



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

Manage Keywords for Stock markets

Training Documents (3) add new

http://demo.viziantcorp.com:8080/vizdemoTrainingDocs/trainingcorpus/Stock markets/doc3.txt
http://demo.viziantcorp.com:8080/vizdemoTrainingDocs/trainingcorpus/Stock markets/doc1.txt
http://demo.viziantcorp.com:8080/vizdemoTrainingDocs/trainingcorpus/Stock markets/doc2.txt

bourse	Manual	100
stock options	Manual	100
stock prices	Manual	100
stock trading	Manual	100
trading in stock	Manual	100
trading in stocks	Manual	100
trading of stock	Manual	100
trading of stocks	Manual	100
stock markets	Manual	100
ecns	Automatic	90.9
Black Monday	Automatic	75
euronext	Automatic	71.43
nasdaq	Automatic	71.43
Dow	Automatic	66.67
Toronto Stock Exchange	Automatic	66.67

Save changes to keywords? cancel save

Machine Learning-Based: Recommind

Recommind CORE® Taxonomy Browser - admin@wkgp

Project Edit Training View Help

Search Taxonomies Document Categories Threshold 12.78 % 500 1 / 1

Path: topics/cpi

Taxonomies:

places

topics

- acq (2538: 2103 / 107 / 328)
- alum (67: 53 / 5 / 9)
- austdlr
- barley (61: 46 / 2 / 13)
- bfr
- bop
- can
- carcass
- castor-oil (121: 2 / 0 / 119)
- castorseed (1: 1 / 0 / 0)
- citruspulp (1: 1 / 0 / 0)
- cocoa (80: 65 / 3 / 12)
- coconut
- coconut-oil
- coffee (160: 139 / 4 / 17)
- copper (87: 71 / 6 / 10)
- copra-cake (5: 3 / 0 / 2)
- corn (269: 208 / 15 / 46)
- corn-oil (1: 1 / 0 / 0)
- corn glutenfeed (43: 2 / 0 / 4)
- cotton (81: 56 / 6 / 19)
- cotton-oil
- cottonseed (1: 1 / 0 / 0)
- cpi (108: 80 / 9 / 19)
- cpu
- crude
- dfl
- dkr
- dlr
- dmk
- earn
- f-cattle
- fishmeal
- fuel

Search:

Document title	Categorization Result	Confidence	Usage	Type
S. AFRICAN PRODUCER PRICE INFLATION FALLS SHARPLY	Suggested	91.76 %	training	Manual
INFLATION STILL A CONCERN, VOLCKER SAYS	Agreed	89.87 %	training	Manual
VENEZUELA APPROVES WAGE INCREASES, PRICE CONTROLS	Agreed	88.75 %	training	Manual
ECUADOR ADOPTS AUSTERITY PROGRAM	Agreed	88.21 %	training	Manual
SWISS GROWTH SEEN SLOWING THIS YEAR AND NEXT	Agreed	86.11 %	training	Manual
SWISS 1988 INFLATION SEEN AT TWO PCT - INSTITUTE	Agreed	84.59 %	training	Manual
SWISS WHOLESALE PRICES RISE 0.1 PCT IN MARCH	Suggested	84.28 %	training	Manual
SPAIN MAINTAINS FIVE PCT INFLATION TARGET	Agreed	84.28 %	test	Manual
NEW ZEALAND CPI RISES 2.3 PCT IN MARCH QUARTER	Agreed	83.27 %	training	Manual
GERMAN FEBRUARY IMPORT PRICES FALL	Suggested	81.78 %	training	Manual
EC ANNUAL INFLATION FALLS IN MAY	Agreed	79.75 %	test	Manual
TURKISH RETAIL PRICES RISE 2.7 PCT IN FEBRUARY	Suggested	79.45 %	training	Manual
FRENCH GDP SHOULD RISE 2.3 PCT IN 1988 - MINISTRY	Suggested	79.38 %	training	Manual
BRAZIL'S SARNEY RENEWS CALL FOR WAR ON INFLATION	Agreed	78.15 %	test	Manual

<document>

<page>1</page>

<numberPages>1</numberPages>

<state state="Suggested" probability="0.917632" test="false" automatic="false">S. AFRICAN PRODUCER PRICE INFLATION FALLS SHARPLY</state>

<field type="main" name="rm_itemtype"> standard </field>

<field type="main" name="rm_doctype"> Standard file/Standard file without attachment </field>

<field name="text"> South African **year-on-year** producer **price inflation** fell to 14.9 **pct** in January against 16.4 **pct** in December, Central Statistics Office **figures** show. The all items **index** (base 1980) rose a monthly 0.8 **pct** in January to 233.9, after also rising 0.8 **pct** in December to 232.1. A **year ago** the **index** stood at 203.6 and **year-on-year** producer **price inflation** at 22.2 **pct**. REUTER </field>

<taxonomy name="rm_mimetype" length="1">

<category name="text%?Exml" displayname="text/xml" probability="1.0" level="0">

1 item selected.

Precision:78.95% Recall:88.24%(estimated)

Machine Learning-Based: Recommend Annotation tool for “tuning” taxonomy terms

MindServer Categorization - Annotation Tool - singleMir

Project Edit View Help

100 1/1 author: By ALAN WHEATLEY, REUTERS; By Alice Ratcliffe, Reuters; By Alison Rea; By Allan Ng, Reuters; By Allan Saunderson and Antonia Sharpe, R...

Document title	Reviewed by	# Modifications	Selection
AMERICAN EXPRESS SAYS ITS STRATEGY IS ST...	-	0	By Alison Rea
ASIAN SMALL MARKETS REEL FROM WALL STR...	-	0	By Andrew Browne
CALL MONEY PRESSURE FROM LARGE GERMA...	-	0	By Allan Saunderson, ...
CATHAY PACIFIC 1986 PROFIT SEEN ABOVE TA...	-	0	By Allan Ng, Reuters
CONABLE WARNS PROTECTIONISM MIGHT SP...	-	0	By Alver Carlson, Reut...
DOLLAR ENDS LOWER IN LACKLUSTRE FRANK...	-	0	By Allan Saunderson, ...
ECONOMIC SPOTLIGHT - FOREIGN BANKS IN G...	-	0	By Allan Saunderson, ...
GERMAN BANKING AUTHORITIES WEIGH SWAP ...	admin	2	By Alice Ratcliffe, Reut...
GERMAN BANKS OUTLOOK CLOUDIER AFTER ...	admin	3	By Allan Saunderson, ...
GERMAN BOND YIELDS SEEN FALLING IN NEA...	-	0	By Alice Ratcliffe, Reut...
GERMAN CAPITAL MARKET LIBERALIZATION ST...	-	0	By Alice Ratcliffe, Reut...

Topics: trade
Places: japan; usa

<document>

<page>1</page>

<numberPages>1</numberPages>

<state state="" probability="0.0" test="false" automatic="false">CONABLE WARNS
PROTECTIONISM MIGHT SPREAD</state>

<field type="main" name="rm_itemtype"> standard </field>

<field type="main" name="rm_doctype"> Standard file/Standard file without attachment
</field>

<field name="title"> CONABLE WARNS PROTECTIONISM MIGHT SPREAD </field>

<field name="text"> World Bank President Barber Conable expressed concern that
trade protectionism, at the heart of a new showdown between the United States and
Japan, might spread throughout the industrial world. But in an interview with Reuters,
Conable said the action by the United States to slap tariffs on certain electronic goods
from Japan did not mean the countries were heading for a full-scale **trade** war. Conable

Reset annotations

Suggested categories 3

Topics (1) Places (2)

Category	Confidence
<input checked="" type="checkbox"/> trade	

Hierarchy: Taxonomy is not hierarchical or contains entities.

Accept



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

conceptSearching
RETRIEVAL JUST GOT SMARTER

kills Keywords

Skills Keywords

- Accounting & Finance (0 of 300)
 - Accounting (140 of 1913)
 - Finance (0 of 1088)
 - Actuary (10 of 10)
 - Billing (73 of 73)**
 - Branch Banking (9 of 9)
 - Budgeting (354 of 389)
 - Core Banking (0 of 0)
 - Corporate Finance (10 of 10)
 - Cost Control (27 of 27)
 - Credit Analysis (1 of 1)
 - Financial Analysis (51 of 51)
 - Financial Modeling (17 of 17)
 - Financial Performance (4 of 4)

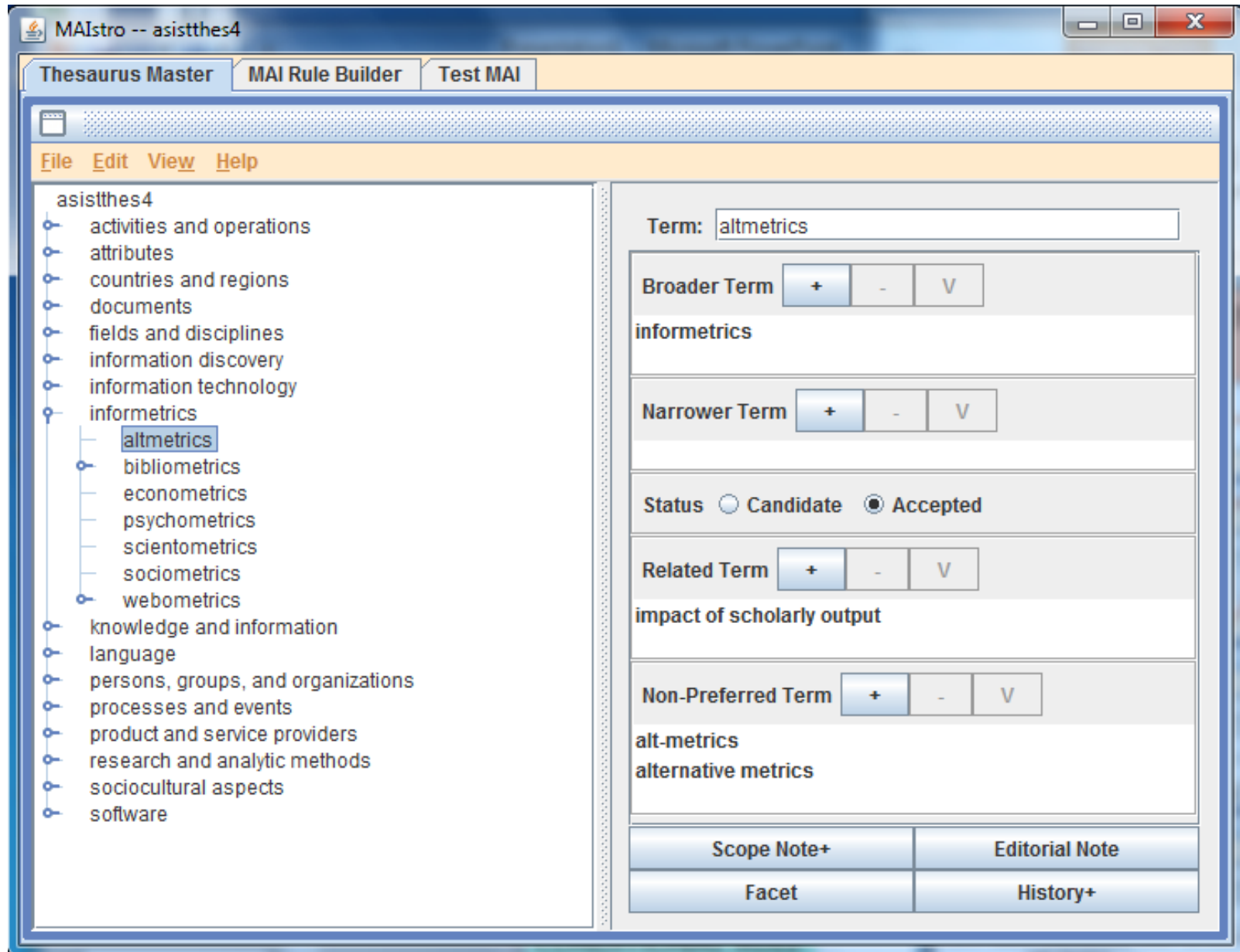
Billing

Showing clues for class

Suggest Clues Search Delete

Type	Clue	Score
Standard		50
<input type="checkbox"/> Standard	Billing Process synonyms languages	45
<input type="checkbox"/> Standard	Billing Services synonyms languages	45
<input type="checkbox"/> Standard	Billing Systems synonyms languages	45
<input type="checkbox"/> Standard	Client Billing synonyms languages	45
<input type="checkbox"/> Standard	Billing synonyms languages	30
<input type="checkbox"/> Class Boost	Accounting & Finance	10
<input type="checkbox"/> Class Boost	47-Accounting	10

Rules Based Auto-Categorization: Data Harmony MAIstro



The image displays four overlapping screenshots of the MAI Rule Builder software interface, illustrating different types of rules that can be created. Each screenshot shows the 'Thesaurus Master' window with tabs for 'Thesaurus Master', 'MAI Rule Builder', and 'Test MAI'. The 'MAI Rule Builder' tab is active in all views.

- Top Screenshot (Identity rule):** The 'Text To Match' field contains 'altmetrics'. An orange arrow points from the text 'Automatically generated rules from term record' to this field. The 'Rule' field is empty.
- Second Screenshot (Synonym rule 1):** The 'Text To Match' field contains 'alt-metrics'. An orange arrow points from the text 'Synonym rule 1' to this field.
- Third Screenshot (Synonym rule 2):** The 'Text To Match' field contains 'alternative metrics'. An orange arrow points from the text 'Synonym rule 2' to this field.
- Bottom Screenshot (Editor-created conditional rule):** The 'Text To Match' field contains 'alternative'. A green arrow points from the text 'Editor-created conditional rule, starts with "IF"' to this field. The 'Rule' field contains the following text:

```
IF (WITH "citation*" AND (AROUND "count*" OR AROUND "number*"))  
    USE altmetrics  
ENDIF
```

© 201

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

The screenshot displays the SAS Content Categorization Studio (Teragram) interface. The window title is "WBG Taxonomy 1 - SAS Content Categorization Studio". The menu bar includes File, Edit, View, Build, Project, Category, Concept, Testing, Document, Server, and Help. The toolbar contains icons for file operations and navigation.

The left pane shows a hierarchical taxonomy tree. The "Education" category is expanded, revealing sub-categories: Abstract, Tertiary Education, Tertiary Education Old Style, and Title. The "Tertiary Education" category is further expanded, showing "test Absgtract", "Test Combo", and "Test Old Style".

The right pane displays a rule editor with the following code:

```
(START_3000,  
(AND,  
  _/add/doc/docty:"Education Sector Review"  
),  
(OR,  
  _/add/doc/display_title:"[Tertiary Education]",  
  _/add/doc/subtopic:"[Tertiary Education]",  
(MINOC_2,_/add/doc/keywd:"[Tertiary Education]"),  
(MINOC_2,_/add/doc/keywordsv2:"[Tertiary Education]"),  
(MINOC_2,_/add/doc/abstracts:"[Tertiary Education]"  
),  
(MINOC_2,_/add/doc/contentTxt:"[Tertiary Education]"  
)  
)) |
```

The bottom of the interface features a toolbar with buttons for "Syntax Check", "Indent", "Text View" (selected), "Tree View", "Load Text...", "Expand Forms", and "Server Query...". Below this is a row of tabs: "Rules", "Testing", "Data", and "Document". At the very bottom, there are tabs for "Taxonomy" and "Dependencies".

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

The screenshot shows the SAS Content Categorization Studio (Teragram) interface. The top menu bar includes View, Build, Project, Category, Concept, Testing, Document, Server, and Help. Below the menu is a toolbar with various icons.

IG Taxonomy 1

English

Categorizer

Top

- Agriculture and Rural Development
 - Agricultural Policy
 - Agricultural Innovation System
 - Agricultural Markets and Risk
 - Agricultural Water Management
 - Agriculture and Climate Change
 - Agriculture and Farming Systems
 - Food Safety
 - Food Security
 - Forestry
 - Land
 - Landscape Approaches
 - Rural Transport
- Debt, Fiscal and Macroeconomic
- Economic Growth
- Education
 - Abstract
 - test Abstract
 - Tertiary Education
 - Test Combo
 - Test Old Style
 - Tertiary Education Old Style
 - Strict Rule
 - Title
 - Energy and Mining
 - Environment
 - Financial and Private Sector

Test File: C:\Image Bank files\imgbankdmp\Education Sector Review\Single Test\Original File 1.xml

Go Stop ...

<?xml version="1.0" encoding="UTF-8" standalone="no"?><doc><field name="id">9919525</field><field name="sourcetype">imagebank</field><field name="itemtype">DOC</field><field name="display_title">Madagascar - **Post primary education**: developing the workforce, shaping the future - transformation of Madagascar's **post-basic education** (2 of 2)</field><field name="bdturl">http://imagebank.worldbank.org/servlet/WDSCContentServer/IW3P/IB/2008/10/08/000333038_20081008031638/Rendered/INDEX/AAA270ESW0VOL110Box334072B01PUBLIC1.bt</field><field name="pdfurl">http://imagebank.worldbank.org/servlet/WDSCContentServer/IW3P/IB/2008/10/08/000333038_20081008031638/Rendered/PDF/AAA270ESW0VOL110Box334072B01PUBLIC1.pdf</field><field name="secd">Public</field><field name="secd_key">82769</field><field name="docdt">2008-09-03T00:00:00Z</field><field name="year">2008</field><field name="irisf">IB</field><field name="irisf_key">545897</field><field name="spaxr">The World</field><field name="spaxr_key">555296</field><field name="lang">English</field><field name="lang_key">120701</field><field name="vemm">Gray cover</field><field name="datee">2008-10-08T00:00:00Z</field><field name="entityid">000333038_20081008031638</field><field name="owner">AFT: Human Development 3 (AFTH3)</field><field name="owner_key">122004</field><field name="docna">Annexes to the main report</field><field name="proid">MG-Madagascar Post Primary Education -- P102241</field><field name="proid_key">908747</field><field name="profler">Mungekar,Santosh Mahadeo</field><field name="profler_key">000333038</field><field name="keywd">Basic Education, Capital Expenditure, choice of employment, Development Finance, Distance education, Earnings, Education Level, Employee, Enrollment, Exchange rates, Expenditure, Expenditures, Higher Education, Investment Climate, Investment Climate Assessment, Junior Secondary, Labor Force, Permanent Employee, Private institutions, private schools, Proprietorship, Public Expenditures, Public Expenditures on Education, public schools, sales, Schools, scientific research, secondary education, Senior, Teachers, vocational education, Working Hours</field><field name="abstracts">The main purpose of this report is to provide analytical inputs for the development of **post-basic education** reforms. Specifically, the report identifies and prioritizes: (i) the need for change in the structure, content and delivery of Madagascar's **post-basic education** and training system, and (ii) the key reforms in financing, governance and sub-sector management required to support changes to the structure, content and delivery of the post-basic system. The Madagascar Action Plan (MAP) outlines an ambitious development strategy, focusing on promoting investment in high growth sectors and regional development. If successful, it will change the demand for skills in fundamental ways. Since 2005, foreign direct investment has increased rapidly. Madagascar's core challenges and the window of opportunity provided by the implementation of basic education reform imply that reform must improve the quality and relevance of **post-basic education**, while putting cost-effective mechanisms for expanding access in place. Post-basic reform should not focus exclusively on a massive expansion of the existing post-basic system. Instead, successful reform will: (i) focus first on improving educational content (structure, curriculum, teaching, and process) and linkages with the economy; (ii) increase coverage, cost-effectively; and (iii) strengthen the enabling framework for reform (governance, finance, and sub-sector management). Reforms aimed at improving educational content must accomplish three objectives: (i) meet the skilled labor requirements of the economy's key growth sectors, in the short to medium term; (ii) gradually build professional capabilities in the key growth sectors, also in the short to medium term; and (iii) help youth to develop the knowledge, skills and attitudes - employability skills - that will allow them to participate in and adapt to the changing labor market over time.</field><field name="subtopic">Tertiary Education</field><field name="subtopic_key">672928</field><field name="subtopic">Access to Finance</field><field name="subtopic_key">880275</field><field name="subtopic">Access & Equity in Basic Education</field><field name="subtopic_key">672932</field><field name="subtopic">Education For All</field><field name="subtopic_key">761316</field><field name="subtopic">m</field><field name="subtopic_key">883293</field><field name="docty">Education Sector Review</field><field name="docty_key">904569</field><field name="admreg">Africa</field><field name="admreg_key">119222</field><field name="admreg">Africa</field><field name="admreg_key">119222</field><field name="count">Madagascar</field><field name="count_key">82548</field><field name="accnm">334072B</field><field name="repro">AAA27</field><field name="repro_key">9919517</field><field name="volnb">2 of 2</field><field name="trustfund">TF056289-EFA FTI EDUCATION PROGRAM DEVELOPMENT FUND - AFRICA</field><field name="trustfund_key">885881</field><field name="trustfund">TF056377-NORWEGIAN POST-PRIMARY EDUCATION TRUST FUND</field><field name="trustfund_key">887380</field><field name="repro">Madagascar - **Post primary education**: developing the workforce, shaping the future - transformation of Madagascar's **post-basic education**</field><field name="teratopic">Education</field><field name="teratopic_key">644301</field><field name="teratopic">Finance and Financial Sector Development</field><field name="teratopic_key">644297</field><field name="proft">IB Document Template</field><field name="proft_key">641021</field><field name="sectr">Education</field><field name="sectr_key">369165</field><field name="subsc">Vocational training</field><field name="subsc_key">646031</field><field name="subsc">Tertiary education</field><field name="subsc_key">646030</field><field name="subsc">Secondary education</field><field name="subsc_key">646029</field><field name="taskm">Bashir,Sajitha</field><field name="Sector_List">SECTOR1\$! EV~SECTOR2\$! ET~SECTOR3\$! ES</field><field name="taskm_key">459240</field><field name="theme">Public expenditure, financial management and procurement</field><field name="theme_key">646089</field><field name="theme">Education for the knowledge economy</field><field name="theme_key">646129</field><field name="theme">Education for all</field><field name="theme_key">646128</field><field name="prdlm">Economic and Sector Work</field><field name="prdlm_key">646128</field></doc></xml>

PASS TEST

Selected category
All categories
All categories and all concepts

Browser View
View Rule Matches

Rules Testing Data Document

Rules Based Auto-Categorization: Smartlogic Semaphore

The screenshot displays the 'FinancialDemo - Semaphore Ontology Manager' application. The interface is divided into a left-hand tree view and a right-hand workspace.

Left Panel (Tree View):

- Content type
- Location
- Organization
 - Apple
 - Google
- Person
- Product
- Sector
- Topic
 - Credit Topics
 - Economics Topics
 - Balance of Payments
 - Commodities
 - Employee Pensions
 - Equities (Economics)
 - Financial Variables
 - FMCJ
 - Forecasts (Economics)
 - Growth
 - Inflation
 - Interest & FX Rates
 - Labour Market
 - Employment
 - Productivity
 - Unemployment
 - Unit Labour Costs
 - Wages & Salaries
 - No Relevant Topic
 - Politics
 - Surveys
- Events
- FX Topics
- Interest Rates Topics

Right Panel (Workspace):

The workspace shows the selected entity 'Apple' with the class 'Organisation'.

Hierarchical Panel:

Type	Term
BT	Organization

Associative Panel:

Type	Term
in Industry	Electronic Equipment
Management	Tim Cook
Produces	Ipad
Produces	Iphone
Produces	Ipod
Produces	mac

Equivalence Panel:

Type	Term
Precluded by	Big Apple
UF	AAPL

Bottom Tabs:

Hierarchical | Alphabetical | Hierarchy Select | Removed | Relationships | Family | Term Information | Term Attributes | Semaphore Settings | Properties

Rules Based Auto-Categorization: Smartlogic Semaphore

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

The screenshot displays the Smartlogic Semaphore Rule Editor interface. The main window shows a list of rules for the file "Apple Inc_18903N257299339bNodryRrz3RXyBa3gcA7_Issuer.xml". The rules are organized into a tree structure under the heading "BODY RULES FOR PTS/NPTs". The rules include "combine", "phrase", "with", "near", and "any" rules, each with associated weights and data fields.

Rule Type	Weight	Data	Stem	Case	Not
BODY RULES FOR PTS/NPTs					
combine	100		0	0	0
All terms - body"					
combine	100		0	0	0
phrase	40		0	1	0
text	100	Big	0	0	1
text	100	Apple	0	1	0
text	100	Inc	0	0	1
with	5		0	0	0
phrase	5		0	1	0
text	100	Big	0	0	1
text	100	Apple	0	1	0
text	100	Inc	0	0	1
text	100	{field_start}	0	0	0
phrase	5		0	1	0
text	100	Big	0	0	1
text	100	Apple	0	1	0
text	100	Inc	0	0	1
near	5		0	0	0
any	100		0	0	0
text	100	Price Target	0	0	0
text	100	cash flow	0	0	0
phrase	5		0	1	0
text	100	Big	0	0	1
text	100	Apple's	0	1	0
text	100	Inc	0	0	1
near	5		0	0	0
any	100		0	0	0
text	100	Price Target	0	0	0

Foreach: 1; Weight: 5; Count: 2

The "Rule Editor" dialog box is open, showing the configuration for a "near" rule. The "Rule Type" is set to "near". The "Weight" is 5, "Foreach" is 1, "Label" is empty, "Count" is 2, "Field" is "body", "Language" is empty, and "Punctuation" is empty. The "near rule" description is visible in the bottom panel of the dialog.

near rule

A **near** rule identifies a set of words occurring near each other in the text of a document. Each word of the near contents should be entered as a **text** rule and inserted as a child rule of the parent **near** rule. The **near** rule will be triggered if all of its child rules are triggered and are within the specified number of words in the document. Unlike the very similar **phrase** rule the order of the children in the document does not matter. To allow unspecified

OK Cancel

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		
Concurrent Terms			Document ↕	First Run ↕	Second Run ▾	Δ ↕	
Term Name ↕	First Run ↕	Second Run ▾	Δ ↕	XML/x-New cash pledge for shires roads...	0.82	0.82	0
Road safety	8	23	15	XML/x-Funding for roads.xml	0.77	0.77	0
Police	4	13	9	XML/x-Stay alert - and stay alive.xml	0.72	0.72	0
Toll Roads	3	13	10	XML/x-			
Victoria	2	12	10	XML/x-			
Motor vehicles	0	11	11	XML/x-			
Drink driving	0	7	7	XML/x-			
Companies	PoV	13		XML/x-			
Prices	PoV	23		XML/x-			
Groups	PoV	10		XML/x-			
Interest rates	PoV	15		XML/x-			
Melbourne	PoV	11		XML/x-			
Sport	PoV	10		XML/x-			
Stocks and Shares	PoV	17		XML/x-			
Domestic animals	PoV	30		XML/x-			
Housing	PoV	9		XML/x-			
Farmers	PoV	10		XML/x-			
COMMUNITY AND SOCIETY	PoV	4		XML/x-			
Universities				XML/x-			
Investment				XML/x-			
Car parks				XML/x-			
Women				XML/x-			
Children	PoV	12		XML/x-			

Document level metadata				
System information and metadata				
Classification results				
Class	Category name	ID	Score	Key
PoV	Bridges	POV698	0.47	10405
PoV	Infrastructure	POV227	0.37	50626
PoV	Planning	POV880	0.38	74910
PoV	Road funding	POV2323	0.56	87022
PoV	Roads	POV376	0.77	87495
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 Bridges fund. ¶				
P : "This funding is delivery of a 2010 Baillieu-Ryan Government election commitment to small country roads and rejuvenate run down country roads and bridges as this government recognises the shortfalls 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 bridge assets throughout the shire." The \$1 million has been allocated to Benalla Rural City to provide 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 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.

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:

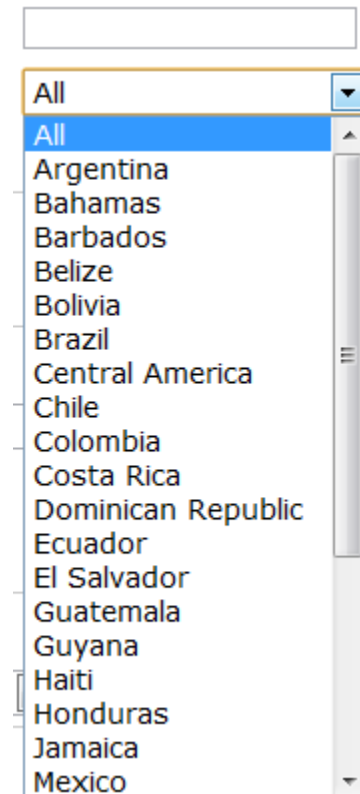
- Term Lists/Pick lists
- Synonym Rings
- Authority Files
- Taxonomies
 - Hierarchical
 - Faceted
- Thesauri
- Ontologies (going beyond a controlled vocabulary)

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

Taxonomy Definitions & Types

Term List/Pick List

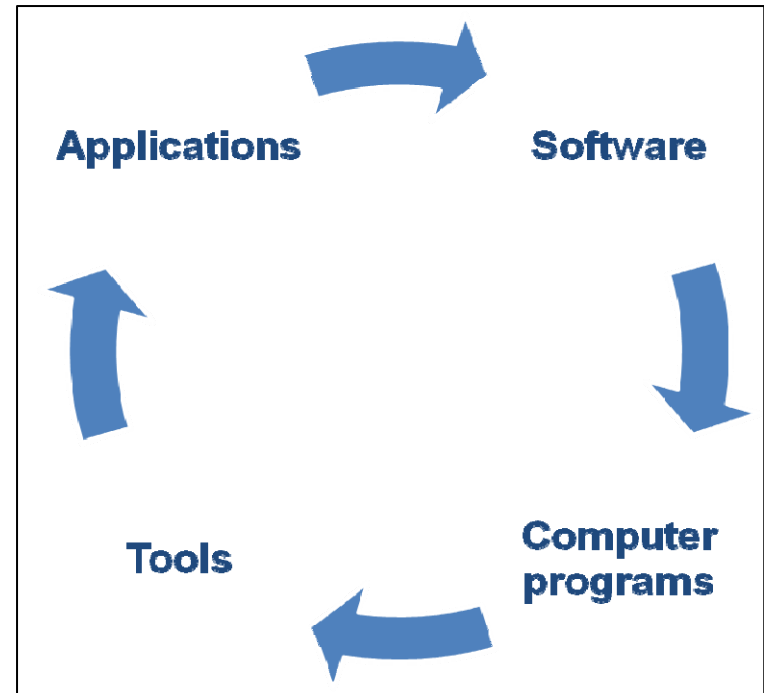
- A simple list of terms
- Lacking synonyms, usually short enough for browsing
- Often displayed in drop-down scroll boxes



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

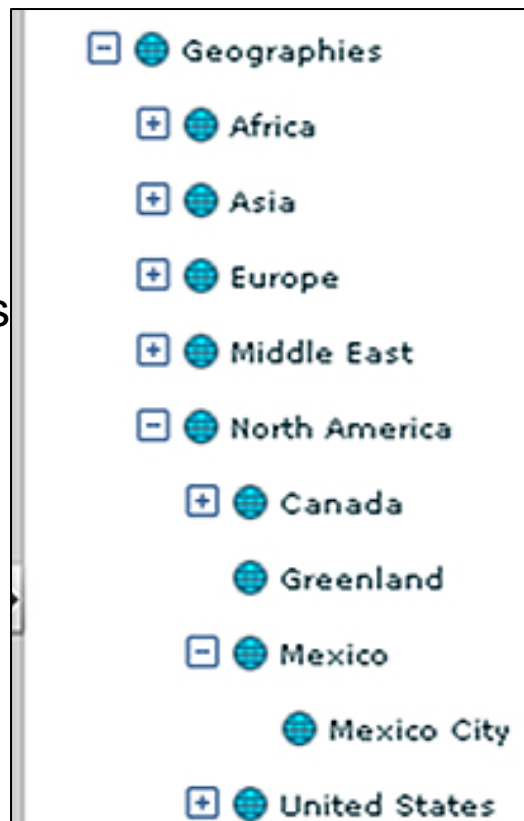
Taxonomy Definitions & Types

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.

Narrow Your Search

+ Search Within Results

- Locations Served
Arizona
Arkansas
California - North
California - South
Colorado
[+] More

+ Search Within # Miles

- Company Type
Manufacturers
Custom Manufacturers
Distributors
Service Companies
Manufacturers' Reps

+ Certifications

- Ownership
Minority-Owned
Woman-Owned
Veteran-Owned
...

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

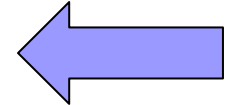
Thesaurus term entry example

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

Users may enter:

Oil industry

Oil & gas industry

Oil & gas industries

Petroleum industry

Taxonomy contains all synonyms:

Oil industry

Oil & gas industry

Oil and gas industry

Oil & gas industries

Oil and gas industries

Petroleum industry

Oil companies

Big oil

Oil producers

Petroleum companies

Text may contain:

Oil and gas industry

Oil companies

Big oil

Oil producers



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.

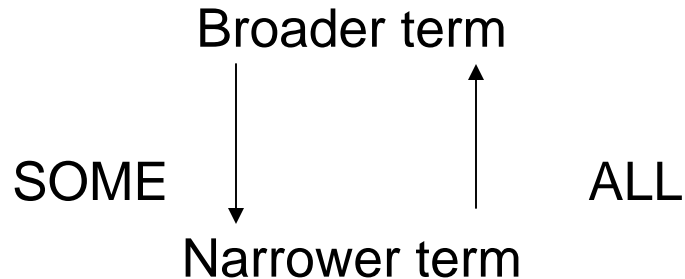
Hierarchy Best Practices

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

> German

Financial services

> 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>



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