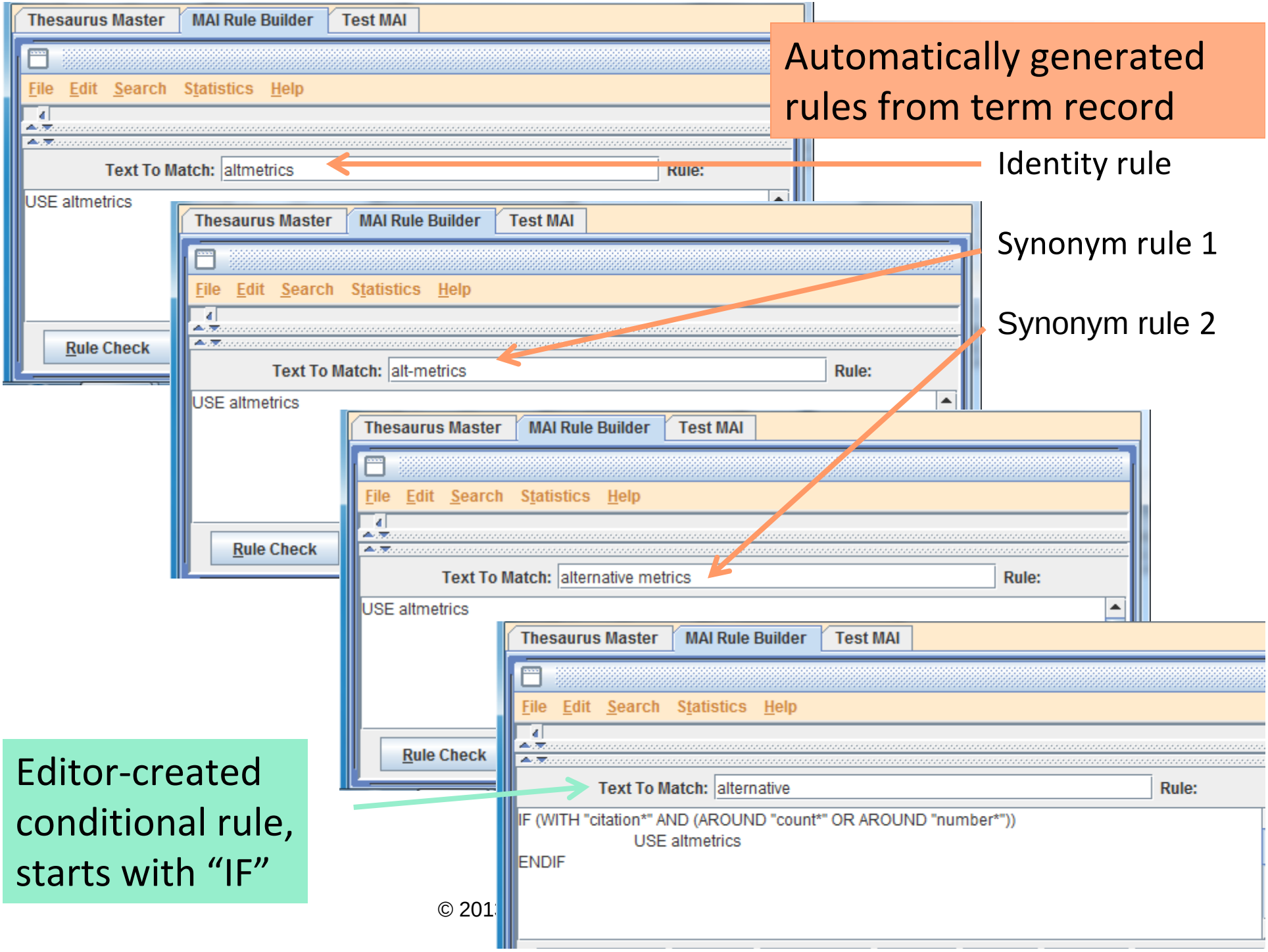
- Rules are created for each taxonomy term.
- Rules are based on synonyms with more conditions.
- Some systems feature weighting of synonyms.
- Some systems feature auto-generated suggested rules for each term/synonym which can be manually edited in addition to writing rules from scratch.
- Some systems feature more sophisticated rule-writing, like advanced Boolean searching (in reverse) and proximity operators or regular expressions.

Rules Based Auto-Categorization: Concept Searching

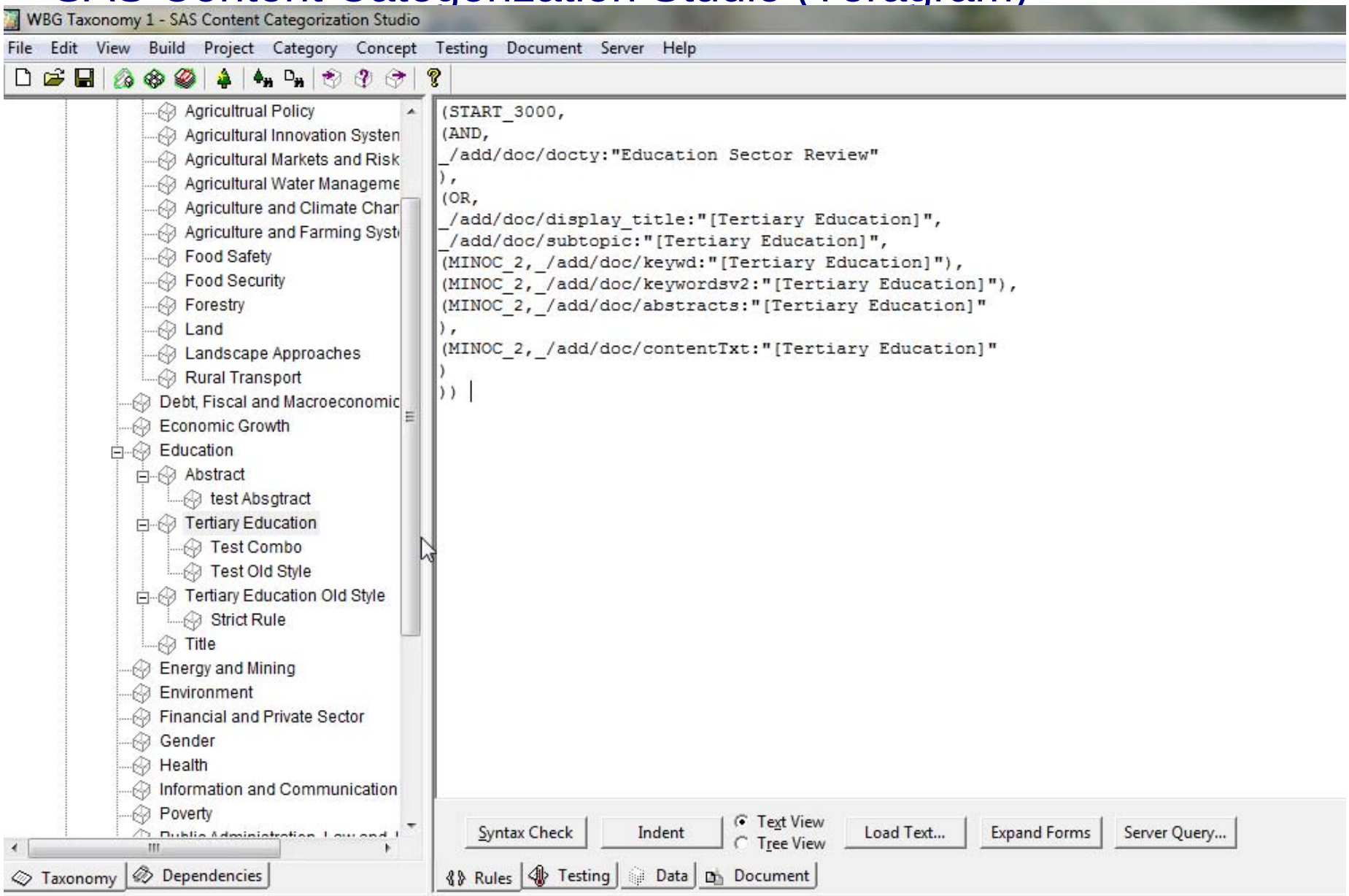


Rules Based Auto-Categorization: Data Harmony MAIstro

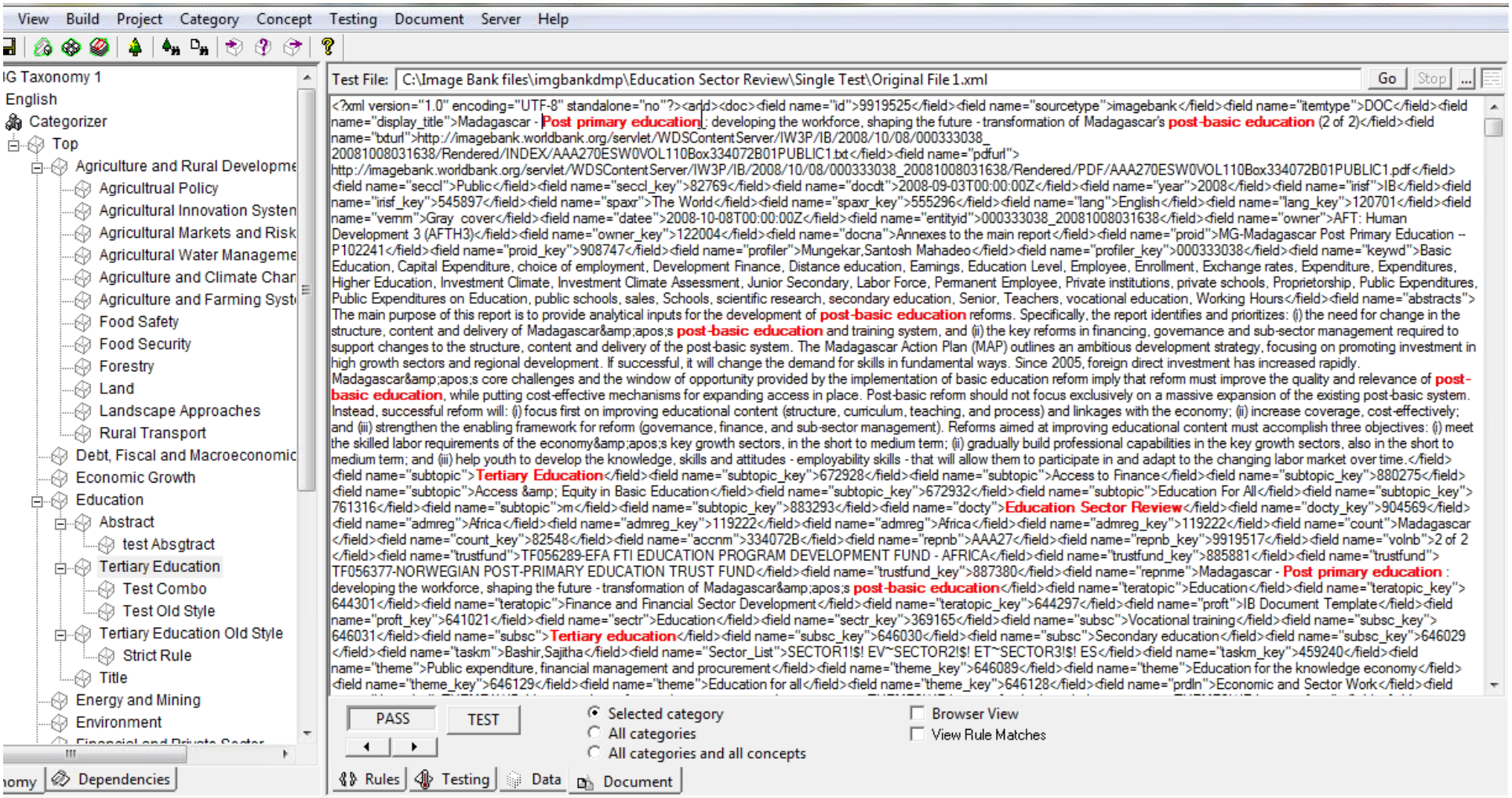



Thesaurus Master
MAI Rule Builder
Test MAI
File Edit Search Statistics Help
Text To Match: altmetrics
Rule:
USE altmetrics
Rule Check
Text To Match: alt-metrics
Rule:
USE altmetrics
Rule Check
Text To Match: alternative metrics
Rule:
USE altmetrics
Rule Check
Text To Match: alternative
Rule:
IF (WITH "citation*" AND (AROUND "count*" OR AROUND "number*"))
USE altmetrics
ENDIF
Automatically generated rules from term record
Identity rule
Synonym rule 1
Synonym rule 2
Editor-created conditional rule, starts with “IF”

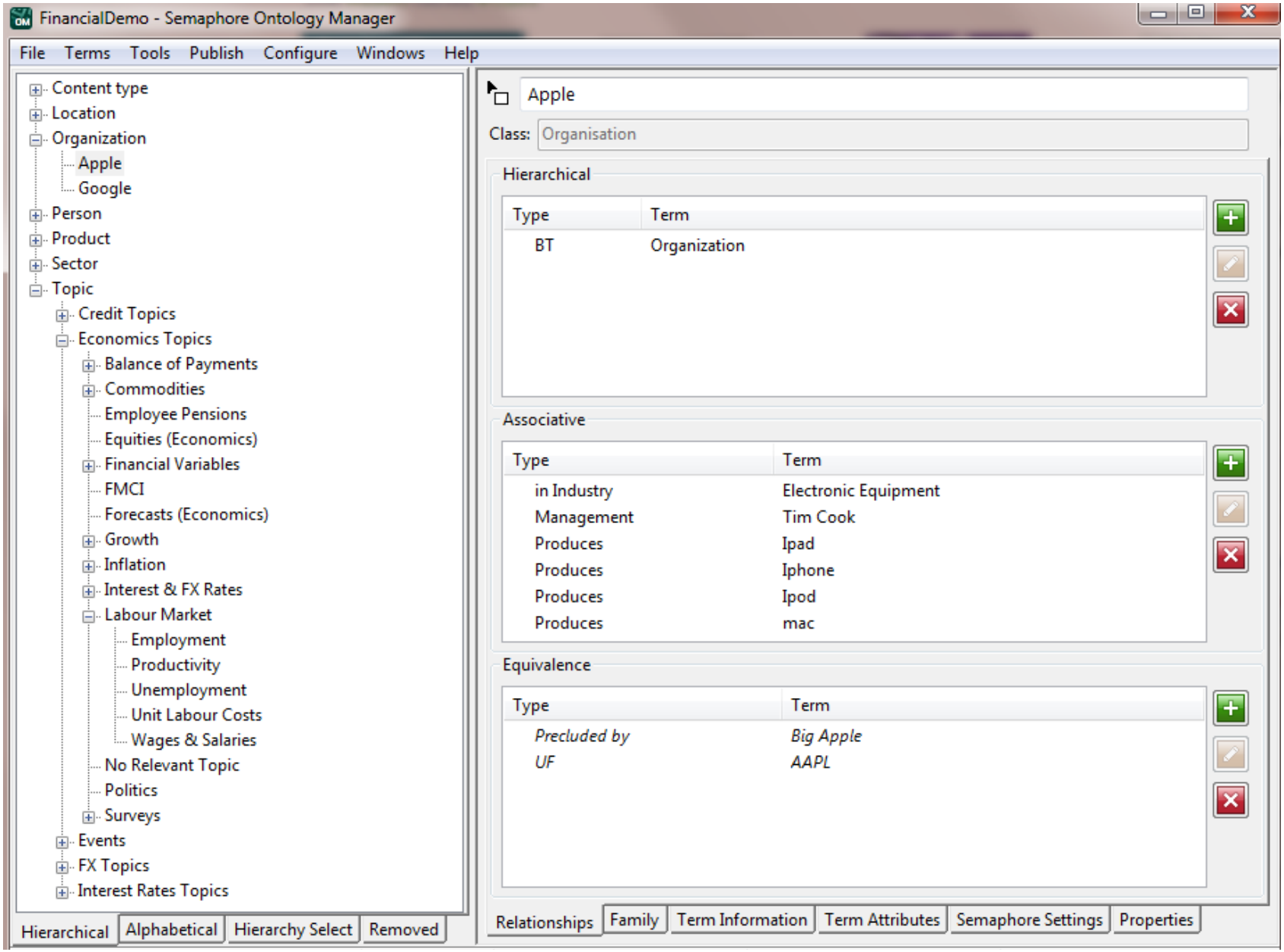
Rules Based Auto-Categorization:
SAS Content Categorization Studio (Teragram)



Rules Based Auto-Categorization: SAS Content Categorization Studio (Teragram)

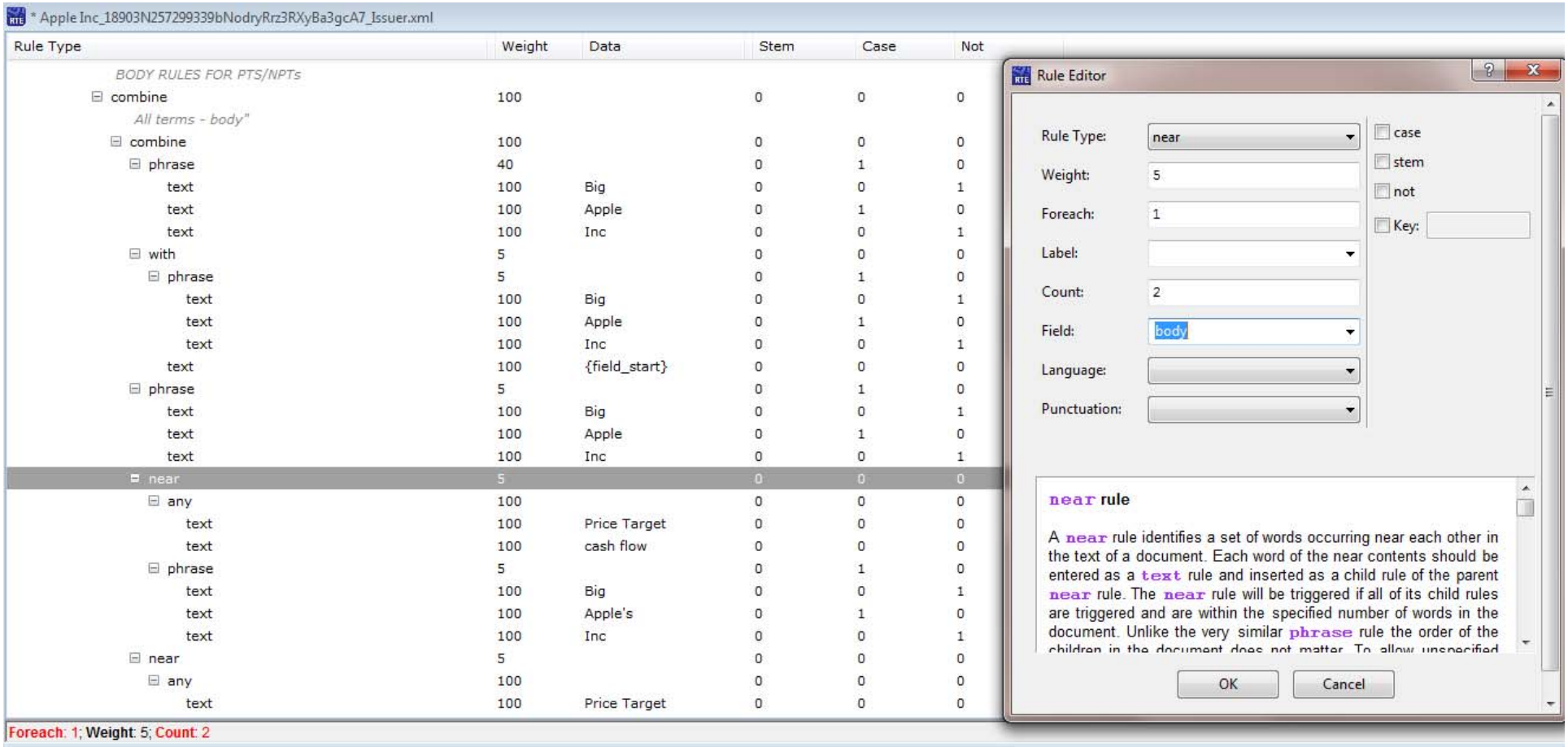


Rules Based Auto-Categorization: Smartlogic Semaphore

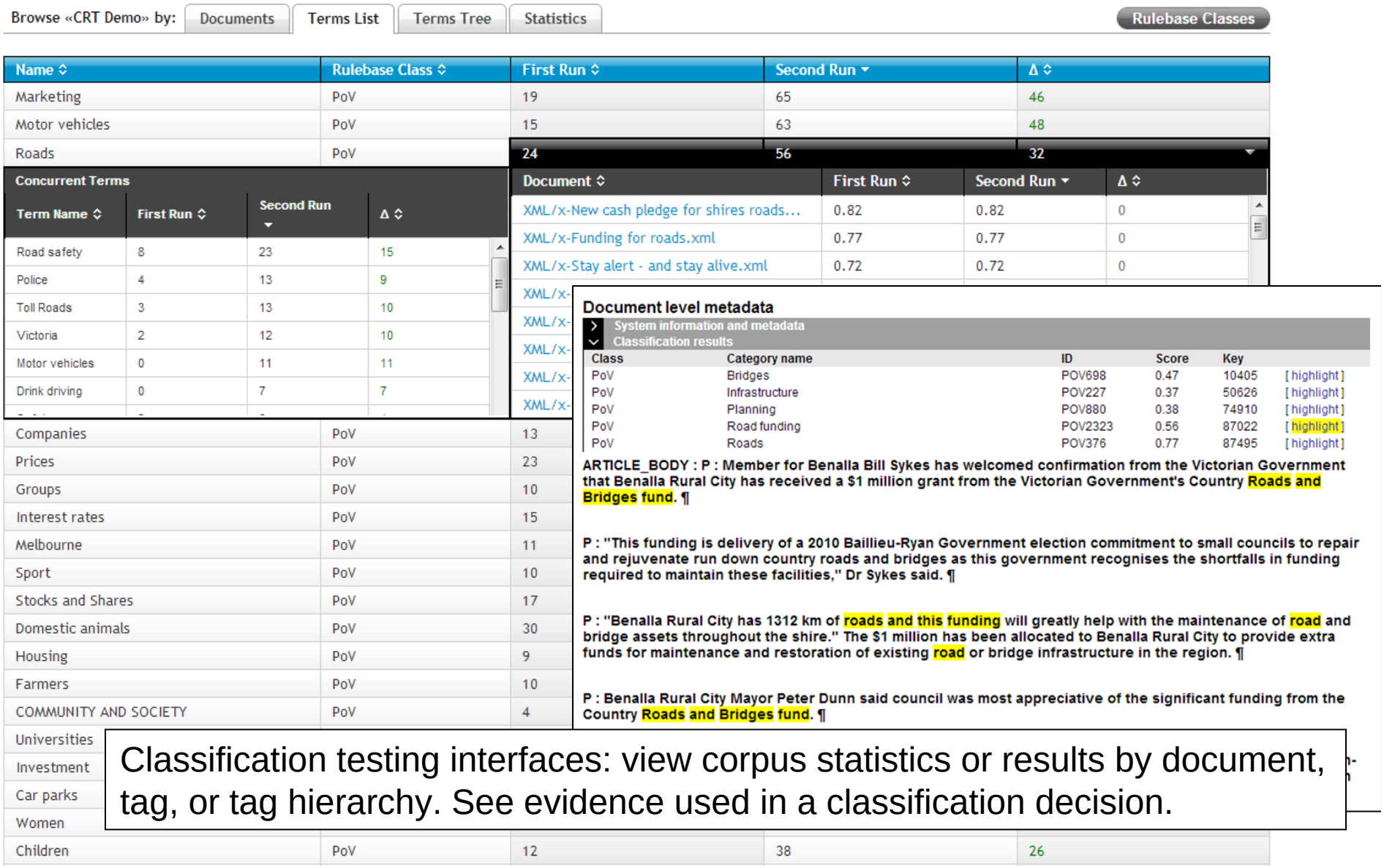


Rules Based Auto-Categorization: Smartlogic Semaphore

Editable rules are automatically created, leveraging content structure, linguistic structure, disambiguation rules, Boolean logic, and term weightings.



Rules Based Auto-Categorization: Smartlogic Semaphore



Browse «CRT Demo» by: Documents | Terms List | Terms Tree | Statistics — Rulebase Classes

Name	Rulebase Class	First Run	Second Run	Δ
Marketing	PoV	19	65	46
Motor vehicles	PoV	15	63	48
Roads	PoV	24	56	32
Companies	PoV	13		
Prices	PoV	23		
Groups	PoV	10		
Interest rates	PoV	15		
Melbourne	PoV	11		
Sport	PoV	10		
Stocks and Shares	PoV	17		
Domestic animals	PoV	30		
Housing	PoV	9		
Farmers	PoV	10		
COMMUNITY AND SOCIETY	PoV	4		
Universities				
Investment				
Car parks				
Women				
Children	PoV	12	38	26

Concurrent Terms

Term Name	First Run	Second Run	Δ
Road safety	8	23	15
Police	4	13	9
Toll Roads	3	13	10
Victoria	2	12	10
Motor vehicles	0	11	11
Drink driving	0	7	7

Document	First Run	Second Run	Δ
XML/x-New cash pledge for shires roads...	0.82	0.82	0
XML/x-Funding for roads.xml	0.77	0.77	0
XML/x-Stay alert - and stay alive.xml	0.72	0.72	0

Document level metadata

System information and metadata

Classification results

Class	Category name	ID	Score	Key	
PoV	Bridges	POV698	0.47	10405	[highlight]
PoV	Infrastructure	POV227	0.37	50626	[highlight]
PoV	Planning	POV880	0.38	74910	[highlight]
PoV	Road funding	POV2323	0.56	87022	[highlight]
PoV	Roads	POV376	0.77	87495	[highlight]

ARTICLE_BODY : P : Member for Benalla Bill Sykes has welcomed confirmation from the Victorian Government that Benalla Rural City has received a $1 million grant from the Victorian Government's Country Roads and Bridges fund. ¶

P : "This funding is delivery of a 2010 Baillieu-Ryan Government election commitment to small councils to repair and rejuvenate run down country roads and bridges as this government recognises the shortfalls in funding required to maintain these facilities," Dr Sykes said. ¶

P : "Benalla Rural City has 1312 km of roads and this funding will greatly help with the maintenance of road and bridge assets throughout the shire." The $1 million has been allocated to Benalla Rural City to provide extra funds for maintenance and restoration of existing road or bridge infrastructure in the region. ¶

P : Benalla Rural City Mayor Peter Dunn said council was most appreciative of the significant funding from the Country Roads and Bridges fund. ¶

Classification testing interfaces: view corpus statistics or results by document, tag, or tag hierarchy. See evidence used in a classification decision.

Manual Tasks for Auto-Categorization

- Continual update work is needed for each new term created.
  - New training documents added and taxonomy terms tuned
  - New rules created or edited

- Feeding and tuning training documents is more appropriate for subject matter experts, editors, indexers.
- Writing rules is more appropriate for information professionals, taxonomists, knowledge engineers.

- Taxonomy should be manually created/edited.
  - Auto-tagging systems may suggest terms, but not structure.

Outline

- Introduction
- Auto-Tagging and Auto-Categorization Methods
- Taxonomy Basics

Taxonomy Basics

- Definition and Types
  - Broad Designations
  - Specific Types
- Purposes and Benefits
- Synonyms for Terms
- Hierarchy Best Practices

Taxonomy Definitions & Types

Broad Designations:

Controlled vocabulary, knowledge organization system, taxonomy

- An authoritative, restricted list of terms (words or phrases)
- Each term for a single unambiguous concept (synonyms/nonpreferred terms, as cross-references, may be included)
- Policies (control) for who, when, and how new terms can be added
- May or may not have structured relationships between terms
- To support indexing/tagging/metadata management of content to facilitate content management and retrieval

Taxonomy Definitions & Types

Specific types:

- Term Lists/Pick lists
- Synonym Rings
- Authority Files
- Taxonomies
  - Hierarchical
  - Faceted
- Thesauri
- Ontologies (going beyond a controlled vocabulary)

Often “taxonomy” is used to mean any controlled vocabulary.

Taxonomy Definitions & Types

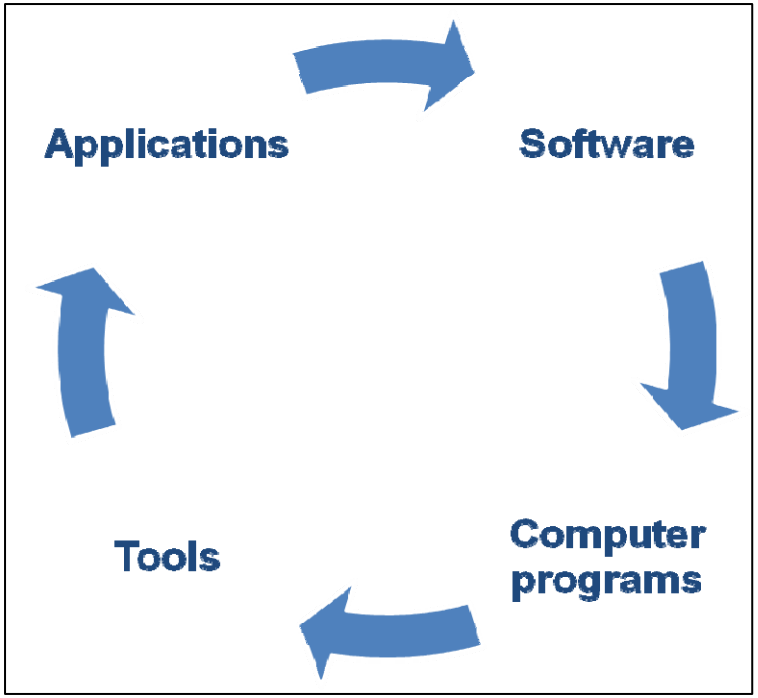
Term List/Pick List

- A simple list of terms
- Lacking synonyms, usually short enough for browsing
- Often displayed in drop-down scroll boxes

Taxonomy Definitions & Types

Synonym Ring

- A controlled vocabulary with synonyms or near-synonyms for each concept
- No designated “preferred” term: All terms are equal and point to each other, as in a ring.
- Table for terms does *not* display to the user



	A	B	C	D
1	applications	software	computer programs	tools
2	administrative agencies	federal agencies	government agencies	
3	civil actions	civil litigation	civil cases	
4	agriculture	farming		
5	Americans with Disabilities Act	ADA		

Taxonomy Definitions & Types

Authority File

Term list, where alternate labels point to the displayed "preferred" term.

Federal Deposit Insurance Corporation

Used from FDIC

Used from Federal Deposit Insurance Corp.

Federal Reserve Board

Used from Federal Reserve

Used from FRB

Office of the Comptroller of the Currency

Used from OCC

Office of Thrift Supervision

Used from OTC

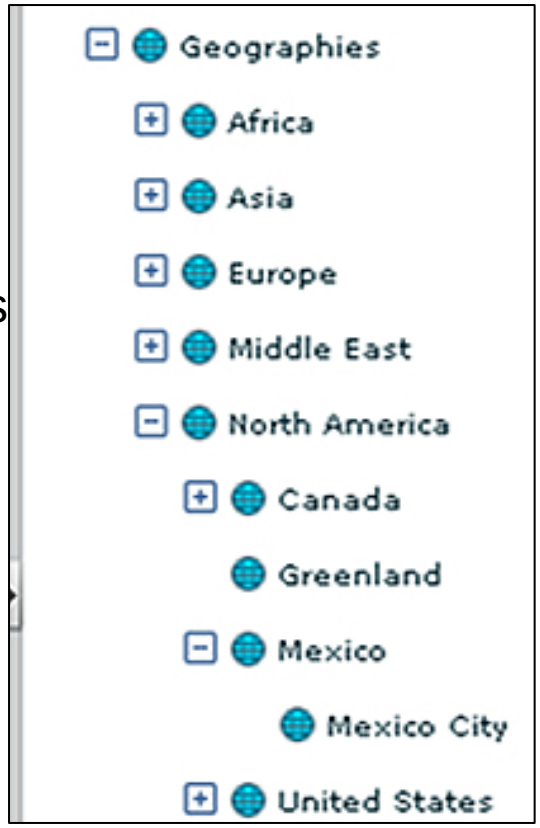
Taxonomy Definitions & Types

Taxonomy

A controlled vocabulary with hierarchical/categorical structure

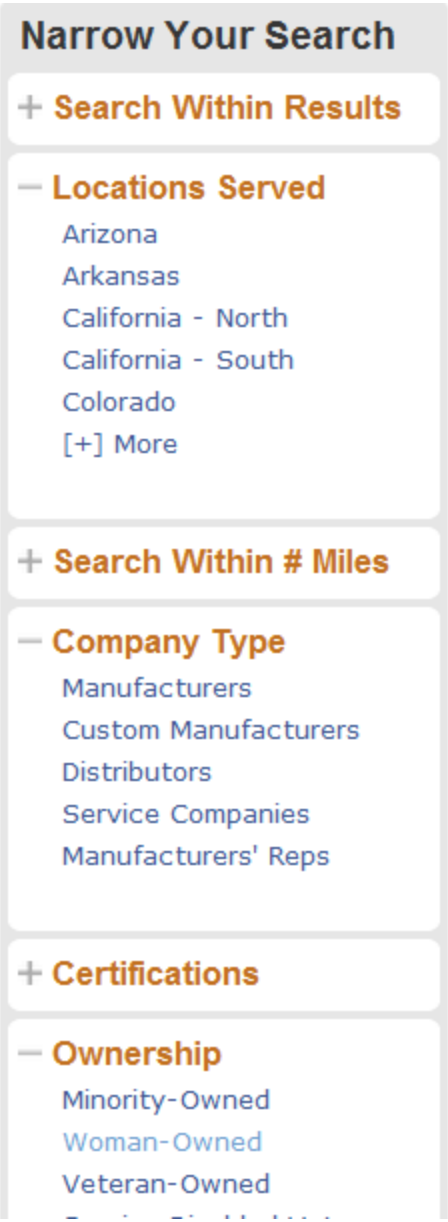
Hierarchical

Has broader term/narrower term relationships that include all terms to create a hierarchical structure



Faceted

Has sets of different types/ aspects which the user selects in combination to refine a search by.



Taxonomy Definitions & Types

Thesaurus

- A controlled vocabulary with standard structured relationships between terms:
  - Hierarchical: broader/narrower terms
  - Equivalence: preferred term/non-preferred term (used from) (alternate labels)
  - Associative: related terms
- Follows ANSI/NISO Z39.19 standards
- May lack the structure of a limited set of top hierarchies
- More suited for alphabetical browsing

Thesaurus term entry example

Government lending
>BT Economic policy
<NT Veterans' loans
RT Agricultural credit
RT Federally-assisted loans
RT Federally-guaranteed loans
RT Government and business
RT Government insurance
RT Loans
RT Student loan funds
UF American domestic economic assistance
UF Federal aid to depressed areas
UF Federal credit programs
UF Federal domestic assistance programs
UF Government loans

BT = Broader term
NT = Narrower term
RT = Related term
UF = Used from

Taxonomy Purposes & Benefits

1. Controlled vocabulary aspect:
   Brings together different wordings (synonyms) for the same concept and disambiguates terms
   - Helps people search for information by different names
   - Helps people retrieve matching concepts, not just words

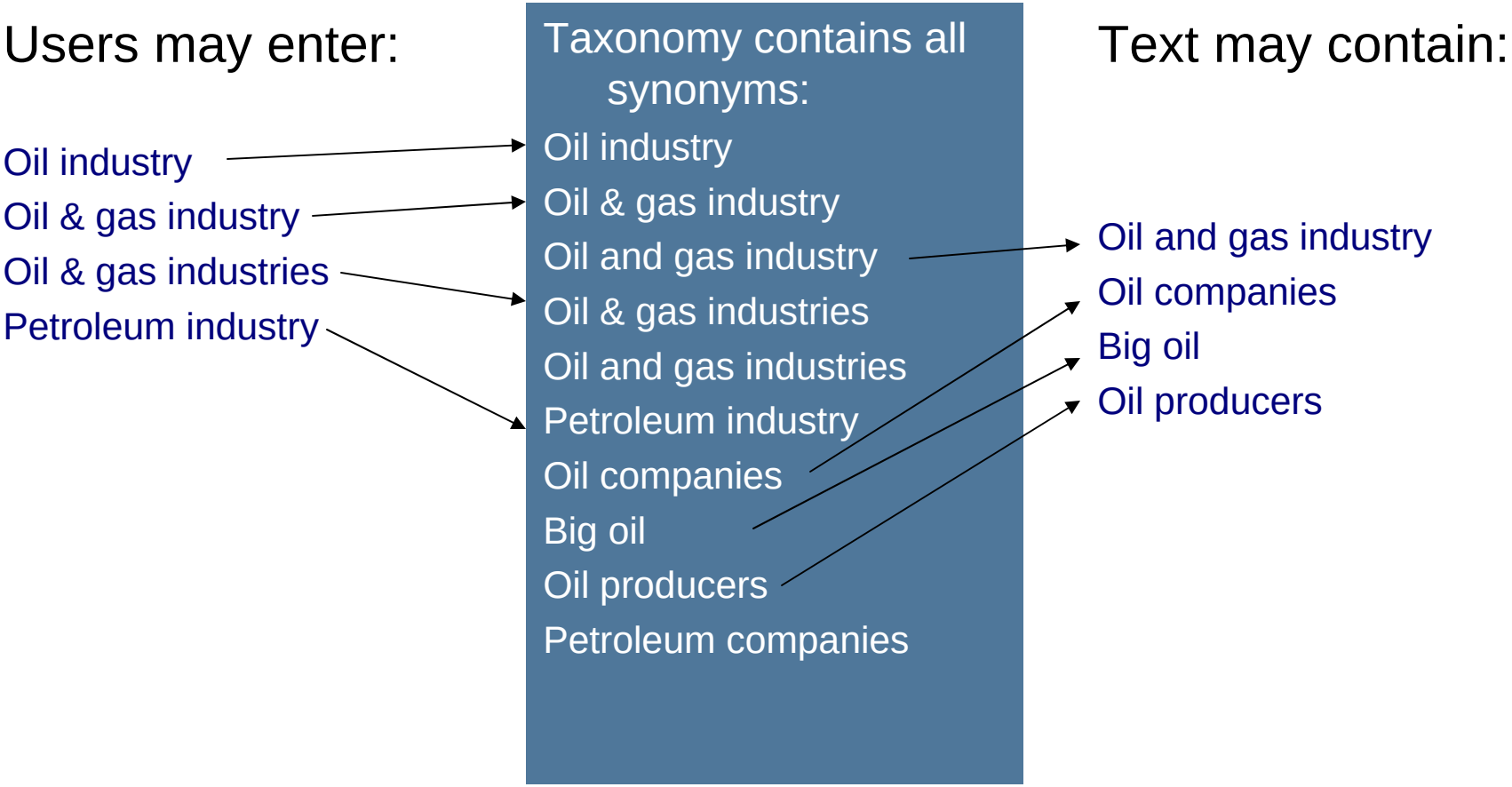
2. Taxonomy or thesaurus structure aspect:
   Organizes information into a logical structure
   - Helps people browse or navigate for information
   - Allows broader concepts to include content of narrower concepts

Taxonomy Purposes & Benefits

- A controlled vocabulary gathers synonyms, acronyms, variant spellings, etc.
  - Content is not missed due to use of different words (e.g. ***Automobiles***, instead of ***Cars***)
  - Without a controlled vocabulary, content would be missed.

- A search restricted on the controlled vocabulary retrieves concepts not just words.
  - Content is excluded for mere text-string matches (e.g. ***monitors*** for computers, not the verb “observes”)
  - Without a controlled vocabulary, too much irrelevant content would be retrieved.

Taxonomy Purposes & Benefits



Synonyms

- Supports search in most controlled vocabulary types: synonym rings, authority files, thesauri, (some taxonomies)
- Anticipating both:
  - varied user search string entries
  - varied forms in the text for the same content
- For both manual and automated indexing
- A concept may have any number of synonyms, but a synonym can point to only one preferred term
- Varied synonym sources:
  - Search analytics records
  - Interviews and use cases
  - Legacy print indexes
  - Obvious patterns (acronyms, phrase inversions, etc.)

Synonyms Creation Tips

Not all are actual “synonyms.” Types include:

- synonyms: Cars USE Automobiles
- near-synonyms: Junior high USE Middle school
- variant spellings: Defence USE Defense
- lexical variants: Hair loss USE Baldness
- foreign language proper nouns: Luftwaffe USE German Air Force
- acronyms/spelled out forms: UN USE United Nations
- scientific/technical names: Neoplasms USE Cancer
- phrase variations: Buses, school USE School buses
- antonyms: Misbehavior USE Behavior
- narrower terms: Alcoholism USE Substance abuse

Also called: variant terms, equivalence terms, non-preferred terms, alternate labels, cross references, etc.

Synonym Creation Tips

Synonym/variant term differences:

For human-indexing

Presidential candidates

Candidates, presidential

For auto-categorization

Presidential candidate

Presidential candidacy

Candidate for president

Candidacy for president

Presidential hopeful

Running for president

Campaigning for president

Presidential nominee

Synonym Creation Tips

- Create as many as possible while maintaining uniqueness
- A synonym can only be used once/can point to only one preferred term...
  Unless, weighting is used. Synonyms of weights less than 100% can be used repeatedly for different preferred terms.
- Variants for Plural/singular?
  Depends on whether the system supports automatic “stemming”
  Stemming might exist for single words but not phrases.
  - **Stations** stems to **Station**
  - **Train stations** may not stem
    Need to add non-preferred term: **Train station**

Hierarchy Best Practices

Hierarchies of terms/concepts:

- Help users browse and navigate to concepts.
- Allow broader concepts to also include content indexed to narrower concepts.
- Provide structure and method for an organization/taxonomist to build and manage comprehensive, in-scope taxonomies.

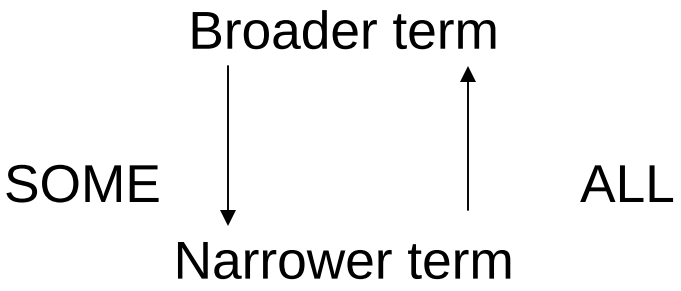
Hierarchy Best Practices

Hierarchical Relationships

Broader Term > Narrower Term

Parent > Child

Superordinate > Subordinate



Two types:

1. Generic > Specific/Instance
2. Whole > Part

Some of *broader term* are/are in *narrower term*.

All of *narrower term* are within *broader term*.

Hierarchy Best Practices

Hierarchical Relationship: Generic > Specific/Instance

Category or class

> members or more specific types

Examples:	
Languages	Financial services
> German	> Investment Services

- Narrower term “is/are a” broader term
- Narrower term “is/are a kind of” broader term

Hierarchy Best Practices

Hierarchical Relationship: Whole > Part

Concept or Entity
> part or subentity

Examples:

U.S. Department of Treasury
> Internal Revenue Service

United States
> California

- Narrower term “is a component of” broader term
- Narrower term “is a sub-unit” of broader term
- Narrower term “is in” broader term

Hierarchy Best Practices

Hierarchical Relationship: Polyhierarchy

Sometimes a term can have two or more broader terms.

- Must be the same term (same ID number)
- Is tagged to the same set of documents
- Is not context-dependent
- Must follow the “is a”/“is a part of” rule for hierarchical relationships in both locations.

Example:

State Laws
> California General Corporate Law

Corporation Laws
> California General Corporate Law

Taxonomies for Auto-Tagging/Categorization

Taxonomies designed for auto-tagging/categorization:

- Need more, varied synonym/variant terms
- Need variant terms of different parts of speech
- Need to be more content-tailored, content-based
- Cannot have subtle differences between concepts
- Should avoid including action (verbal) terms

  For example both **Investing** and **Investments**

Taxonomy Resources

- ANSI/NISO Z39.19 (2005) *Guidelines for Construction, Format, and Management of Monolingual Controlled Vocabularies.* Bethesda, MD: NISO Press. www.niso.org
- Hedden, Heather. (2010) *The Accidental Taxonomist.* Medford, NJ: Information Today Inc. www.accidental-taxonomist.com
- American Society for Indexing: Taxonomies and Controlled Vocabularies Special Interest Group www.taxonomies-sig.org
- Special Libraries Association (SLA): Taxonomy Division http://wiki.sla.org/display/SLATAX
- Taxonomy Community of Practice discussion group http://finance.groups.yahoo.com/group/TaxoCoP
- "Taxonomies and Controlled Vocabularies" Simmons College Graduate School of Library and Information Science Continuing Education Program, 5 weeks. $250. November 2013. http://alanis.simmons.edu/ceweb/byinstructor.php#9

47

Questions/Contact

Heather Hedden
Hedden Information Management
Carlisle, MA
heather@hedden.net
978-467-5195
www.hedden-information.com
www.linkedin.com/in/hedden
twitter.com/hhedden
accidental-taxonomist.blogspot.com