Automated Metadata Generation and the Critical Role of Catalogers and Indexers in Technical Services of the Future

Denise A. D. Bedford, Ph.d.

Goodyear Professor of Knowledge Management

Kent State University

Kent Ohio

Technical Services Renaissance, Ohio Valley Technical Services Group

Miami University, May 16, 2011

Five Questions

- Making the Case: Catalogers as Knowledge Engineers
- The Future is Semantic
- Automated Metadata Generation – What is it and How does it Work?
- The Role of Human Knowledge in Automated Metadata Generation

CATALOGERS AS KNOWLEDGE ENGINEERS

Little History

- 1979 Director’s Office at Stanford University Libraries

- Cataloging Backlog Analysis and Resolution

- Workflow Investigations, Technical Services Organizational Structure Review …….

- Only way to reduce the backlog and prevent its continual growth was to reduce the unit processing time for cataloging

Little More History …..

- University of California Berkeley Ph.d. program and Stanford courses in operations research, artificial intelligence and natural language processing, and programming-engineering systems

- U. C. Systemwide Administration saw the same continuous growth in number of resources requiring metadata – first online catalogs

- Advances in semantic analysis methods (early 1980's through 2000's) and natural language processing

- Increased value of and demand for metadata to support information management and access due to the emerging semantic web

Early Fascination with Natural Language Processing

- In the 1980s, I was part of the community that looked to natural language processing to produce significant improvements in all aspects of information management and access

- It soon became clear to me, though, that most of the technologies were not going to get us where we needed to be

- Most of the technologies either used a statistical approach or took a simplistic approach to leveraging Knowledge Organization Systems

- Different approach was called for ….. What was that approach?

Today's Dynamic Information Landscape

- Demand for metadata is increasing exponentially today
  - for richer and more granular metadata
  - more resources and more types of resources to proc ess

- Personalization is also a growing factor

- Focus on "content" not just the "package" – which brings design opportunities

- Knowledge is not static – neither is any reference source or knowledge base – need to continuously update our sources

Meeting the Challenge

- Automated Metadata Generation allows us to:
  - Increase number of resources that have metadata
  - Increase the availability of metadata – at the whole and also at the part level
  - Increase the number of values for metadata attributes
  - Increase the number of attributes (i.e., access points)
  - Decrease the time devoted to creating metadata (avg. time from 20 mins. to 2 seconds)
  - Improve the quality and consistency of metadata generated
  - Meet the increasing demand for personalized views of information

But, It Doesn't Happen "Automagically"

- Each productive use of technology requires use of existing human knowledge - there is no such thing as a technology that works well "automagically" without human training or design

- And, no single technology is suited to any or all knowledge processing challenges – each knowledge processing challenge requires that we stop and think about how humans do the task – in order to model how the technology will support it

- Beware of the "I have a hammer, so everything looks like a nail" syndrome

- There are also different levels of support – some solutions may be full automated, whereas others may simply provide assistance to the person performing the task

The Cost – Teaching Technology to be Smart

- Technology can only behave intelligently – by human standards – when it has human intelligence to work with. Just because technology produces a result doesn't mean that it is a good result

- Challenge we face in making technology smart is figuring out (1) how to teach technology what we know and (2) how we think about things

- Artificial intelligence, psychology, philosophy, communications, education – all have contributed to our understanding of what technology is and is not capable of doing

- People share what they know, express what they know and record what they know using language – to process information, we need to start at the point of teaching technology how to understand language

Part 2

THE FUTURE IS SEMANTIC

Semantic Analysis

- Semantic only means that there is some “meaningful” and “understandable” approach involved to solving the problem – can be performed by people and machines

- Relies on formal models or representations of knowledge of language and leverages knowledge of phonology, phonetics, morphology, syntax, semantics, pragmatics and discourse

- Formal models used to capture knowledge include state machines, formal rule systems, logic and probabilistic models

- The foundations of technology based semantic analysis lie in computer science, linguistics, mathematics, electrical engineering and psychology

It’s All About Semantic Analysis

- Good automated metadata generation is grounded in quality semantic analysis

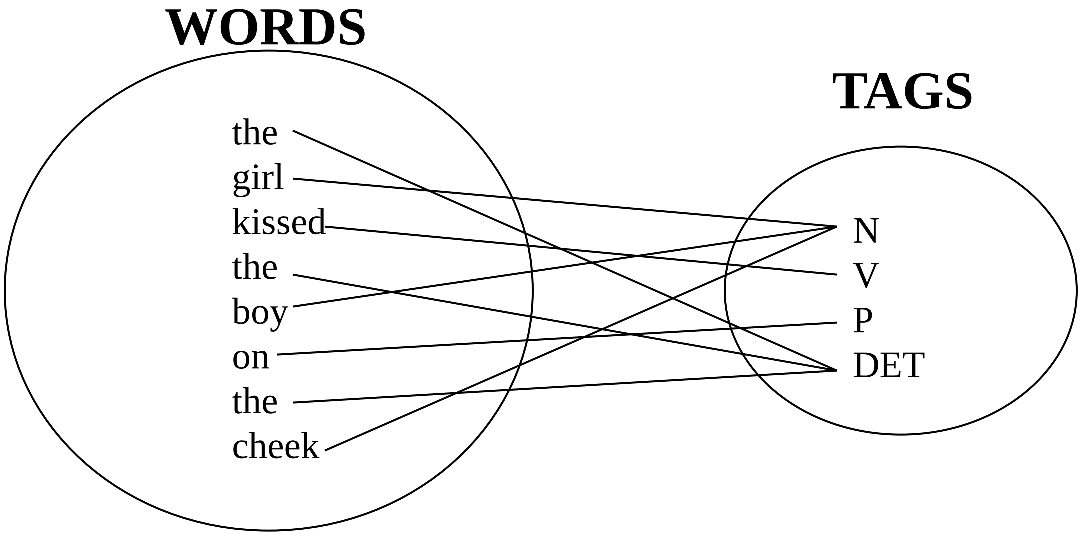
- Semantic analysis can be performed by both people and machines. In both cases, the problem being analyzed is by definition defined by a human expert.

- Always model the human process
  - People have a rich store of linguistic and domain knowledge to draw upon
  - Computers need to be able to have all of that linguistic and domain knowledge encoded and also the rules for using that knowledge

Step 1: Natural Language Processing

This is what a computer does to get to the level of understanding where it can take and act upon our instructions

“The process of assigning a part-of-speech or other lexical class marker to each word in a corpus” [or text] (Jurafsky and Martin)



Part of Speech Tagging

- In order to POS tag content, we need to have a framework or set of tags

- The tagset should include all possible combinations of category values for a given language. A tagset is generally represented by a string of letters or digits:
  - NNS (gen. noun, plural)
  - AAMP3----2A---- (gen. Adj., Masc., Pl., 3rd case (dative), comparative (2nd degree of comparison), Affirmative (no negation))

- Sample tagsets include those developed at Brown, Penn, Multext

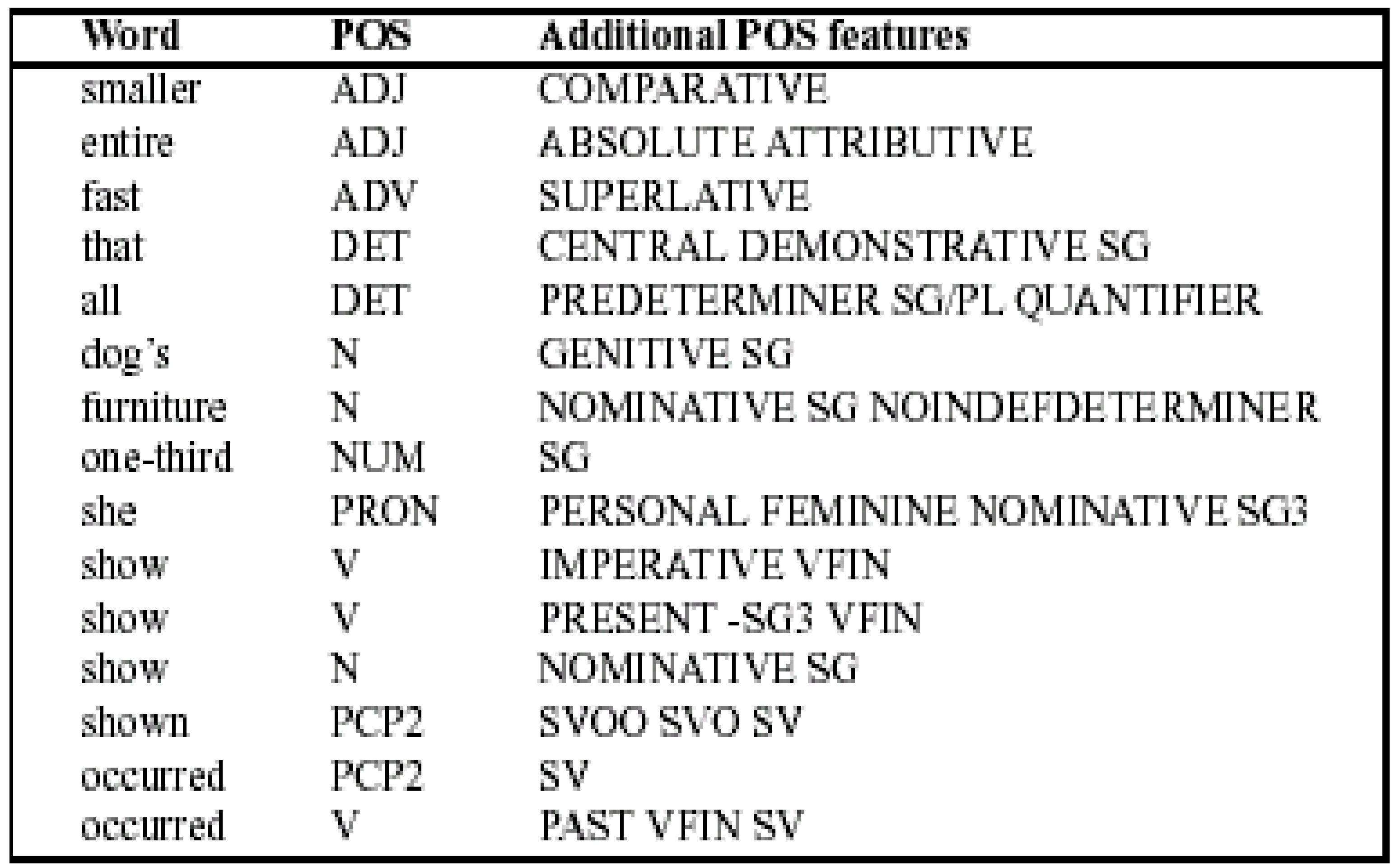
Xerox Tagset

WORD	LEMMA	TAG
the	the	+DET
girl	girl	+NOUN
kissed	kiss	+VPAST
the	the	+DET
boy	boy	+NOUN
on	on	+PREP
the	the	+DET
cheek	cheek	+NOUN

From: http://www.xrce.xerox.com/competencies/content-analysis/fsnlp/tagger.en.html

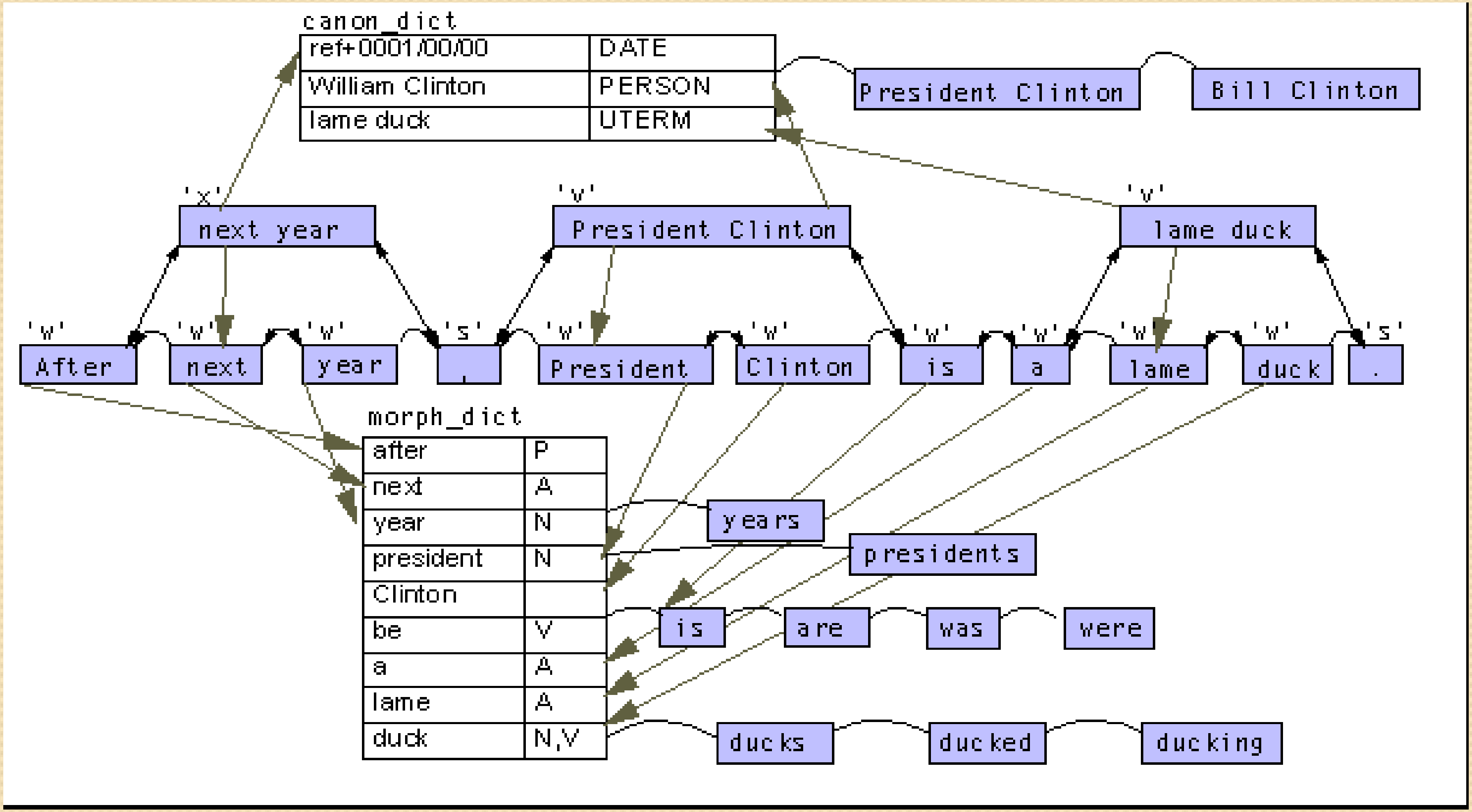
ENGTWOL Lexicon

http://www.lingsoft.fi/cgi-bin/engtwol



Word	POS	Additional POS features
smaller	ADJ	COMPARATIVE
entire	ADJ	ABSOLUTE ATTRIBUTIVE
fast	ADV	SUPERLATIVE
that	DET	CENTRAL DEMONSTRATIVE SG
all	DET	PREDETERMINER SG/PL QUANTIFIER
dog's	N	GENITIVE SG
furniture	N	NOMINATIVE SG NOINDEFDETERMINER
one-third	NUM	SG
she	PRON	PERSONAL FEMININE NOMINATIVE SG3
show	V	IMPERATIVE VFIN
show	V	PRESENT -SG3 VFIN
show	N	NOMINATIVE SG
shown	PCP2	SVOO SVO SV
occurred	PCP2	SV
occurred	V	PAST VFIN SV

POS Tagging Example



canon_dict

ref+0001/00/00	DATE
William Clinton	PERSON
lame duck	UTERM

morph_dict

after	P
next	A
year	N
president	N
Clinton	
be	V
a	A
lame	A
duck	N,V

Step 2: Building the Knowledge Base(s)

- Catalogers use many different sources of knowledge to make decisions, to reason about issues, to determine what next step to take in the process, and even when to discard knowledge

- A cataloger's underlying tacit knowledge must be integrated into a system that generates metadata automatically if the process is to be performed as effectively by technology as by a person

- The design challenge here is a significant one – simply representing a word, or a concept or linking concepts in a structure does not assume it can be effectively used by a computer – neither is simply plugging in a thesaurus or classification scheme the same as a "cataloger's brain"

How People Classify

- Let’s go back to the most important question – how does a cataloger do it?
- First, we develop knowledge of the classification scheme to which we’re classifying - the better a person’s knowledge of the scheme and the better their knowledge of the object, the better judgment they can make
- Second, we analyze the object that we’re classifying
- Third, we make a judgment as to the best fit of the object we’re classifying to all the classes that are available to us –

Caution About Some Technologies

- Rule based classification implies that we have a scheme and defined classes to which to assign entities or objects

- This is a different process than defining classes to constitute a classification scheme – most of the tools do this today

- Much of the "semantic analysis" literature focuses on how to define classes from a set of information – deductively – and then to classify the entities in that set back to the scheme

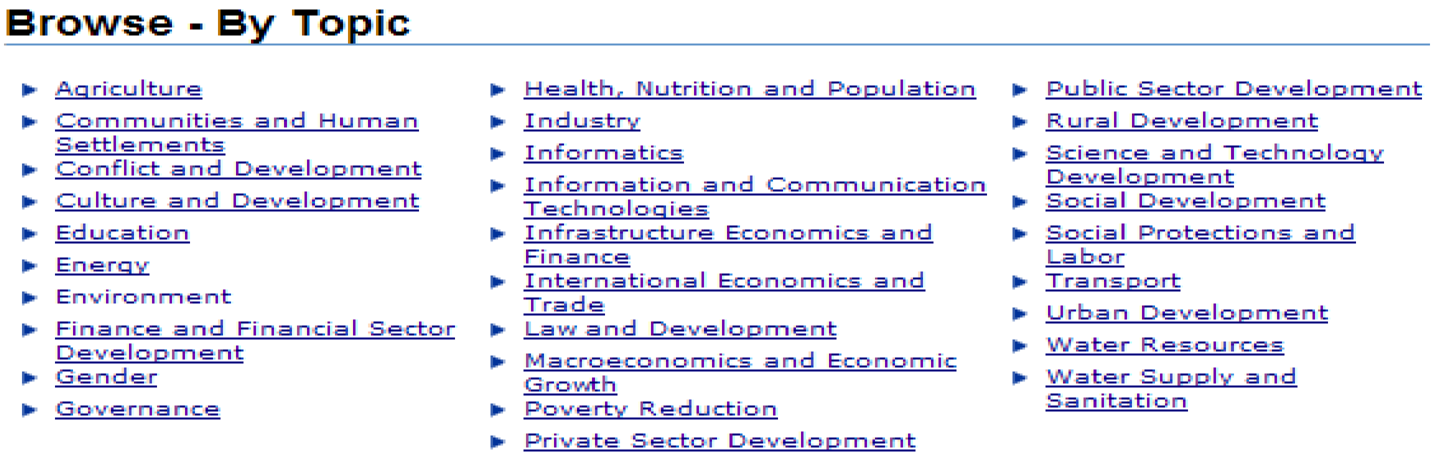
How a Machine Selects a Class

- From the choices we give them, based on what we tell them about the choices, and the rules we give them to make the selection

- They will choose poorly,
  - if we give them a poorly defined or unbalanced scheme
  - if we tell them nothing or very little about the classes
  - If the manual rules are not rigorous

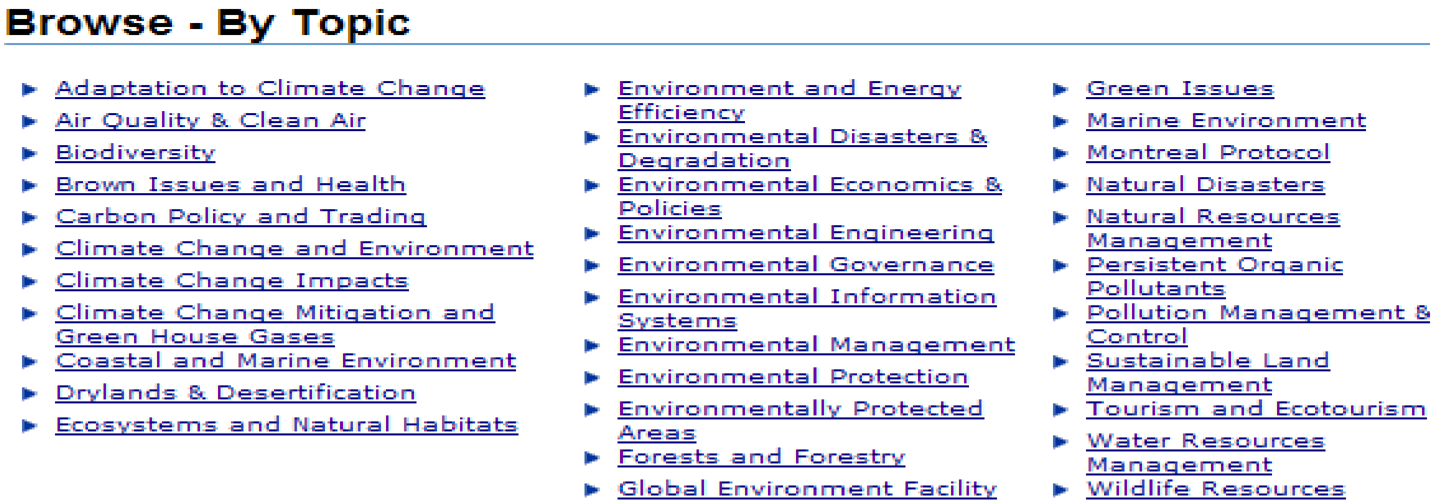
- You may be surprised to find how often a cataloger is subconsciously compensating for a poorly formed classification scheme…..

A Real Life Example:
Topic Classification Scheme



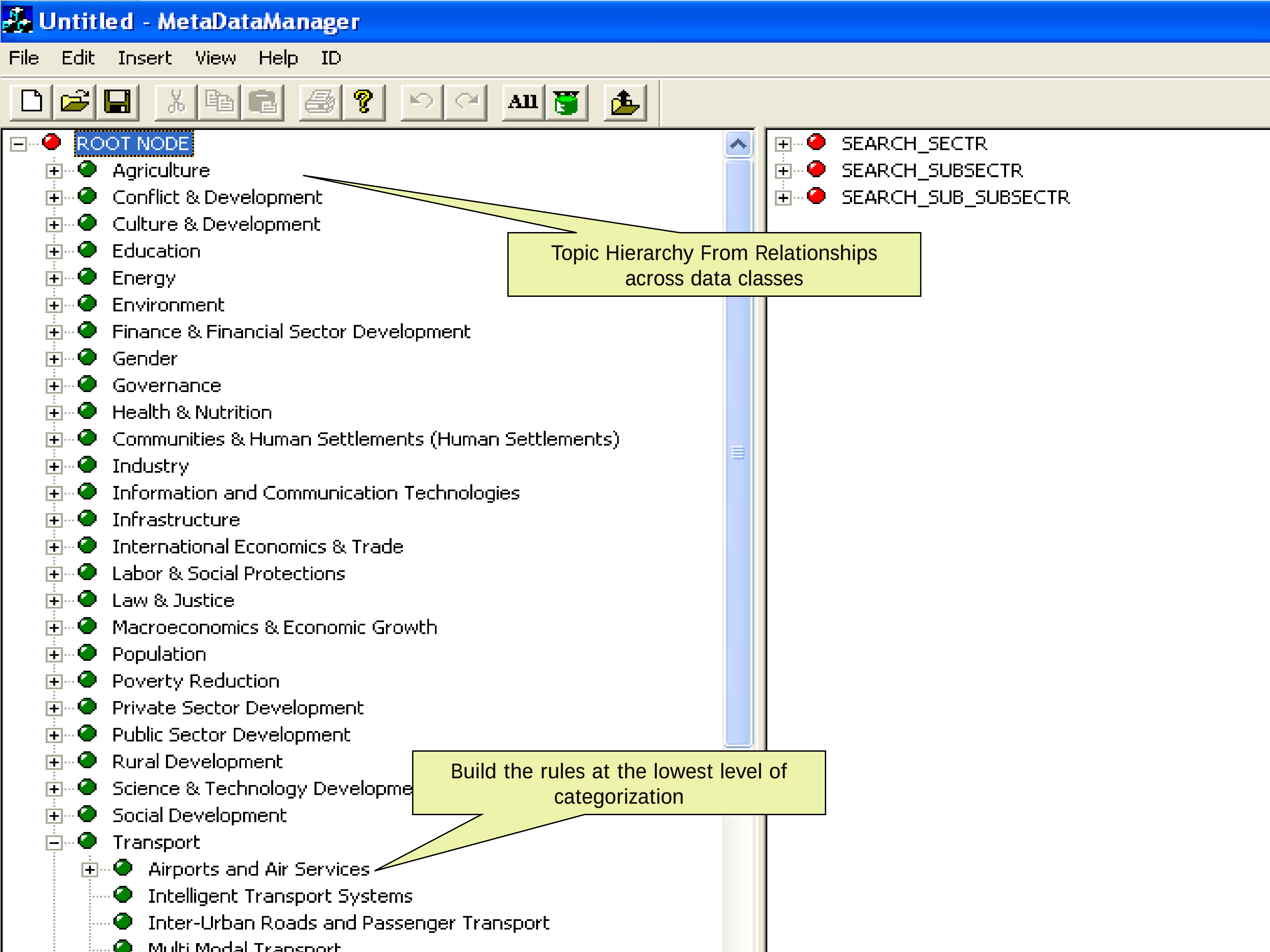
Browse - By Topic

- Agriculture
- Communities and Human Settlements
- Conflict and Development
- Culture and Development
- Education
- Energy
- Environment
- Finance and Financial Sector Development
- Gender
- Governance
- Health, Nutrition and Population
- Industry
- Informatics
- Information and Communication Technologies
- Infrastructure Economics and Finance
- International Economics and Trade
- Law and Development
- Macroeconomics and Economic Growth
- Poverty Reduction
- Private Sector Development
- Public Sector Development
- Rural Development
- Science and Technology Development
- Social Development
- Social Protections and Labor
- Transport
- Urban Development
- Water Resources
- Water Supply and Sanitation



Browse - By Topic

- Adaptation to Climate Change
- Air Quality & Clean Air
- Biodiversity
- Brown Issues and Health
- Carbon Policy and Trading
- Climate Change and Environment
- Climate Change Impacts
- Climate Change Mitigation and Green House Gases
- Coastal and Marine Environment
- Drylands & Desertification
- Ecosystems and Natural Habitats
- Environment and Energy Efficiency
- Environmental Disasters & Degradation
- Environmental Economics & Policies
- Environmental Engineering
- Environmental Governance
- Environmental Information Systems
- Environmental Management
- Environmental Protection
- Environmentally Protected Areas
- Forests and Forestry
- Global Environment Facility
- Green Issues
- Marine Environment
- Montreal Protocol
- Natural Disasters
- Natural Resources Management
- Persistent Organic Pollutants
- Pollution Management & Control
- Sustainable Land Management
- Tourism and Ecotourism
- Water Resources Management
- Wildlife Resources



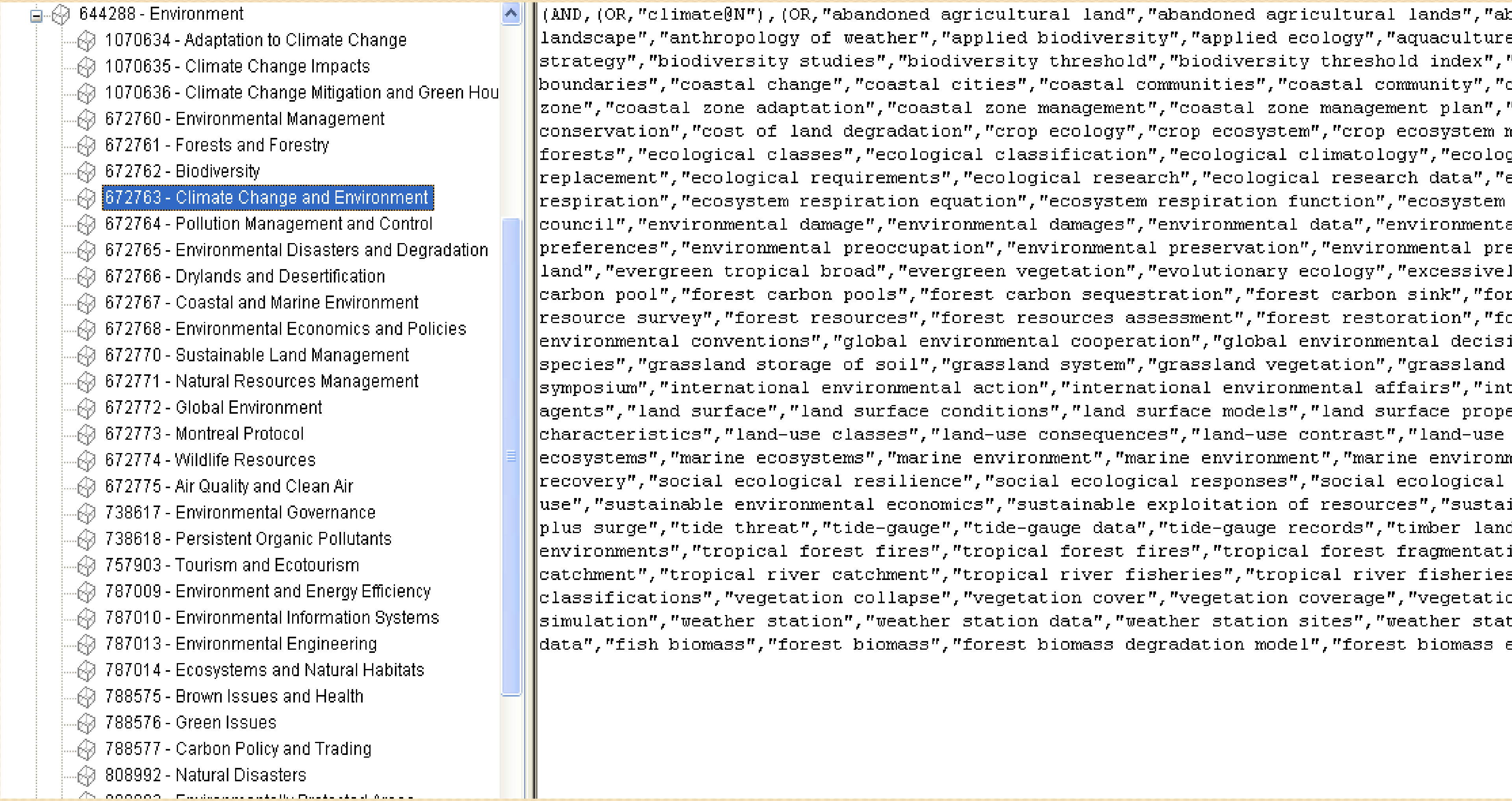
- ROOT NODE
  - Agriculture
  - Conflict & Development
  - Culture & Development
  - Education
  - Energy
  - Environment
  - Finance & Financial Sector Development
  - Gender
  - Governance
  - Health & Nutrition
  - Communities & Human Settlements (Human Settlements)
  - Industry
  - Information and Communication Technologies
  - Infrastructure
  - International Economics & Trade
  - Labor & Social Protections
  - Law & Justice
  - Macroeconomics & Economic Growth
  - Population
  - Poverty Reduction
  - Private Sector Development
  - Public Sector Development
  - Rural Development
  - Science & Technology Developme
  - Social Development
  - Transport
    - Airports and Air Services
    - Intelligent Transport Systems
    - Inter-Urban Roads and Passenger Transport
    - Multi Modal Transport

- SEARCH_SECTR
- SEARCH_SUBSECTR
- SEARCH_SUB_SUBSECTR

Topic Hierarchy From Relationships across data classes

Build the rules at the lowest level of categorization

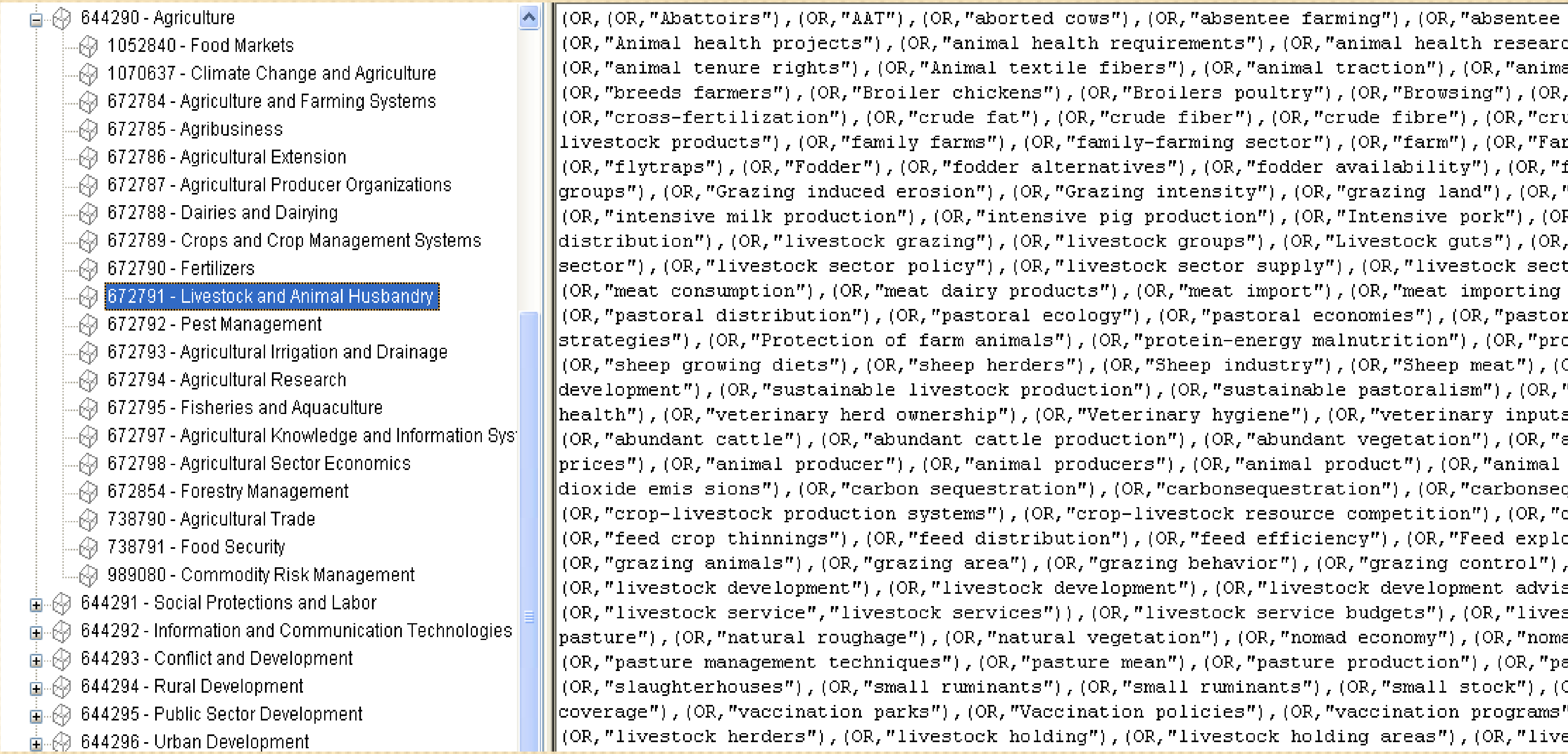
Sample Definition of Subclass Climate Change and Environment



- 644288 - Environment
  - 1070634 - Adaptation to Climate Change
  - 1070635 - Climate Change Impacts
  - 1070636 - Climate Change Mitigation and Green Hou
  - 672760 - Environmental Management
  - 672761 - Forests and Forestry
  - 672762 - Biodiversity
  - 672763 - Climate Change and Environment
  - 672764 - Pollution Management and Control
  - 672765 - Environmental Disasters and Degradation
  - 672766 - Drylands and Desertification
  - 672767 - Coastal and Marine Environment
  - 672768 - Environmental Economics and Policies
  - 672770 - Sustainable Land Management
  - 672771 - Natural Resources Management
  - 672772 - Global Environment
  - 672773 - Montreal Protocol
  - 672774 - Wildlife Resources
  - 672775 - Air Quality and Clean Air
  - 738617 - Environmental Governance
  - 738618 - Persistent Organic Pollutants
  - 757903 - Tourism and Ecotourism
  - 787009 - Environment and Energy Efficiency
  - 787010 - Environmental Information Systems
  - 787013 - Environmental Engineering
  - 787014 - Ecosystems and Natural Habitats
  - 788575 - Brown Issues and Health
  - 788576 - Green Issues
  - 788577 - Carbon Policy and Trading
  - 808992 - Natural Disasters
  - 808993 - Environmentally Protected Areas

```
(AND,(OR,"climate@N"),(OR,"abandoned agricultural land","abandoned agricultural lands","ab
landscape","anthropology of weather","applied biodiversity","applied ecology","aquaculture
strategy","biodiversity studies","biodiversity threshold","biodiversity threshold index","
boundaries","coastal change","coastal cities","coastal communities","coastal community","c
zone","coastal zone adaptation","coastal zone management","coastal zone management plan","
conservation","cost of land degradation","crop ecology","crop ecosystem","crop ecosystem m
forests","ecological classes","ecological classification","ecological climatology","ecolog
replacement","ecological requirements","ecological research","ecological research data","e
respiration","ecosystem respiration equation","ecosystem respiration function","ecosystem
council","environmental damage","environmental damages","environmental data","environmenta
preferences","environmental preoccupation","environmental preservation","environmental pre
land","evergreen tropical broad","evergreen vegetation","evolutionary ecology","excessivel
carbon pool","forest carbon pools","forest carbon sequestration","forest carbon sink","for
resource survey","forest resources","forest resources assessment","forest restoration","fo
environmental conventions","global environmental cooperation","global environmental decisi
species","grassland storage of soil","grassland system","grassland vegetation","grassland
symposium","international environmental action","international environmental affairs","int
agents","land surface","land surface conditions","land surface models","land surface prope
characteristics","land-use classes","land-use consequences","land-use contrast","land-use
ecosystems","marine ecosystems","marine environment","marine environment","marine environm
recovery","social ecological resilience","social ecological responses","social ecological
use","sustainable environmental economics","sustainable exploitation of resources","sustai
plus surge","tide threat","tide-gauge","tide-gauge data","tide-gauge records","timber land
environments","tropical forest fires","tropical forest fires","tropical forest fragmentati
catchment","tropical river catchment","tropical river fisheries","tropical river fisheries
classifications","vegetation collapse","vegetation cover","vegetation coverage","vegetatio
simulation","weather station","weather station data","weather station sites","weather stat
data","fish biomass","forest biomass","forest biomass degradation model","forest biomass e
```

Sample Definition of Subclass Livestock and Animal Husbandry



644290 - Agriculture
- 1052840 - Food Markets
- 1070637 - Climate Change and Agriculture
- 672784 - Agriculture and Farming Systems
- 672785 - Agribusiness
- 672786 - Agricultural Extension
- 672787 - Agricultural Producer Organizations
- 672788 - Dairies and Dairying
- 672789 - Crops and Crop Management Systems
- 672790 - Fertilizers
- 672791 - Livestock and Animal Husbandry
- 672792 - Pest Management
- 672793 - Agricultural Irrigation and Drainage
- 672794 - Agricultural Research
- 672795 - Fisheries and Aquaculture
- 672797 - Agricultural Knowledge and Information Sys
- 672798 - Agricultural Sector Economics
- 672854 - Forestry Management
- 738790 - Agricultural Trade
- 738791 - Food Security
- 989080 - Commodity Risk Management

644291 - Social Protections and Labor

644292 - Information and Communication Technologies

644293 - Conflict and Development

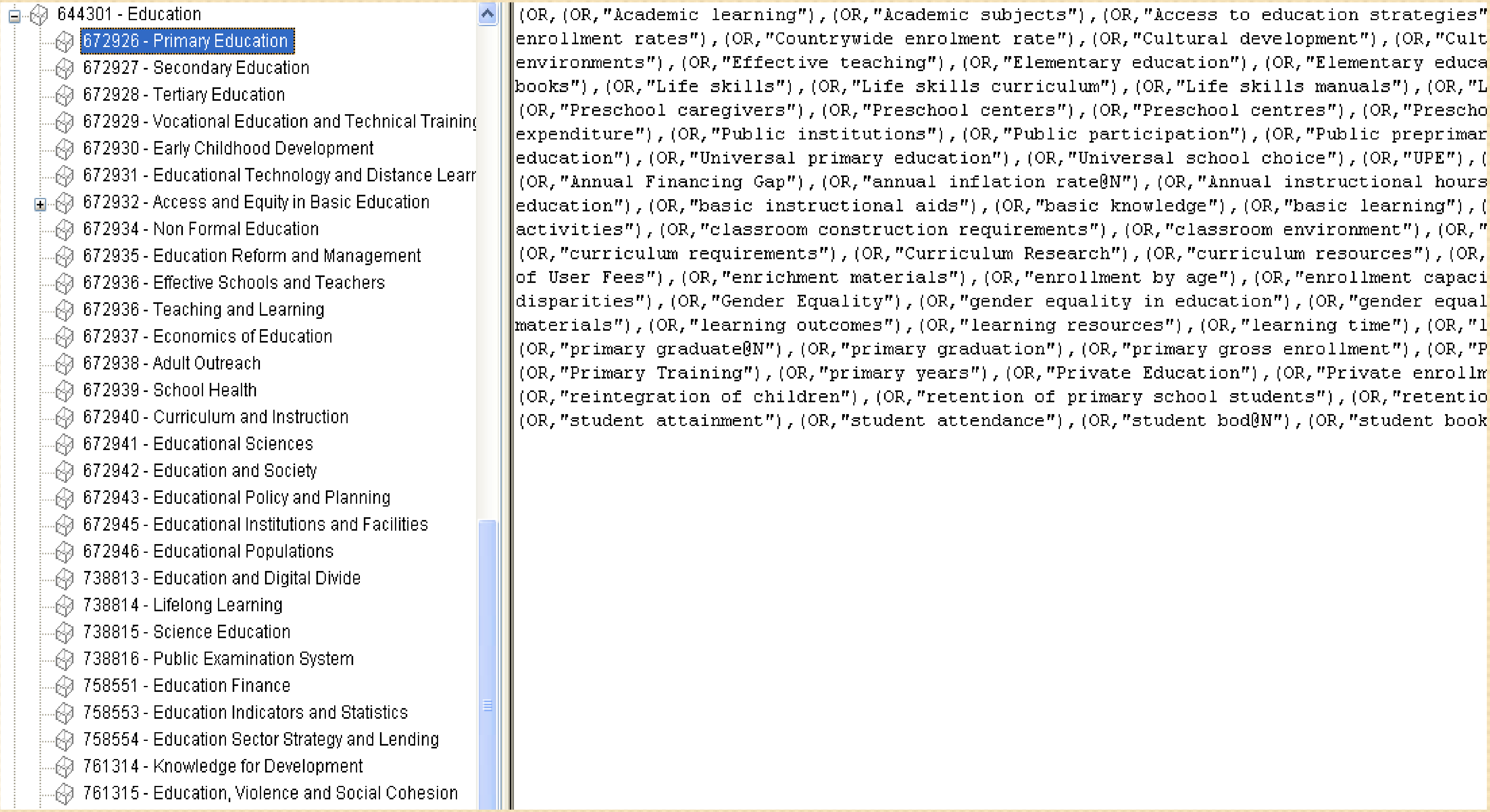
644294 - Rural Development

644295 - Public Sector Development

644296 - Urban Development

```
(OR,(OR,"Abattoirs"),(OR,"AAT"),(OR,"aborted cows"),(OR,"absentee farming"),(OR,"absentee
(OR,"Animal health projects"),(OR,"animal health requirements"),(OR,"animal health researc
(OR,"animal tenure rights"),(OR,"Animal textile fibers"),(OR,"animal traction"),(OR,"anima
(OR,"breeds farmers"),(OR,"Broiler chickens"),(OR,"Broilers poultry"),(OR,"Browsing"),(OR,
(OR,"cross-fertilization"),(OR,"crude fat"),(OR,"crude fiber"),(OR,"crude fibre"),(OR,"cru
livestock products"),(OR,"family farms"),(OR,"family-farming sector"),(OR,"farm"),(OR,"Far
(OR,"flytraps"),(OR,"Fodder"),(OR,"fodder alternatives"),(OR,"fodder availability"),(OR,"f
groups"),(OR,"Grazing induced erosion"),(OR,"Grazing intensity"),(OR,"grazing land"),(OR,"
(OR,"intensive milk production"),(OR,"intensive pig production"),(OR,"Intensive pork"),(OR
distribution"),(OR,"livestock grazing"),(OR,"livestock groups"),(OR,"Livestock guts"),(OR,
sector"),(OR,"livestock sector policy"),(OR,"livestock sector supply"),(OR,"livestock sect
(OR,"meat consumption"),(OR,"meat dairy products"),(OR,"meat import"),(OR,"meat importing
(OR,"pastoral distribution"),(OR,"pastoral ecology"),(OR,"pastoral economies"),(OR,"pastor
strategies"),(OR,"Protection of farm animals"),(OR,"protein-energy malnutrition"),(OR,"pro
(OR,"sheep growing diets"),(OR,"sheep herders"),(OR,"Sheep industry"),(OR,"Sheep meat"),(
development"),(OR,"sustainable livestock production"),(OR,"sustainable pastoralism"),(OR,"
health"),(OR,"veterinary herd ownership"),(OR,"Veterinary hygiene"),(OR,"veterinary inputs
(OR,"abundant cattle"),(OR,"abundant cattle production"),(OR,"abundant vegetation"),(OR,"a
prices"),(OR,"animal producer"),(OR,"animal producers"),(OR,"animal product"),(OR,"animal
dioxide emis sions"),(OR,"carbon sequestration"),(OR,"carbonsequestration"),(OR,"carbonsec
(OR,"crop-livestock production systems"),(OR,"crop-livestock resource competition"),(OR,"c
(OR,"feed crop thinnings"),(OR,"feed distribution"),(OR,"feed efficiency"),(OR,"Feed explo
(OR,"grazing animals"),(OR,"grazing area"),(OR,"grazing behavior"),(OR,"grazing control"),
(OR,"livestock development"),(OR,"livestock development"),(OR,"livestock development advis
(OR,"livestock service","livestock services")),(OR,"livestock service budgets"),(OR,"lives
pasture"),(OR,"natural roughage"),(OR,"natural vegetation"),(OR,"nomad economy"),(OR,"noma
(OR,"pasture management techniques"),(OR,"pasture mean"),(OR,"pasture production"),(OR,"pa
(OR,"slaughterhouses"),(OR,"small ruminants"),(OR,"small ruminants"),(OR,"small stock"),(O
coverage"),(OR,"vaccination parks"),(OR,"Vaccination policies"),(OR,"vaccination programs"
(OR,"livestock herders"),(OR,"livestock holding"),(OR,"livestock holding areas"),(OR,"live
```

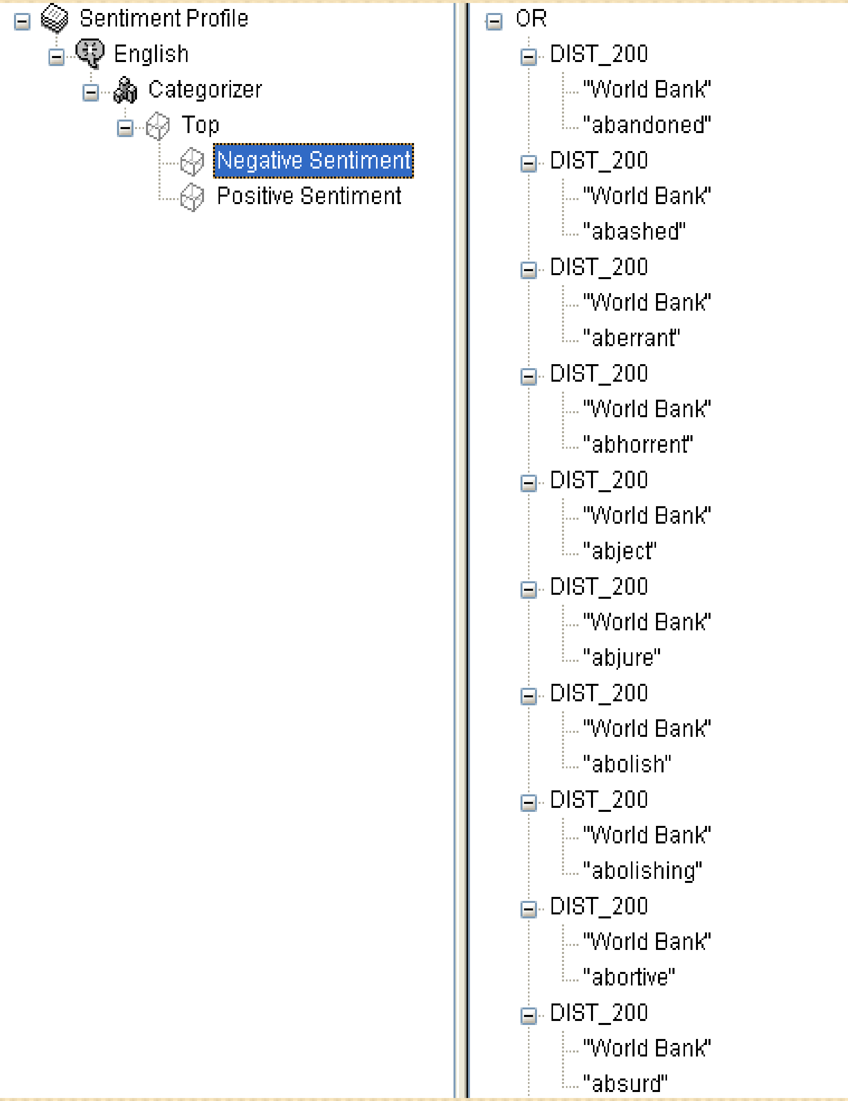
Sample Definition of Subclass Primary Education



- 644301 - Education
  - 672926 - Primary Education
  - 672927 - Secondary Education
  - 672928 - Tertiary Education
  - 672929 - Vocational Education and Technical Training
  - 672930 - Early Childhood Development
  - 672931 - Educational Technology and Distance Learr
  - 672932 - Access and Equity in Basic Education
  - 672934 - Non Formal Education
  - 672935 - Education Reform and Management
  - 672936 - Effective Schools and Teachers
  - 672936 - Teaching and Learning
  - 672937 - Economics of Education
  - 672938 - Adult Outreach
  - 672939 - School Health
  - 672940 - Curriculum and Instruction
  - 672941 - Educational Sciences
  - 672942 - Education and Society
  - 672943 - Educational Policy and Planning
  - 672945 - Educational Institutions and Facilities
  - 672946 - Educational Populations
  - 738813 - Education and Digital Divide
  - 738814 - Lifelong Learning
  - 738815 - Science Education
  - 738816 - Public Examination System
  - 758551 - Education Finance
  - 758553 - Education Indicators and Statistics
  - 758554 - Education Sector Strategy and Lending
  - 761314 - Knowledge for Development
  - 761315 - Education, Violence and Social Cohesion

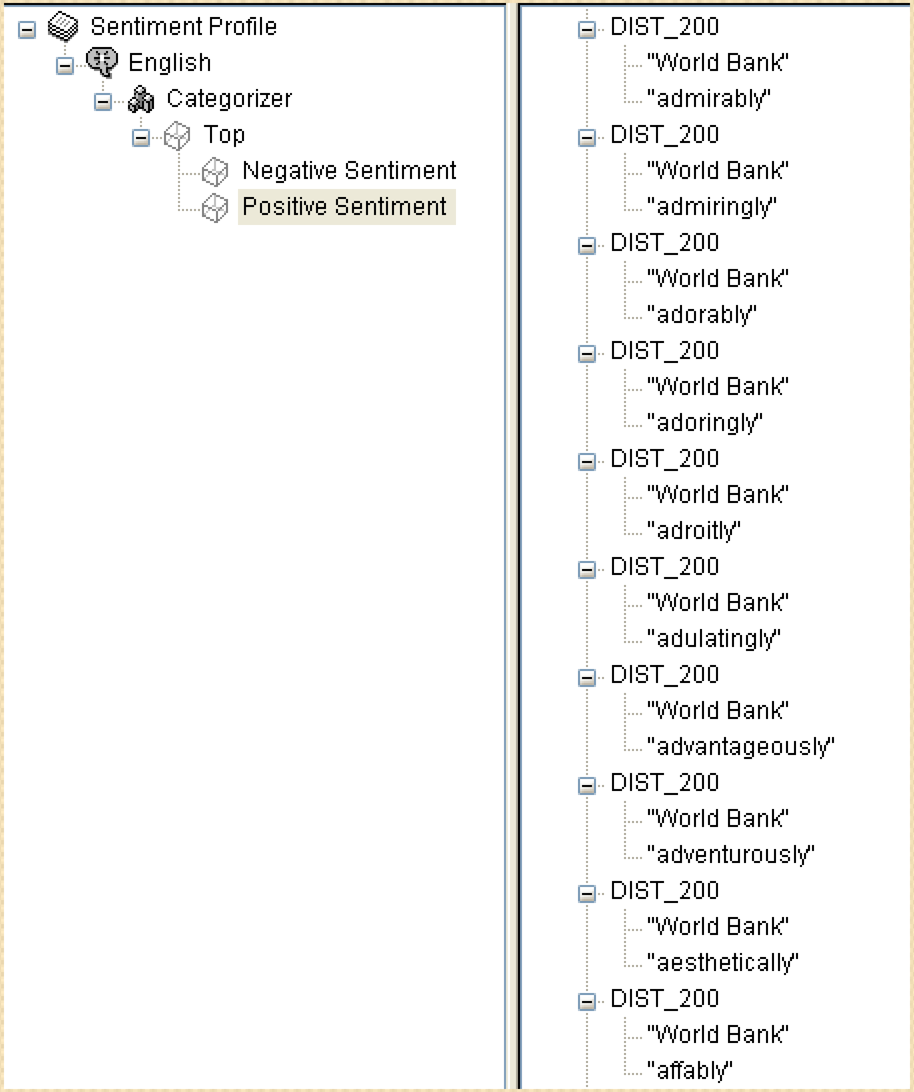
```
(OR,(OR,"Academic learning"),(OR,"Academic subjects"),(OR,"Access to education strategies"
enrollment rates"),(OR,"Countrywide enrolment rate"),(OR,"Cultural development"),(OR,"Cult
environments"),(OR,"Effective teaching"),(OR,"Elementary education"),(OR,"Elementary educa
books"),(OR,"Life skills"),(OR,"Life skills curriculum"),(OR,"Life skills manuals"),(OR,"L
(OR,"Preschool caregivers"),(OR,"Preschool centers"),(OR,"Preschool centres"),(OR,"Prescho
expenditure"),(OR,"Public institutions"),(OR,"Public participation"),(OR,"Public preprimar
education"),(OR,"Universal primary education"),(OR,"Universal school choice"),(OR,"UPE"),(
(OR,"Annual Financing Gap"),(OR,"annual inflation rate@N"),(OR,"Annual instructional hours
education"),(OR,"basic instructional aids"),(OR,"basic knowledge"),(OR,"basic learning"),(
activities"),(OR,"classroom construction requirements"),(OR,"classroom environment"),(OR,"
(OR,"curriculum requirements"),(OR,"Curriculum Research"),(OR,"curriculum resources"),(OR,
of User Fees"),(OR,"enrichment materials"),(OR,"enrollment by age"),(OR,"enrollment capaci
disparities"),(OR,"Gender Equality"),(OR,"gender equality in education"),(OR,"gender equal
materials"),(OR,"learning outcomes"),(OR,"learning resources"),(OR,"learning time"),(OR,"l
(OR,"primary graduate@N"),(OR,"primary graduation"),(OR,"primary gross enrollment"),(OR,"P
(OR,"Primary Training"),(OR,"primary years"),(OR,"Private Education"),(OR,"Private enrollm
(OR,"reintegration of children"),(OR,"retention of primary school students"),(OR,"retentio
(OR,"student attainment"),(OR,"student attendance"),(OR,"student bod@N"),(OR,"student book
```

Sample Sentiment Analysis Profile



```
Sentiment Profile
  English
    Categorizer
      Top
        Negative Sentiment
        Positive Sentiment
```

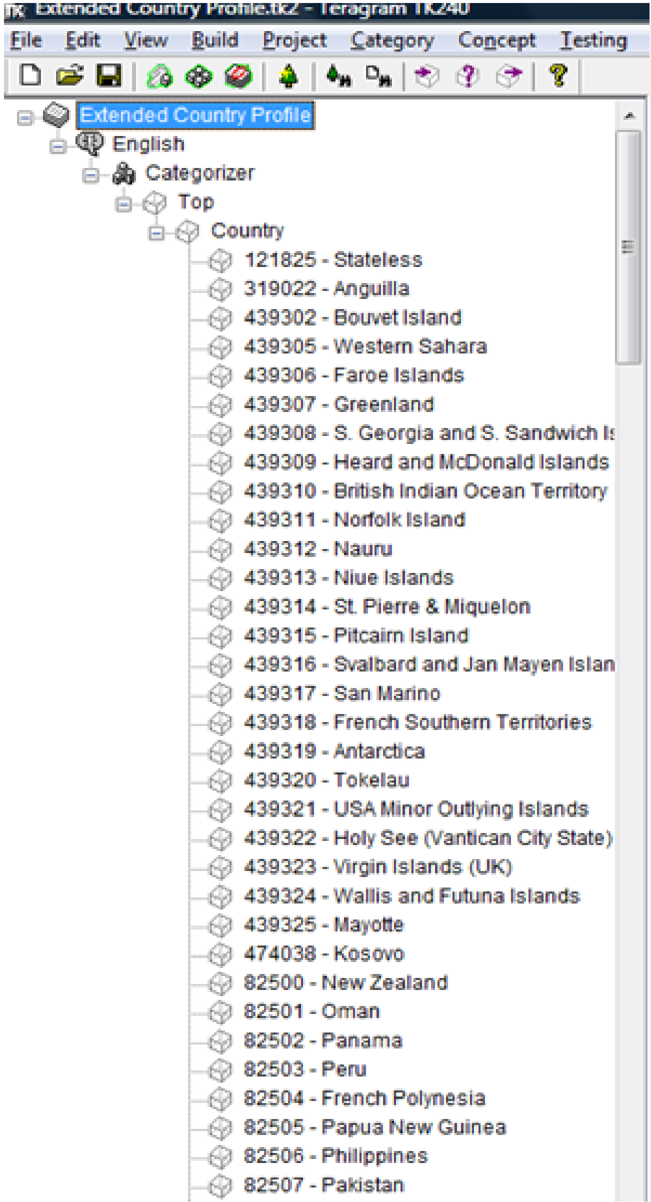
```
OR
  DIST_200
    "World Bank"
    "abandoned"
  DIST_200
    "World Bank"
    "abashed"
  DIST_200
    "World Bank"
    "aberrant"
  DIST_200
    "World Bank"
    "abhorrent"
  DIST_200
    "World Bank"
    "abject"
  DIST_200
    "World Bank"
    "abjure"
  DIST_200
    "World Bank"
    "abolish"
  DIST_200
    "World Bank"
    "abolishing"
  DIST_200
    "World Bank"
    "abortive"
  DIST_200
    "World Bank"
    "absurd"
```



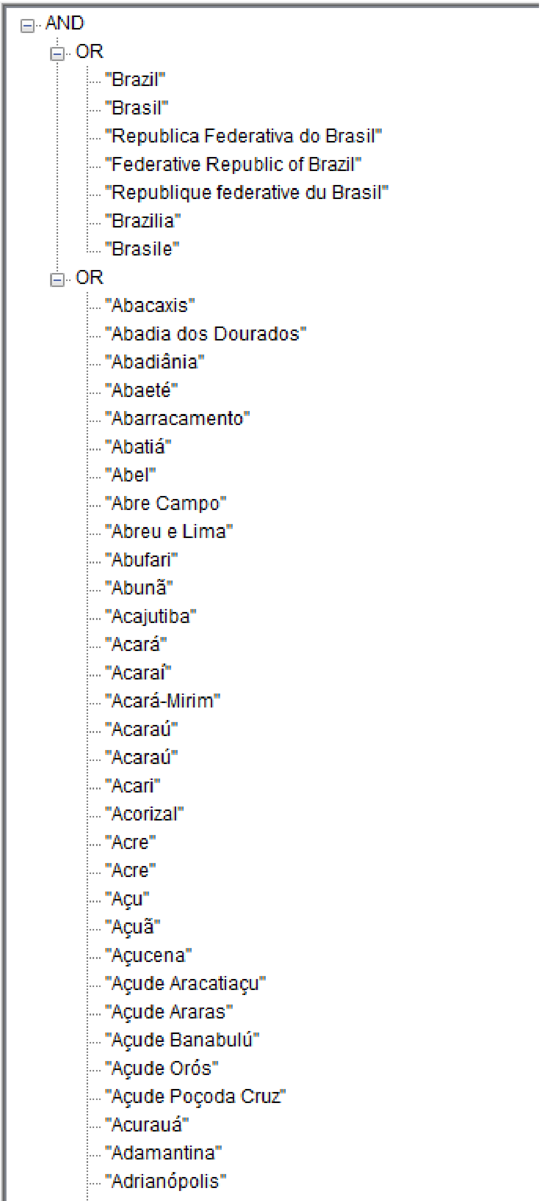
```
Sentiment Profile
  English
    Categorizer
      Top
        Negative Sentiment
        Positive Sentiment
```

```
  DIST_200
    "World Bank"
    "admirably"
  DIST_200
    "World Bank"
    "admiringly"
  DIST_200
    "World Bank"
    "adorably"
  DIST_200
    "World Bank"
    "adoringly"
  DIST_200
    "World Bank"
    "adroitly"
  DIST_200
    "World Bank"
    "adulatingly"
  DIST_200
    "World Bank"
    "advantageously"
  DIST_200
    "World Bank"
    "adventurously"
  DIST_200
    "World Bank"
    "aesthetically"
  DIST_200
    "World Bank"
    "affably"
```

Another Example:
Country Categorization and City Extraction



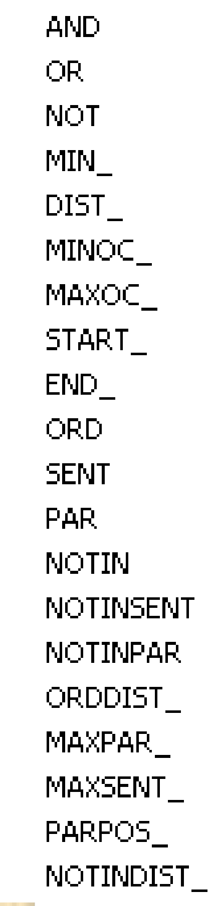
```
Extended Country Profile
  English
    Categorizer
      Top
        Country
          121825 - Stateless
          319022 - Anguilla
          439302 - Bouvet Island
          439305 - Western Sahara
          439306 - Faroe Islands
          439307 - Greenland
          439308 - S. Georgia and S. Sandwich Is
          439309 - Heard and McDonald Islands
          439310 - British Indian Ocean Territory
          439311 - Norfolk Island
          439312 - Nauru
          439313 - Niue Islands
          439314 - St. Pierre & Miquelon
          439315 - Pitcairn Island
          439316 - Svalbard and Jan Mayen Islan
          439317 - San Marino
          439318 - French Southern Territories
          439319 - Antarctica
          439320 - Tokelau
          439321 - USA Minor Outlying Islands
          439322 - Holy See (Vantican City State)
          439323 - Virgin Islands (UK)
          439324 - Wallis and Futuna Islands
          439325 - Mayotte
          474038 - Kosovo
          82500 - New Zealand
          82501 - Oman
          82502 - Panama
          82503 - Peru
          82504 - French Polynesia
          82505 - Papua New Guinea
          82506 - Philippines
          82507 - Pakistan
```



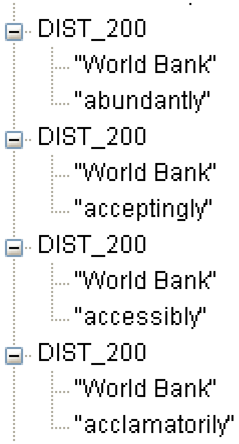
```
AND
  OR
    "Brazil"
    "Brasil"
    "Republica Federativa do Brasil"
    "Federative Republic of Brazil"
    "Republique federative du Brasil"
    "Brazilia"
    "Brasile"
  OR
    "Abacaxis"
    "Abadia dos Dourados"
    "Abadiânia"
    "Abaeté"
    "Abarracamento"
    "Abatiá"
    "Abel"
    "Abre Campo"
    "Abreu e Lima"
    "Abufari"
    "Abunã"
    "Acajutiba"
    "Acará"
    "Acaraí"
    "Acará-Mirim"
    "Acaraú"
    "Acaraú"
    "Acari"
    "Acorizal"
    "Acre"
    "Acre"
    "Açu"
    "Açuã"
    "Açucena"
    "Açude Aracatiaçu"
    "Açude Araras"
    "Açude Banabulú"
    "Açude Orós"
    "Açude Poçoda Cruz"
    "Acurauá"
    "Adamantina"
    "Adrianópolis"
```

Operator and Condition Based Matching

Common Matching Operators

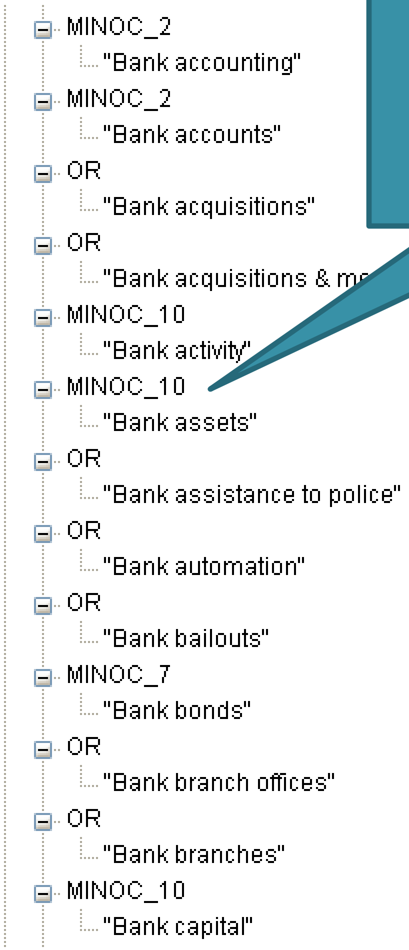


AND
OR
NOT
MIN_
DIST_
MINOC_
MAXOC_
START_
END_
ORD
SENT
PAR
NOTIN
NOTINSENT
NOTINPAR
ORDDIST_
MAXPAR_
MAXSENT_
PARPOS_
NOTINDIST_



```
DIST_200
    "World Bank"
    "abundantly"
DIST_200
    "World Bank"
    "acceptingly"
DIST_200
    "World Bank"
    "accessibly"
DIST_200
    "World Bank"
    "acclamatorily"
```

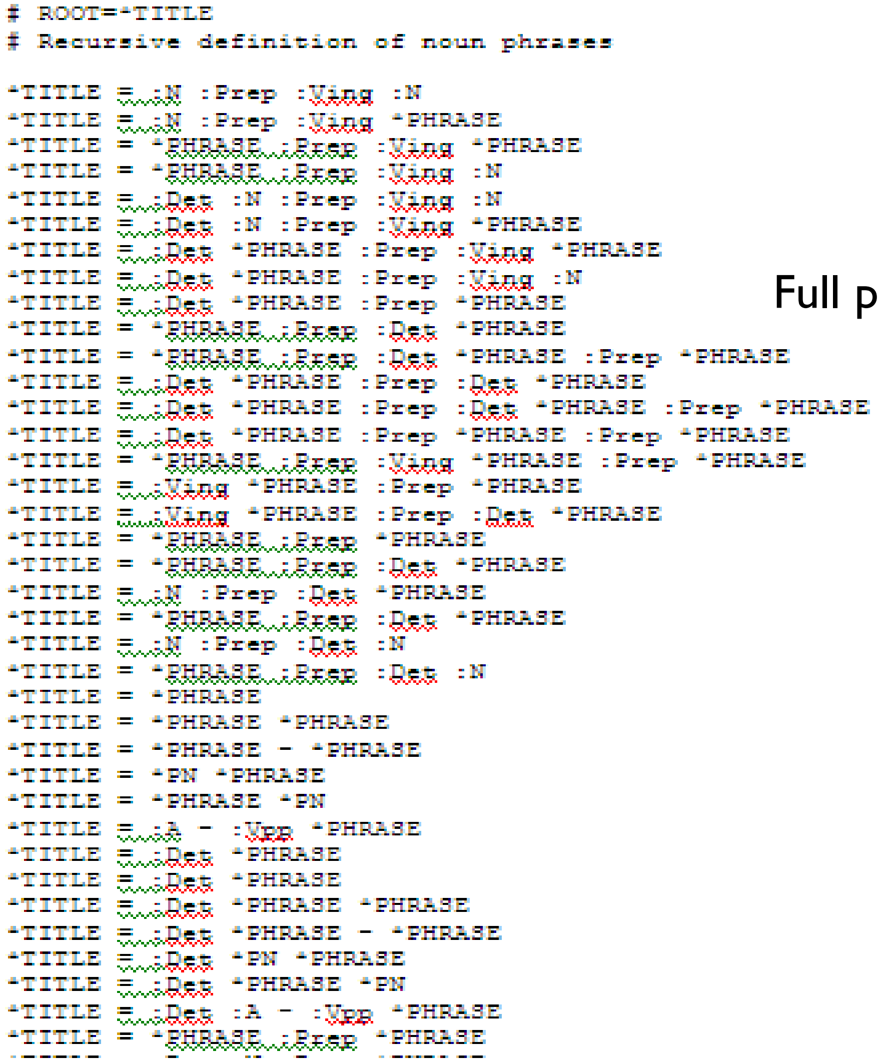
If you find this word within 200 characters of "World Bank" then score as one match



```
MINOC_2
    "Bank accounting"
MINOC_2
    "Bank accounts"
OR
    "Bank acquisitions"
OR
    "Bank acquisitions & me
MINOC_10
    "Bank activity"
MINOC_10
    "Bank assets"
OR
    "Bank assistance to police"
OR
    "Bank automation"
OR
    "Bank bailouts"
MINOC_7
    "Bank bonds"
OR
    "Bank branch offices"
OR
    "Bank branches"
MINOC_10
    "Bank capital"
```

Do not match on this concept unless there are a minimum of 10 occurrences in the entity.

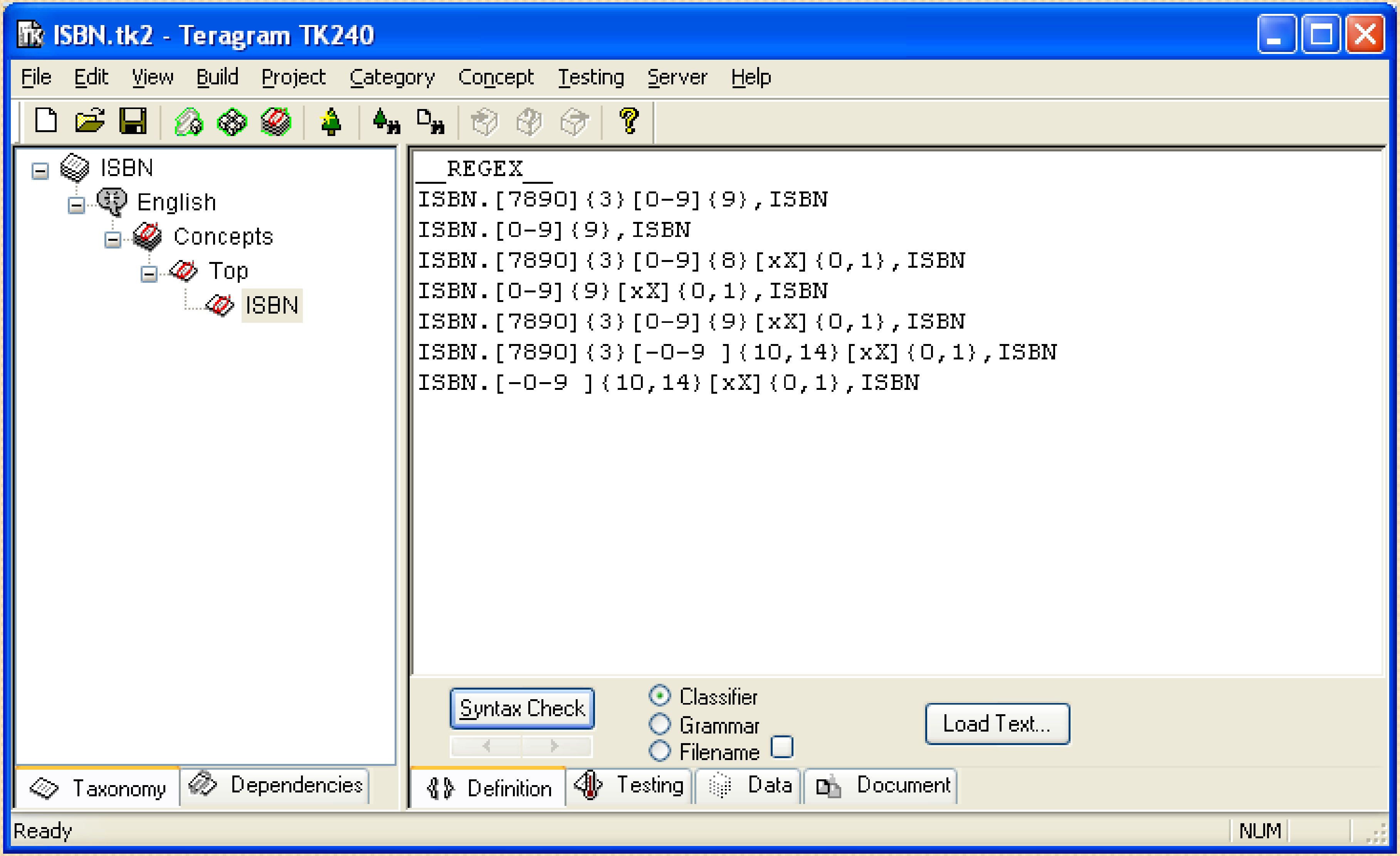
Example 3:
Partial Grammatical Concept Extraction for Titles



```
# ROOT=*TITLE
# Recursive definition of noun phrases

*TITLE = :N :Prep :Ving :N
*TITLE = :N :Prep :Ving *PHRASE
*TITLE = *PHRASE :Prep :Ving *PHRASE
*TITLE = *PHRASE :Prep :Ving :N
*TITLE = :Det :N :Prep :Ving :N
*TITLE = :Det :N :Prep :Ving *PHRASE
*TITLE = :Det *PHRASE :Prep :Ving *PHRASE
*TITLE = :Det *PHRASE :Prep :Ving :N
*TITLE = :Det *PHRASE :Prep *PHRASE
*TITLE = *PHRASE :Prep :Det *PHRASE
*TITLE = *PHRASE :Prep :Det *PHRASE :Prep *PHRASE
*TITLE = :Det *PHRASE :Prep :Det *PHRASE
*TITLE = :Det *PHRASE :Prep :Det *PHRASE :Prep *PHRASE
*TITLE = :Det *PHRASE :Prep *PHRASE :Prep *PHRASE
*TITLE = *PHRASE :Prep :Ving *PHRASE :Prep *PHRASE
*TITLE = :Ving *PHRASE :Prep *PHRASE
*TITLE = :Ving *PHRASE :Prep :Det *PHRASE
*TITLE = *PHRASE :Prep *PHRASE
*TITLE = *PHRASE :Prep :Det *PHRASE
*TITLE = :N :Prep :Det *PHRASE
*TITLE = *PHRASE :Prep :Det *PHRASE
*TITLE = :N :Prep :Det :N
*TITLE = *PHRASE :Prep :Det :N
*TITLE = *PHRASE
*TITLE = *PHRASE *PHRASE
*TITLE = *PHRASE - *PHRASE
*TITLE = *PN *PHRASE
*TITLE = *PHRASE *PN
*TITLE = :A - :Vpp *PHRASE
*TITLE = :Det *PHRASE
*TITLE = :Det *PHRASE
*TITLE = :Det *PHRASE *PHRASE
*TITLE = :Det *PHRASE - *PHRASE
*TITLE = :Det *PN *PHRASE
*TITLE = :Det *PHRASE *PN
*TITLE = :Det :A - :Vpp *PHRASE
*TITLE = *PHRASE :Prep *PHRASE
```

Full profile is about 4 pages long

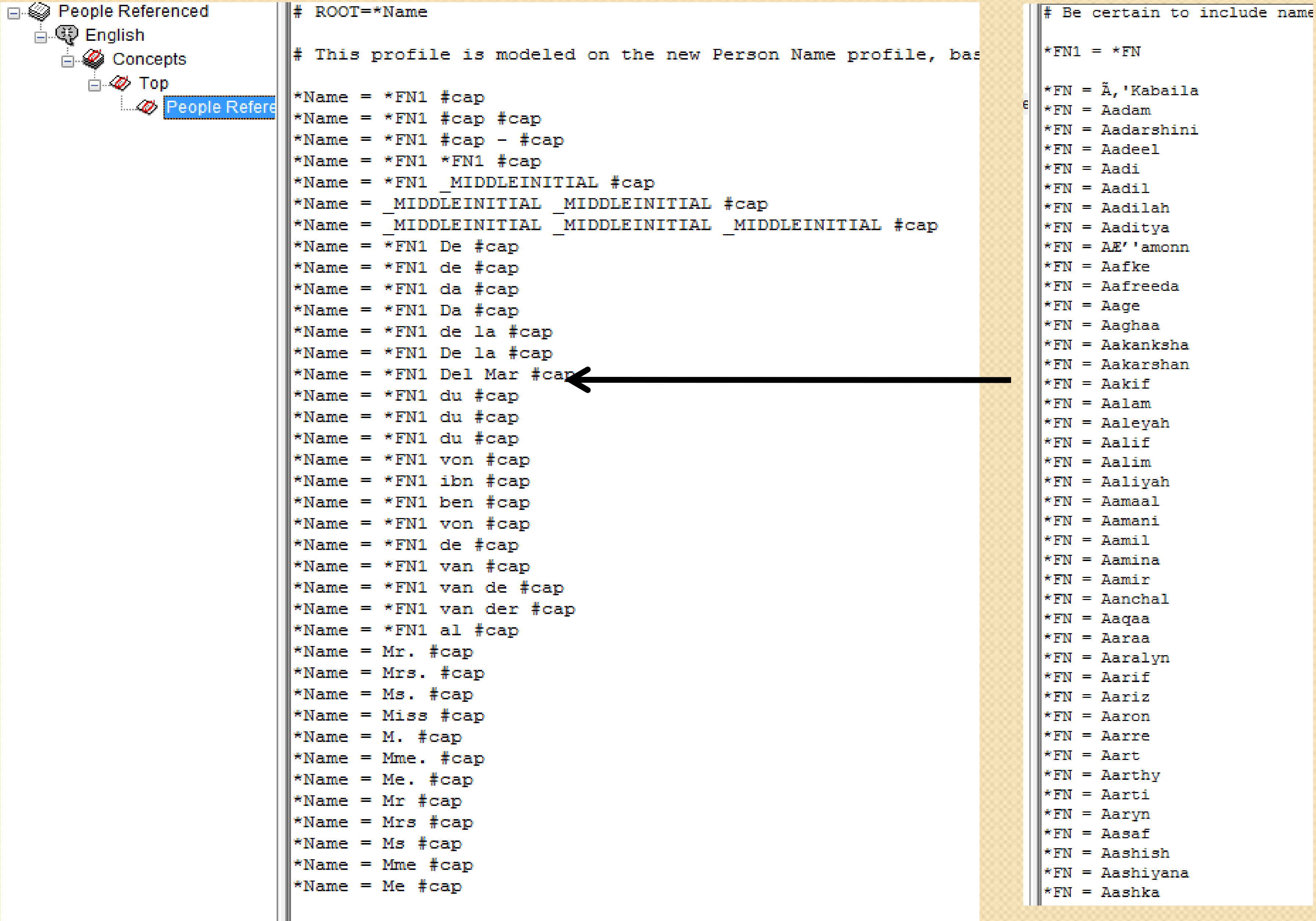
Example 4:
ISBN Concept Extraction Profile



```
__REGEX__
ISBN.[7890]{3}[0-9]{9},ISBN
ISBN.[0-9]{9},ISBN
ISBN.[7890]{3}[0-9]{8}[xX]{0,1},ISBN
ISBN.[0-9]{9}[xX]{0,1},ISBN
ISBN.[7890]{3}[0-9]{9}[xX]{0,1},ISBN
ISBN.[7890]{3}[-0-9 ]{10,14}[xX]{0,1},ISBN
ISBN.[-0-9 ]{10,14}[xX]{0,1},ISBN
```

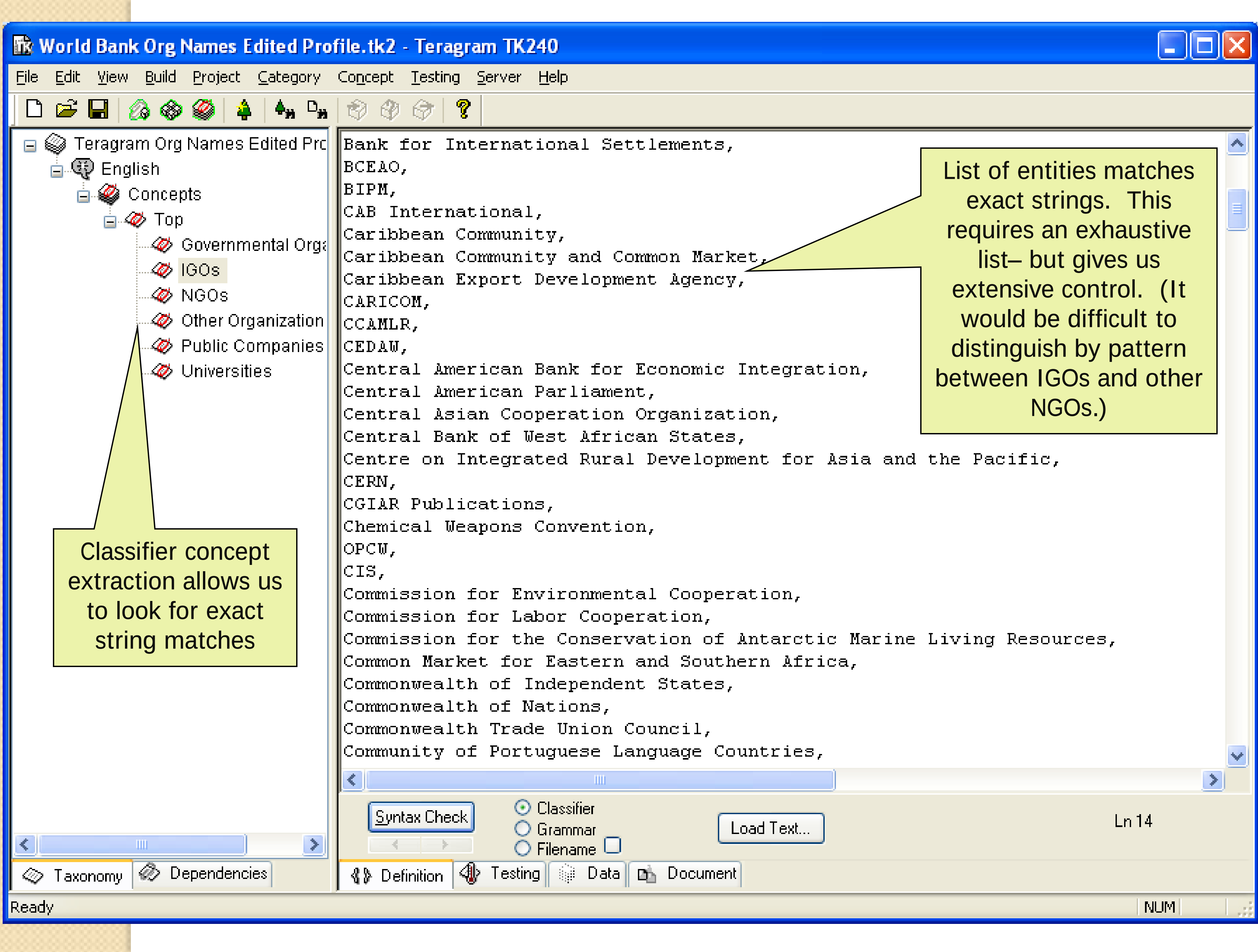
● Classifier
○ Grammar
○ Filename □

Example 5:
People Profile With Authority File of First Names



```
# ROOT=*Name

# This profile is modeled on the new Person Name profile, bas

*Name = *FN1 #cap
*Name = *FN1 #cap #cap
*Name = *FN1 #cap - #cap
*Name = *FN1 *FN1 #cap
*Name = *FN1 _MIDDLEINITIAL #cap
*Name = _MIDDLEINITIAL _MIDDLEINITIAL #cap
*Name = _MIDDLEINITIAL _MIDDLEINITIAL _MIDDLEINITIAL #cap
*Name = *FN1 De #cap
*Name = *FN1 de #cap
*Name = *FN1 da #cap
*Name = *FN1 Da #cap
*Name = *FN1 de la #cap
*Name = *FN1 De la #cap
*Name = *FN1 Del Mar #cap
*Name = *FN1 du #cap
*Name = *FN1 du #cap
*Name = *FN1 du #cap
*Name = *FN1 von #cap
*Name = *FN1 ibn #cap
*Name = *FN1 ben #cap
*Name = *FN1 von #cap
*Name = *FN1 de #cap
*Name = *FN1 van #cap
*Name = *FN1 van de #cap
*Name = *FN1 van der #cap
*Name = *FN1 al #cap
*Name = Mr. #cap
*Name = Mrs. #cap
*Name = Ms. #cap
*Name = Miss #cap
*Name = M. #cap
*Name = Mme. #cap
*Name = Me. #cap
*Name = Mr #cap
*Name = Mrs #cap
*Name = Ms #cap
*Name = Mme #cap
*Name = Me #cap
```

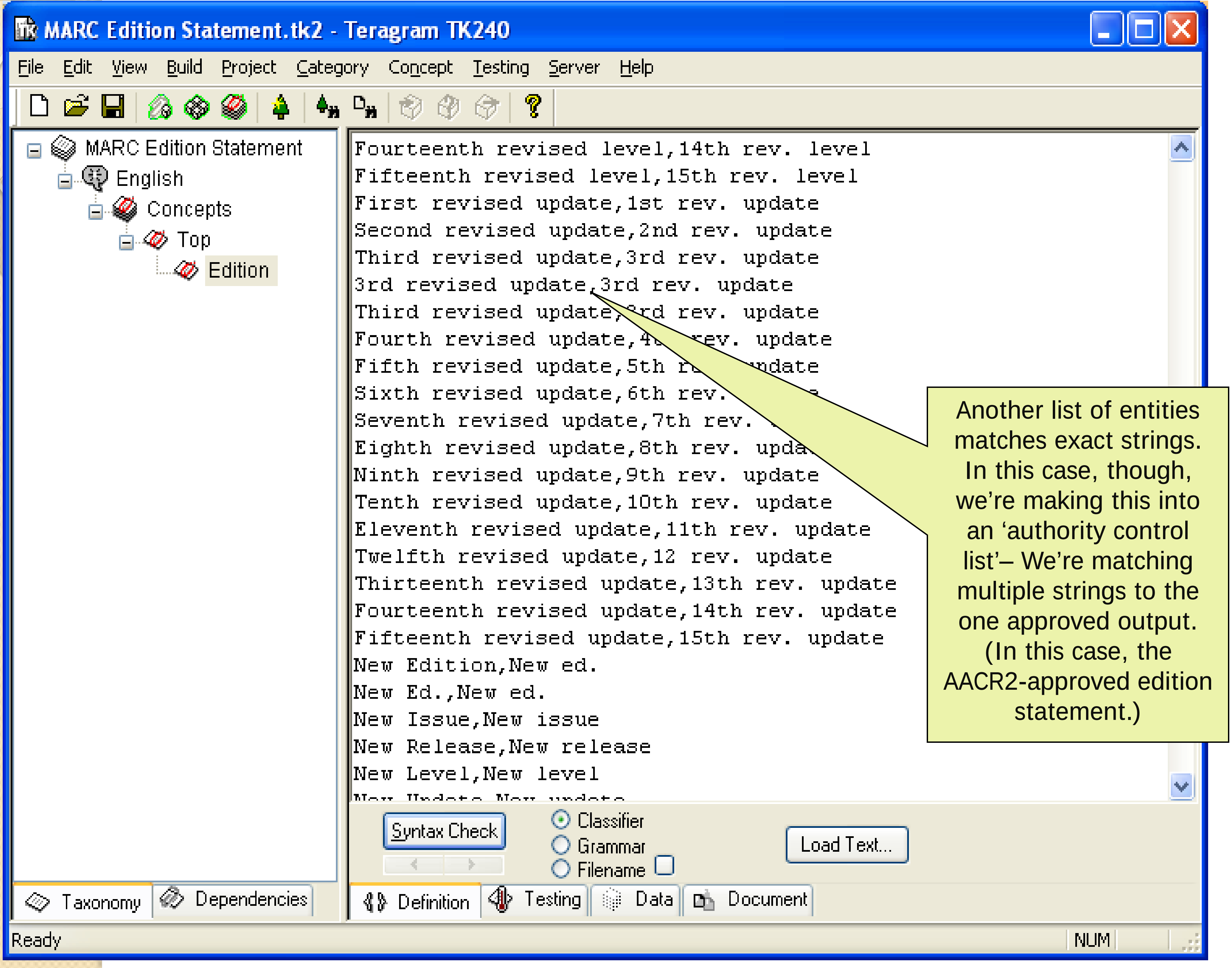
```
# Be certain to include name

*FN1 = *FN

*FN = Ã,'Kabaila
*FN = Aadam
*FN = Aadarshini
*FN = Aadeel
*FN = Aadi
*FN = Aadil
*FN = Aadilah
*FN = Aaditya
*FN = AÆ''amonn
*FN = Aafke
*FN = Aafreeda
*FN = Aage
*FN = Aaghaa
*FN = Aakanksha
*FN = Aakarshan
*FN = Aakif
*FN = Aalam
*FN = Aaleyah
*FN = Aalif
*FN = Aalim
*FN = Aaliyah
*FN = Aamaal
*FN = Aamani
*FN = Aamil
*FN = Aamina
*FN = Aamir
*FN = Aanchal
*FN = Aaqaa
*FN = Aaraa
*FN = Aaralyn
*FN = Aarif
*FN = Aariz
*FN = Aaron
*FN = Aarre
*FN = Aart
*FN = Aarthy
*FN = Aarti
*FN = Aaryn
*FN = Aasaf
*FN = Aashish
*FN = Aashiyana
*FN = Aashka
```



```
Bank for International Settlements,
BCEAO,
BIPM,
CAB International,
Caribbean Community,
Caribbean Community and Common Market,
Caribbean Export Development Agency,
CARICOM,
CCAMLR,
CEDAW,
Central American Bank for Economic Integration,
Central American Parliament,
Central Asian Cooperation Organization,
Central Bank of West African States,
Centre on Integrated Rural Development for Asia and the Pacific,
CERN,
CGIAR Publications,
Chemical Weapons Convention,
OPCW,
CIS,
Commission for Environmental Cooperation,
Commission for Labor Cooperation,
Commission for the Conservation of Antarctic Marine Living Resources,
Common Market for Eastern and Southern Africa,
Commonwealth of Independent States,
Commonwealth of Nations,
Commonwealth Trade Union Council,
Community of Portuguese Language Countries,
```

List of entities matches exact strings. This requires an exhaustive list– but gives us extensive control. (It would be difficult to distinguish by pattern between IGOs and other NGOs.)

Classifier concept extraction allows us to look for exact string matches



```
Fourteenth revised level,14th rev. level
Fifteenth revised level,15th rev. level
First revised update,1st rev. update
Second revised update,2nd rev. update
Third revised update,3rd rev. update
3rd revised update,3rd rev. update
Third revised update, rd rev. update
Fourth revised update,4      ev. update
Fifth revised update,5th re      pdate
Sixth revised update,6th rev.
Seventh revised update,7th rev.
Eighth revised update,8th rev. upda
Ninth revised update,9th rev. update
Tenth revised update,10th rev. update
Eleventh revised update,11th rev. update
Twelfth revised update,12 rev. update
Thirteenth revised update,13th rev. update
Fourteenth revised update,14th rev. update
Fifteenth revised update,15th rev. update
New Edition,New ed.
New Ed.,New ed.
New Issue,New issue
New Release,New release
New Level,New level
New Update,New update
```

Another list of entities matches exact strings. In this case, though, we're making this into an 'authority control list'– We're matching multiple strings to the one approved output. (In this case, the AACR2-approved edition statement.)

THE ROLE OF HUMAN KNOWLEDGE IN METADATA GENERATION

No Semantic Future Without Catalogers…..

- Catalogers need to be involved in configuring and designing the semantic applications
  - Identifying the best sources of reference knowledge
  - Serving as the "experts" for "expert systems" development
  - Performing quality control on processes

- In the future, catalogers' knowledge and ways of thinking and working will be the basis of well designed semantic analysis applications

- Both the need for catalogers and the role they play will become critical in the future

Catalogers as Knowledge Engineers

- Role of the cataloger will be shifted in the future from a "Doer" to a "Designer" -- "Knowledge Engineer"

- Designing the context, the content and taking a more proactive role in *engineering* access to not only information but knowledge

- Future information landscape is inherently "semantic" which aligns very closely with a cataloger's tacit knowledge

- Cataloger's tacit knowledge includes rules of thumb, interpretation of guidelines, knowledge of sources, and knowledge of domains

Catalogers as Knowledge Engineers

- This shift will mean:
  - Learning how to design and build the reference sources, how to develop and apply guiding principles and how to manage reference sources
  - Teaching semantic analysis methods and knowledge organization systems
  - Putting the tools in the hands of catalogers
  - Involving catalogers in the semantic analysis design and development process

- In many professional schools, we only teach catalogers how to "use" general purpose reference sources – that source is designed for one area of practice and one general audience - this does not fully leverage our professional knowledge

dbedfor3@kent.edu
dbedford@worldbank.org

THANK YOU!

QUESTIONS & COMMENTS?