Redefining the Search Question

Denise Bedford
World Bank

Presentation to Canadian Metadata Forum
September 27, 2005
Library and Archives of Canada

Search – The Essential Question

- The question is no longer full text versus catalogue-type searching. The question is how and when you will implement semantic searching.
- Semantic searching combines aspects of both full text and catalogue searching, but it moves above the negative aspects of both.
- It leverages a metadata architecture base, deep conceptual indexing, programmatic methods of capturing metadata, works across and within languages, and promotes interoperability.
- Semantic search is also the foundation upon which to build an enterprise search architecture.
- For varying reasons, neither full text nor catalogue searching will scale to enterprise search.

Overview

- What do users need and how can you build a search system to support the need?
- What is Enterprise Search?
- What is Semantic Search?
- How to Fuel Your Semantic Search with Metadata

What do users want?

Basic Assumptions About Search Systems

- Search systems are not WYSIWYG – what you see is not what you get
- Search systems are more like ice-bergs - most of the search system is below the surface – you can't see it, you don't know how it has been configured or what components it has
- Search systems have some basic components, but there are some significant differences as well
- You need to know what your users want before you decide what kind of a search system you need

Begin at the beginning…

- The most valuable advice I can offer to you based on my experience working with search over the past 30 years is to make sure you know what problem you're solving before you apply a solution
- While this seems straight forward and logical, it is rarely the approach we take to solving search problems
- Before you select a search solution, make sure you know what kind of a search problem you have and whether the solution fits the problem
- A few words about the kind of search challenge we are facing

The Environment

- Like other libraries we acquire, create, license, access information in 30 subject domains – ranging from Law & Justice, to Transport, to Environment, to Agriculture, to Education, Health & Nutrition, etc.
- We have 500 different kinds of content/document types
- We have a set of business processes which represent the way we do our work
- We have six working languages – but actually working in many more
- We have a rich history – content dating back to the 1940s
- Our priorities change over time
- Everyone in the Bank is a recognized international expert

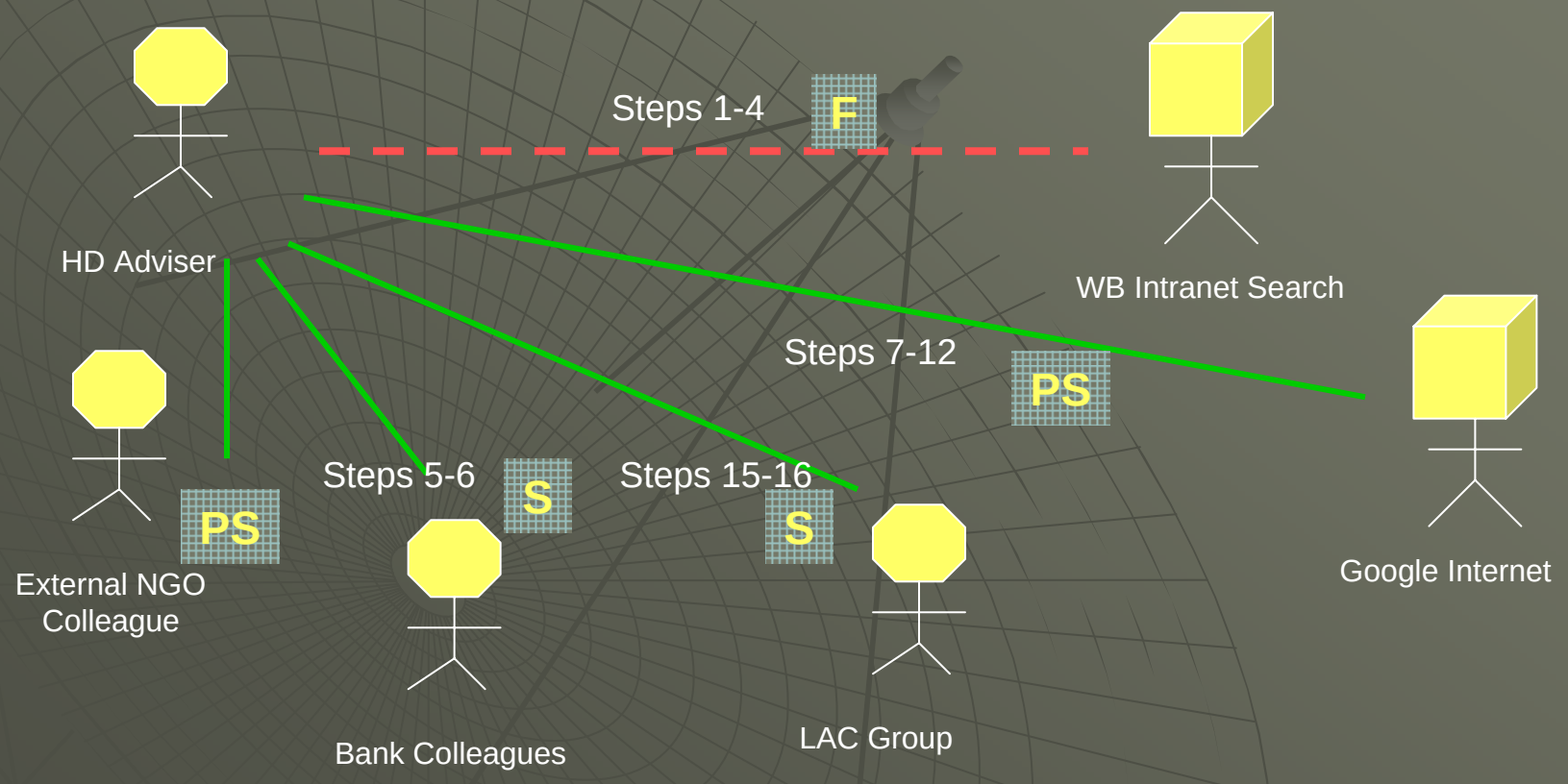
Search Environment

- It is very challenging to design a search system which meets the needs of this complex environment
- Internal Experts do ‘known item’ searching and look for other experts to talk to
- General public looks for information ‘about’ something
- Most people don’t know all the dimensions of the Bank
- People need to search in other than English and find content written in other languages
- Our new Knowledge and Learning initiative is moving us towards a W3 working environment - What information I need when I need it, where I need it

Search Has Always Been Problemmatic

- Despite any reports you may read in the general literature, buying a fast crawler (Google) definitely did not solve our search problems
- Implementing a fast crawler simply surfaced our information management and data quality challenges directly to the users
- The crawler approach provided high hit rates, and very low relevancy rates AND very low precision rates (it created more problems than it solved)
- Our users were very unhappy with the search system and requested a new solution
- Some on the development side didn't understand the 'search problem' so we did some 'use cases'
- Here's a sample of what we learned….

Use Case Identifier: 1.15
Use Case Name: Racial Inequality Speech Preparation
Goal: HD Adviser wants to find social indicators, relevant quotes and Bank's position on racism, as well as learn what the Bank has done in this area in order to prepare a 30 min. speech for a WB Managing Director for conference.



Use Case Outcomes: Success
Use Case Metrics:
Actors/Agents: 5

Search System Factors:
# of Steps: 16
Avg. Steps per Agent: 3.2

What We Learned

- The absolute success rate per search task was low at 43.18%
- The Search experience generally consists of multiple steps and multiple searches within and across sources
- The source with the fewest number of steps was a Colleague or Personal Contact. The source with the greatest number of steps was External Web Search Browse
- Each source has its own behavior, business rules, functional architectures – users need to learn each system
- The Intranet search was selected as a logical place to look for information in over 65% of the use cases. However, the success rate for the Intranet Google search was only 35%
- In the past three months we have extended our analysis to look at 'user relevancy' levels – we also found that the relevancy rate is low – between 30% – 47% of the top 30 results

What is Enterprise Search?

And, what it is not….

Enterprise Search Solution

- What Enterprise Search is and what it is not
- Distinguish between the web services platform for accessing Enterprise Search and the enterprise's full store of content (its not only electronic, and its not all in html format!!)
- What kind of an architecture do you need to support enterprise search?
- What kind of tools do you need to support enterprise search? Its much more than just a web crawler….
- What do you need to support true concept level cross-language searching?
- How does Enterprise Search respect security classifications with/without restricting knowledge of what exists?

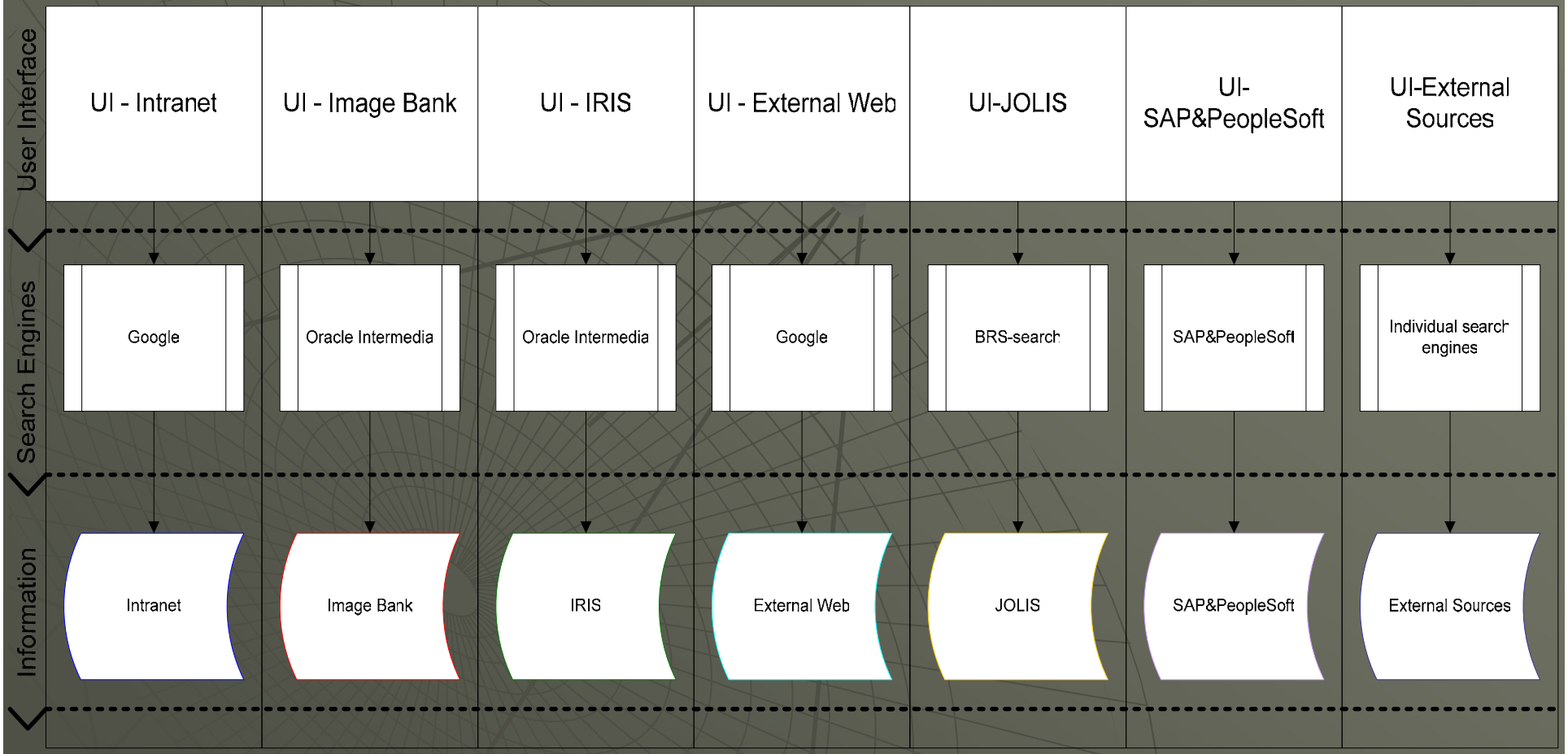
What Enterprise Search IS

- It is comprehensive in coverage of all your enterprise's information resources – regardless of the kind of information, the format, or the language
- It supports access to information based on privileges or security classifications
- It is integrating in the way it discovers and presents information to searchers
- It supports both simple searching and fielded or faceted searching
- It supports the contextualization of information by users
- It supports both searching of all content, and the creation of 'frontier' search systems – a parameterization of all the content

What Enterprise Search Is NOT

- It is NOT limited to just a web crawler which covers only your web published electronic content
- It is NOT a search which goes against a simple huge full-text index
- It is NOT a subset of your enteprise's information which only allows searching of publicly available content (this is a frontier search)
- It is NOT a federated or partitioned search which presents results back to users in silos which represent how/where the information is stored

Underlying Search Architecture



Silos of Information = Multiple search engines and user interfaces

Search Engine Parts

Contextualization, Personalization, Recommender, Content Syndication, Q&A Systems, Intelligent Search

Search Outputs

Query Manipulation

Simple search

Fielded Search

Query Processing Algorithms

Search term assistance (thesaurus, dictionaries)

Search language selection

Sources selection

Display Results

Integrated search results

Search results sorting

Search results contextualization

Search results relevancy ranking

What the user sees

Query Matching Algorithms

Boolean matching

Exact phrase matching

Fuzzy matching

Term matching + synonym expansion

Term matching + cross language expansion

Term weighted matching

Query term root matching & dictionary expansion

Wild card matching

Term proximity matching

Neural network expansion

Genetic algorithm expansion

The basic search programs

Search System Inputs

Metadata Repository for Enterprise Search

Index Architecture & Construction

Vocabulary Support (thesaurus)

Classification Schemes (taxonomy)

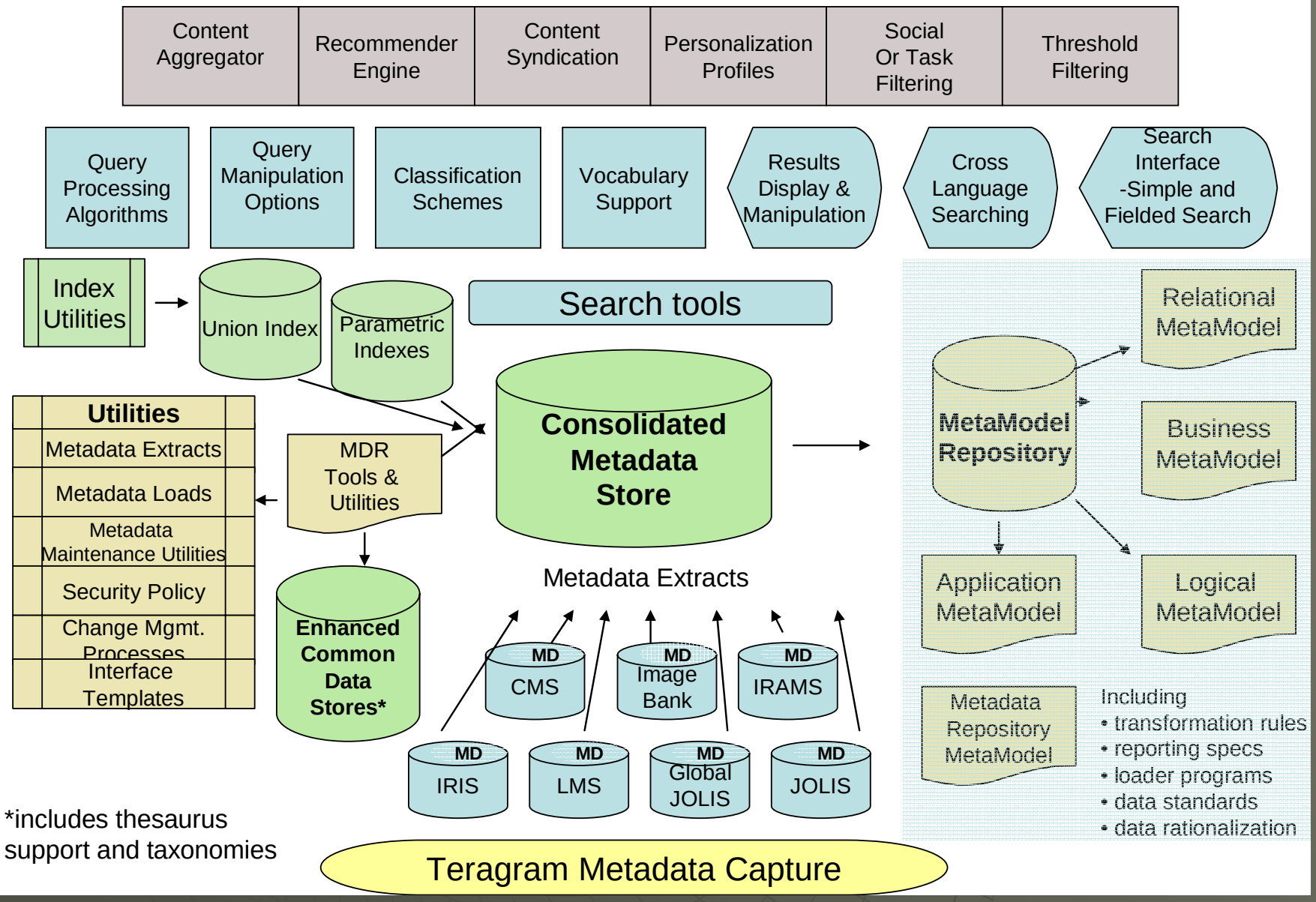
Foundation or Baseline of The Search System

Proposed Solution: Data Driven Enterprise Search

- Architecture anchored on
  - Enterprise Master Data Store which contains and supports all of the Bank's taxonomies
  - Enterprise Metadata Repository

- Tools
  - Oracle Intermedia Search Engine/UltraSearch
  - Teragram programmatic metadata capture
  - Search Portal
  - Enterprise Content Management system in future for managing our electronic content – one which supports metadata capture and more sophisticated information access and publishing needs

Enterprise Search Functional Architecture



Semantic Search

What is it and how do we achieve it?

Begin with a challenge…

- I provide a challenge to the group to find me a web search engine which supports a true advanced search – one which allows you to do fielded searching and to contextualize your search ……
- Any environment that supports research or experts needs to be able to contextualize search….
- What do you need to architect an advanced search?
- Why don’t web search engines support this type of searching?

Ontologies and Semantic Search

- Semantic Search leverages ontologies – but we need to be clear about what an ontology is…Ontology is an integrated design of different kinds of taxonomies, rules and entities
- Let's see what a semantic search interface might look like
- Then let's deconstruct what's behind the interface

Advanced Search

Search | Clear | Search Help

Text/Keyword [ ] All words

Title [ ]

Author [ ]

Ring taxonomy

Date From [ ] To [ ]

Country
Albania
Bosnia-Herzegovina
Congo, Democratic Republic
Egypt, Arab Republic of
Albania

Ring taxonomy

Region
Select All
Africa - Sub-Saharan
East Asia & the Pacific
Europe & Central Asia
Latin America & Caribbean

Topic
Select All
Agriculture & Rural Development
Children & Youth
Economic Policy & Debt
Education

Subtopic
Select All
Environmental & Management
Forests & Forestry
Biodiversity
Climate Change

Document Type: Select Document Type

Flat taxonomy

Language Any Language

Hierarchical taxonomy

Numbers Report [ ] Loan [ ] Credit [ ]
Project ID [ ] Trust Funds [ ]

Sort results by Date Type Relevence

Display results in the sets of 10 20 50

Search | Clear

Fielded Search = Faceted Taxonomy

Results for "biodiversity"

Search | New Search | Search Help | Saved Search

Did you mean: "biodiversity"

Ring Taxonomy

Metadata

Narrow your search by:

- All
- Country
- Document Type
- Topic

Sort results by: Relevance | Date | Title | Document Type

Results 1 - 10 of 1220 | Page: 1 2 3 4 5 | Show in sets of: 10 20 50

Email Description *(Please check against the result(s) and click on the "Email Description" to email)*

1. **Romania - Danube Delta Biodiversity Project** | View as HTML
   Type: Implementation Completion Report. **Date:** 10-Jul-2005
   Biodiversity Biological diversity - or biodiversity - refers to the variety of lifeon Earth, and is often categorized at the levels of genetics, species, and ...

2. **World Bank Group | Biodiversity** (Request)
   Type: Visual Material. **Date:** 10-Jul-2005
   Biodiversity and Environmental Assessment. Score Card to Assess Progress in Achieving Management Effectiveness Goals for Marine Protected Areas, 2004

3. **Protecting Asia's Biodiversity – From Crouching ...**
   Type: Press Release. **Date:** 09-Jul-2005
   Protecting Asia's Biodiversity – From Crouching Tigers to Hidden Langurs. World Bank launches new report responding to East Asia's **Biodiversity** Challenges. ...

4. **News - Biodiversity 'Hotspots' are of New Partnership**
   Type: News Story. **Date:** 09-Jul-2005
   News highlights on development issues ... Critical Ecosystem Partnership Fund Will Better Protect **Biodiversity** Hotspots. A new $150 million fund ...

5. **Upcoming Biodiversity Meetings**
   Type: Event. **Date:** 08-Jul-2005
   Access to the Annual Meetings venues and related events ... the DM and its upcoming Global Competition please ... protected areas and **biodiversity** conservation. It will ...

6. **Uruguay: Integrated Natural Resources and Biodiversity ...**
   Type: Project Appraisal Document. **Date:** 08-Jul-2005
   Uruguay: Integrated Natural Resources and **Biodiversity** Management Project. WASHINGTON ...

Try these other searches:

- Coastal biodiversity
- Dispersal biodiversity
- Domestic biodiversity
- Forest biodiversity
- Freshwater biodiversity

Network Taxonomy

Results for "biodiversity"

Search | New Search | Search Help | Saved Search

Did you mean: "biodiversity"

Narrow your search by:

- All
- Country
- Document Type
- Topic

Try these other searches:

- Coastal biodiversity
- Dispersal biodiversity
- Domestic biodiversity
- Forest biodiversity
- Freshwater biodiversity

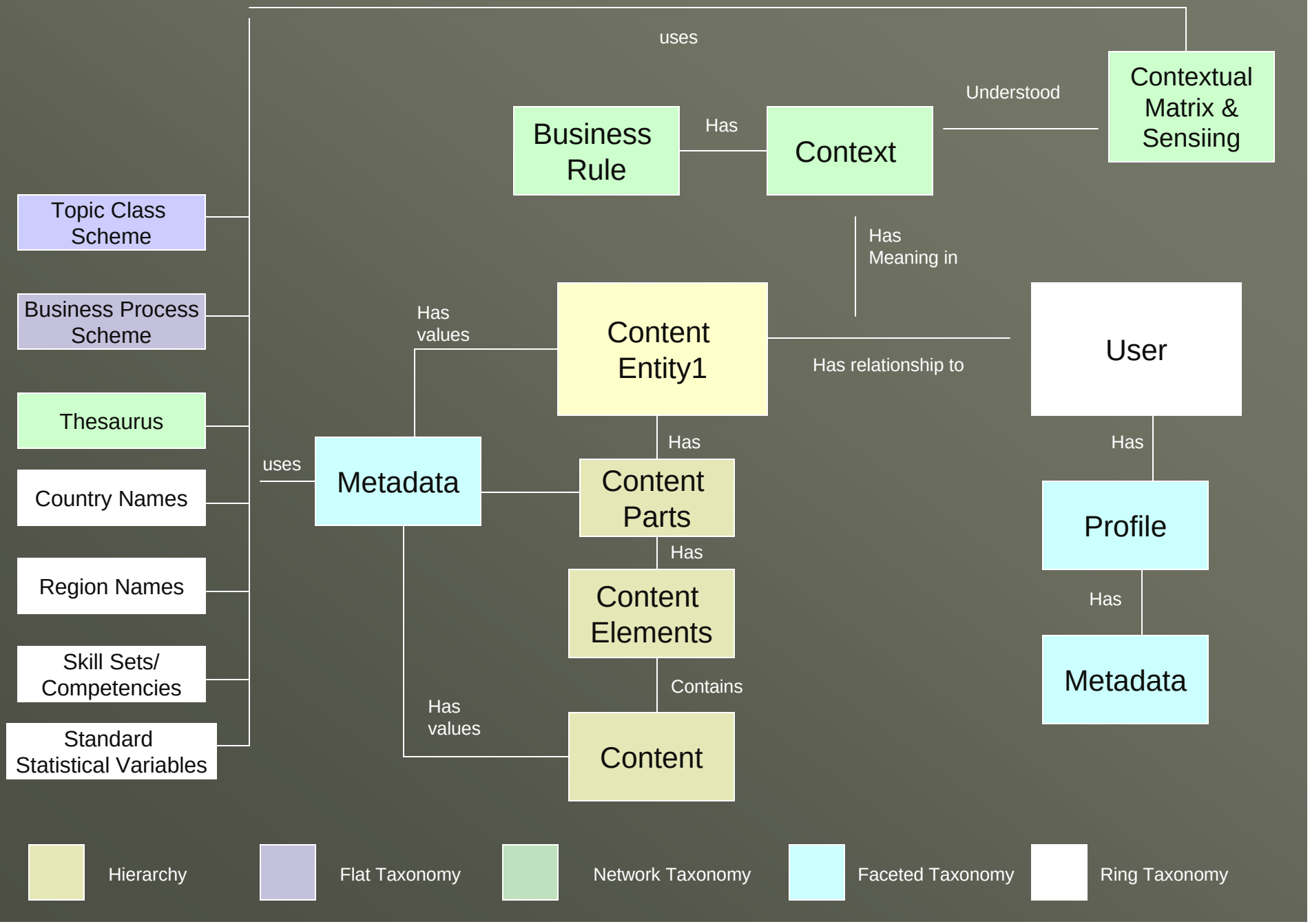
More explicit View of faceted taxonomy

Sort results by: Relevance | Date | Title | Document Type

Email Description | Results 1 - 10 of 1220 | Page: 1 2 3 4 5 | Show in sets of: 10 20 50

Save	Title	Date	Type	Description	Info
1.	Romania - Danube Delta Biodiversity Project	12-Jul-2005	Implementation Completion Report	**Biodiversity** Biological diversity - or biodiversity - refers to the variety of lifeon Earth, and is often categorized at the levels of genetics, species, and ...	
2.	World Bank Group | Biodiversity (Hardcopy)	10-Jul-2005	Visual Material	**Biodiversity** and Environmental Assessment. Score Card to Assess Progress in Achieving Management Effectiveness Goals for Marine Protected Areas, 2004	
3.	Protecting Asia's Biodiversity – From Crouching ... | View as HTML	10-Dec-2004	Press Release	Protecting Asia's Biodiversity – From Crouching Tigers to Hidden Langurs. World Bank launches new report responding to East Asia's **Biodiversity** Challenges. ...	
4.	News - Biodiversity 'Hotspots' are of New Partnership	08-Jan-2004	News Story	News highlights on development issues ... Critical Ecosystem Partnership Fund Will Better Protect **Biodiversity** Hotspots. A new $150 million fund ...	

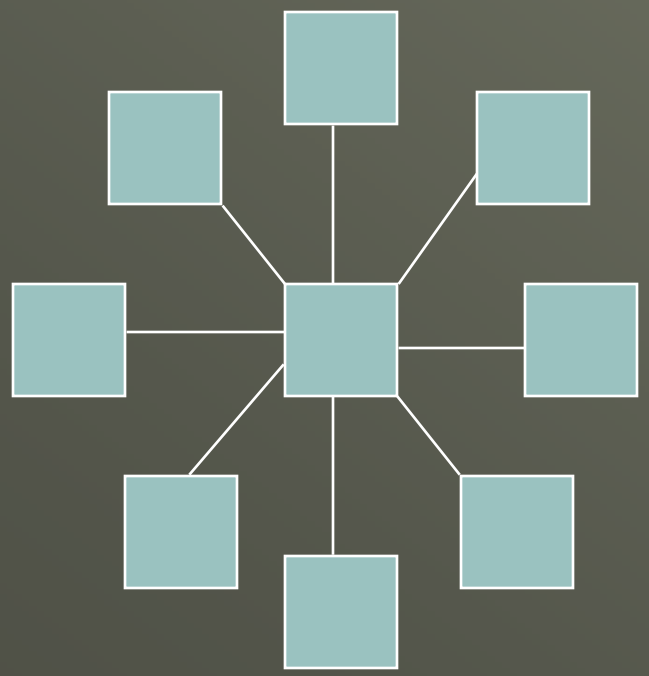
Ontology Design from 50,000 Feet



Taxonomies at Work in the Ontology

- Controlled vocabularies – guiding users through the search process (flat taxonomies)
- Classification schemes – browsing, navigation, syndication, contextualization (hierarchical taxonomies)
- Metadata – supporting fielded search (faceted taxonomies)
- Thesauri – supporting knowledge discovery, preventing Zero results, supporting cross-language searching (network taxonomies)
- Synonym Rings – improving relevancy, reducing information scattering, managing recall (ring taxonomies)

Facet Taxonomies



Faceted taxonomy represented as a star data structure. Each node in the start structure is liked to the center focus.
Any node can be linked to other nodes in other stars.
Appears simple, but becomes complex quickly.

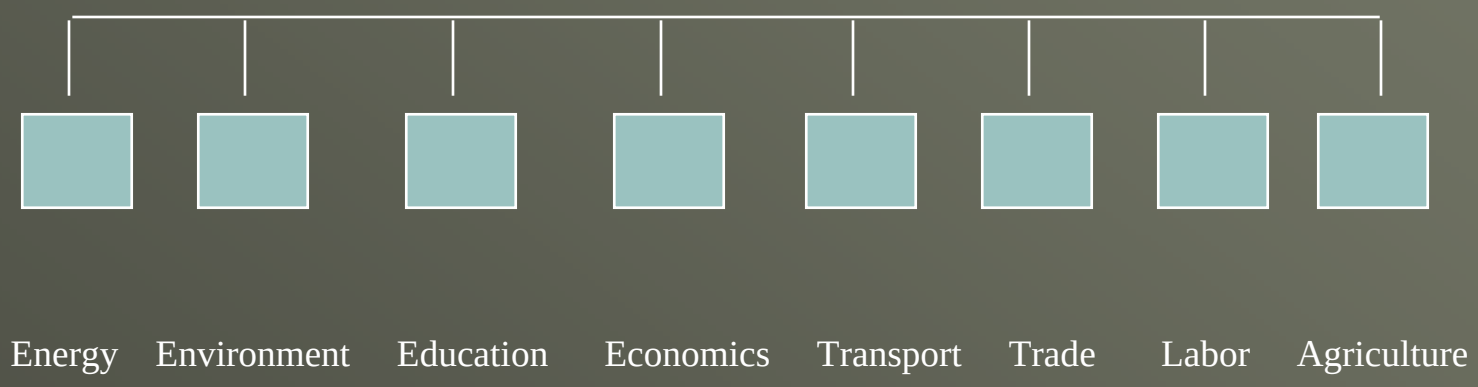
Core Metadata Strategy

- What is a core metadata strategy?
- What's the process you use to discover your organization's core metadata strategy?
- Libraries have many core metadata standards – COSATI, Dublin Core, MARC, MODS, COSATI, AIIM TR48
- The important question for metadata in search is how are you using your metadata to support search?
- Too often we put a 'dumb' search engine on top of 'smart metadata' and do nothing more with it than 'publish' it
- It's time to think smarter about how we use our metadata

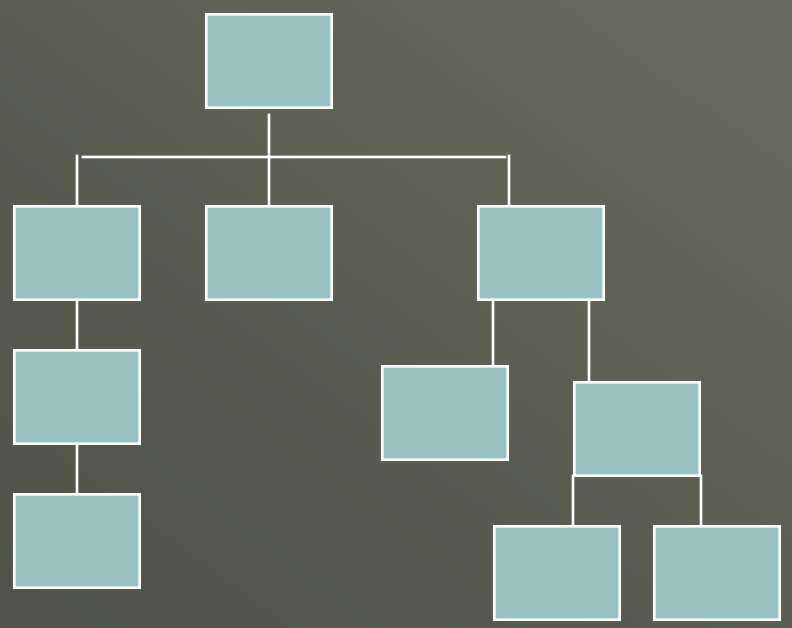
Purpose of Bank Metadata

Identification/ Distinction	Search & Browse	Use Management	Compliant Document Management
Agent	Country	Authorized By	Record Identifier
Title	Region	Rights Management	Disposal Status
Date	Abstract/ Summary	Access Rights	Disposal Review Date
Format	Keywords	Location	Management History
Publisher	Subject-Sector-Theme-Topic	Use History	Retention Schedule/Mandate
Language	Business Function		Preservation History
Version		Disclosure Status	Aggregation Level
Series & Series #		Disclosure Review Date	Relation
Content Type			

Flat Taxonomy Structure



Hierarchical Taxonomy

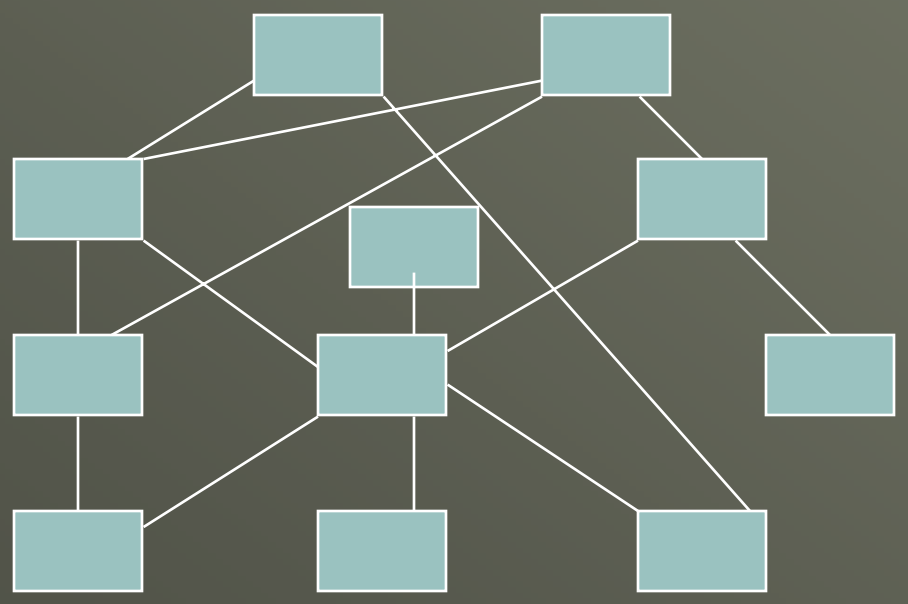


A hierarchical taxonomy is represented as a tree data structure in a database application. The tree data structure consists of nodes and links. In an RDBMS environment, the relationships become associations. In a hierarchical taxonomy, a node can have only one parent.

Hierarchical Taxonomies – Classification Schemes

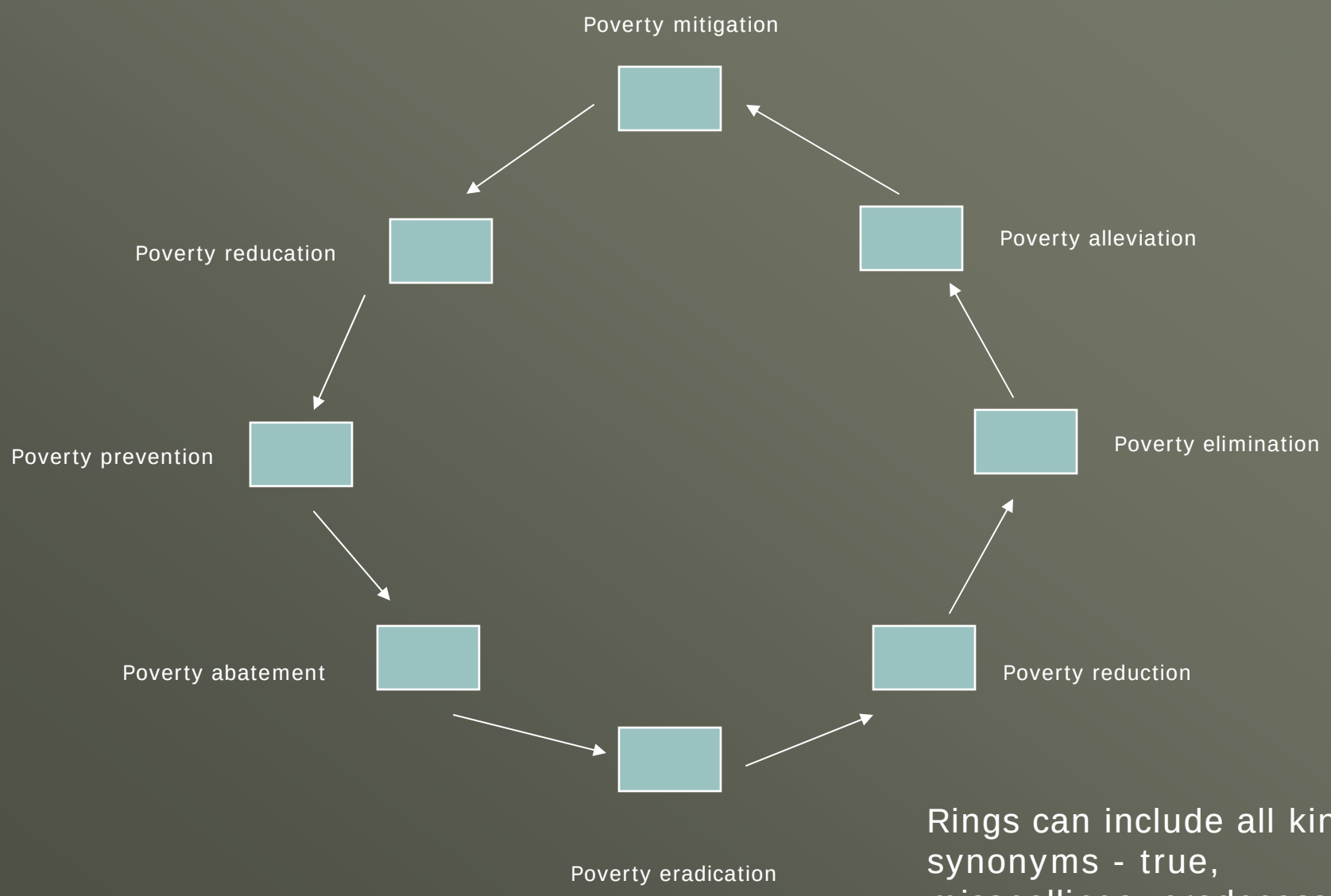
- Ranganathan is the supreme authority on how to create a well-design classification scheme
- Most of what we teach in graduate school, though, is how to use a classification scheme not how to create one
- Classification schemes are the controlling reference source for metadata attributes
- For example, the Enterprise Topic Classification Scheme is the reference source for the attribute Topic
- We use tools to help us discover what the scheme should be and also to help us classify content

Network Taxonomies



A network taxonomy is a plex data structure. Each node can have more than one parent. Any item in a plex structure can be linked to any other item. In plex structures, links can be meaningful & different.

Ring Taxonomy



Rings can include all kinds of synonyms - true, misspellings, predecessors, abbreviations

Fueling Semantic Search With Metadata

Or, ….if Metadata is Dead, Semantic Web and Semantic Search Are Dead

Building and Maintaining Taxonomies

- Moving towards automated metadata generation means that catalogers shift their effort to reviewing the metadata generated and to more fully developing and maintaining subject headings/thesauri and classification schemes as part of a suite of categorization tools
- Level of effort shifts to training and developing the tools and away from original cataloging and metadata capture
- Continue to work closely with subject experts to define the controlled vocabularies and classification schemes
- It means that you have to have a metadata infrastructure that looks something like that ontology we just reviewed
- There is no silver bullet ontology tool out there that will do this work for you – your knowledge and skills are critical

Metadata Capture Methods

Identification / Distinction	Search & Browse	Use Management	Compliant Document Management
Agent	Country	Authorized By	Record Identifier
Title	Region	Rights Management	Disposal Status
Date	Abstract/ Summary	Access Rights	Disposal Review Date
Format	Keywords	Location	Management History
Publisher	Subject-Sector-Theme-Topic	Use History	Retention Schedule/Mandate
Language	Business Function		Preservation History
Version			Aggregation Level
Series & Series #			Relation
Content Type			

Human Capture

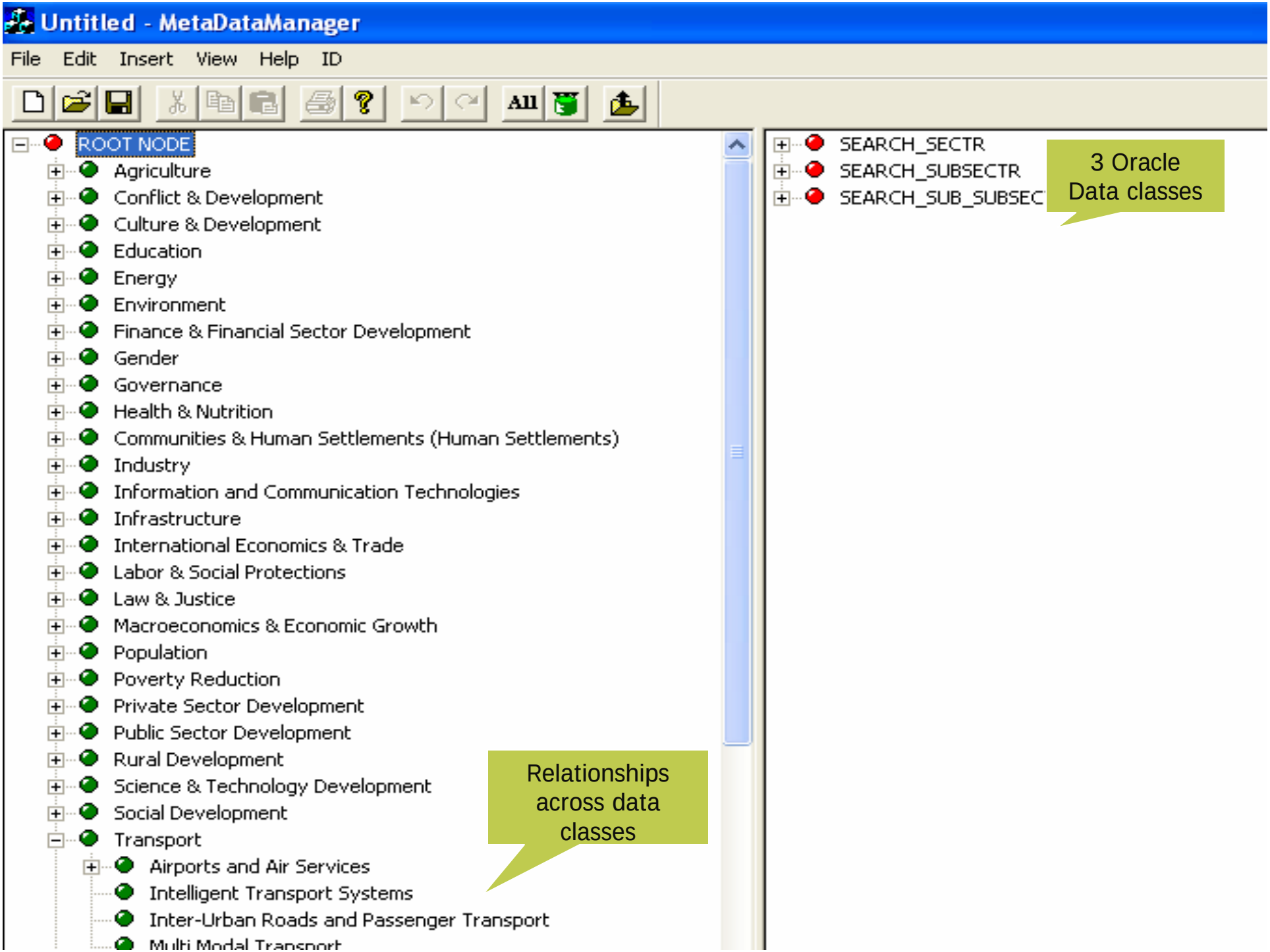
Programmatic Capture

Extrapolate from Business Rules

Inherit from System Context

Smart Use of Technologies

- Sample structure – Bank Topics Classification Scheme (hierarchical taxonomy)
  - Oracle data classes used to represent Topic Classification scheme
    - hierarchical taxonomy as reference source for the attribute – Topic
    - used for Browse, Search, Content Syndication, Personalization
  - 1st challenge is to architect the hierarchy correctly
    - 3 distinct data classes, not a tree structure with inheritance
    - Allows you to use the three data classes for distinct functions across systems but still enforce relationships across the classes


Untitled - MetaDataManager
File Edit Insert View Help ID
All
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 Development
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_SUBSEC
3 Oracle Data classes
Relationships across data classes

Untitled - MetaDataManager

File Edit Insert View Help ID

ROOT NODE

SEARCH_SECTR

- Infrastructure
- Water Resources
- 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
- International Economics & Trade
- Labor & Social Protections
- Law & Justice
- Macroeconomics & Economic Growth
- Population
- Poverty Reduction
- Private Sector Development
- Public Sector Development
- Rural Development
- Science & Technology Development
- Social Development
- Transport
- Urban Development
- Water Supply & Sanitation

Topic data class

- ROOT NODE

- SEARCH_SUBSECTR
  - A
    - AIDS HIV
    - Access of Poor to Social Services
    - Access to Markets
    - Adolescent Health
    - Agricultural Growth and Rural Development
    - Agricultural Trade
    - Alcohol and Substance Abuse
    - Anti-Money Laundering
    - Arbitration
    - Archives Management
    - Artisans
    - Access & Equity in Basic Education
    - Accommodation & Tourism Industry
    - Administrative & Civil Service Reform
    - Administrative & Civil Service Reform (Administrative & Civil Service Reform - 19.08.00)
    - Administrative & Regulatory Law
    - Adult Outreach
    - Agribusiness
    - Agribusiness & Markets
    - Agricultural Extension
    - Agricultural Industry
    - Agricultural Irrigation and Drainage
    - Agricultural Knowledge & Information Systems
    - Agricultural Knowledge & Information Systems (Agricultural Knowledge & Information Systems -
    - Agricultural Producer Organizations
    - Agricultural Research
    - Agricultural Sector Economics
    - Agriculture & Farming Systems
    - Air Quality & Clean Air
    - Airports and Air Services
    - Anthropology
    - Arts & Music
  - B

Subtopic Data Class

Untitled - MetaDataManager

File Edit Insert View Help ID

- ROOT NODE

- SEARCH_SECTR
- SEARCH_SUBSECTR
- SEARCH_SUB_SUBSECTR
  - A
    - Access to Municipal Financial Markets
    - Access to Urban Housing
    - Adult Education Approaches and Program
    - Air Services Competition and Regulation
    - Air Traffic Management
    - Air Transport Infrastructure Finance
    - Air Transport Security
    - Aircraft and Route Licensing
    - Airports and Land Access
    - Alternative Education
    - Artisanal Mining
  - B
  - C
  - D
  - E
  - F
  - G
  - H
  - I
  - J
  - L
  - M
  - N
  - O
  - P
  - Q

Subsubtopic Data class

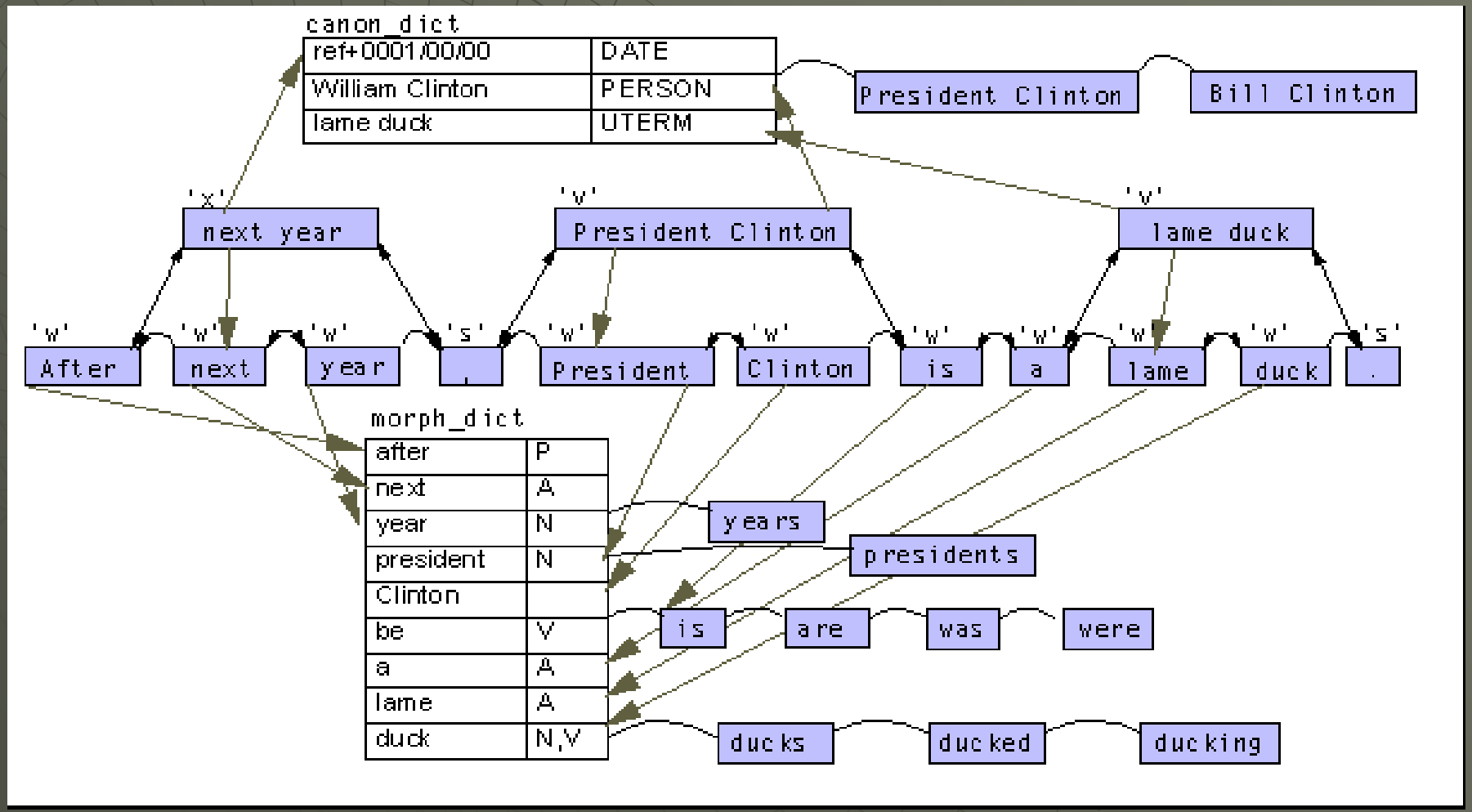
Categorizing Content

- Let's look at how we're categorizing our content to this structure automatically
- Topic classification, geographical region assignment, keywording examples
- Can apply this approach to any kind of content
- Enables us to build a robust metadata repository model, with strong metadata quality, to move towards SI at the functional level
- Also note that we can do this across many languages

Semantic Analysis – Using The Technologies to Best Advantage

- Semantic analysis tools which support concept extraction, categorization, summarization and pattern matching rules engines
- Teragram works in 23 languages
- Use categorization to capture Topics, Business Activities, Regions, Sectors, Themes, etc.
- Use Concept Extraction to capture keywords
- Use Rules Engine to capture Loan #, Credit #, Project ID, Trust Fund #, etc.
- Use Summarization to generate a ‘gist’ of the content

How does semantic analysis work?



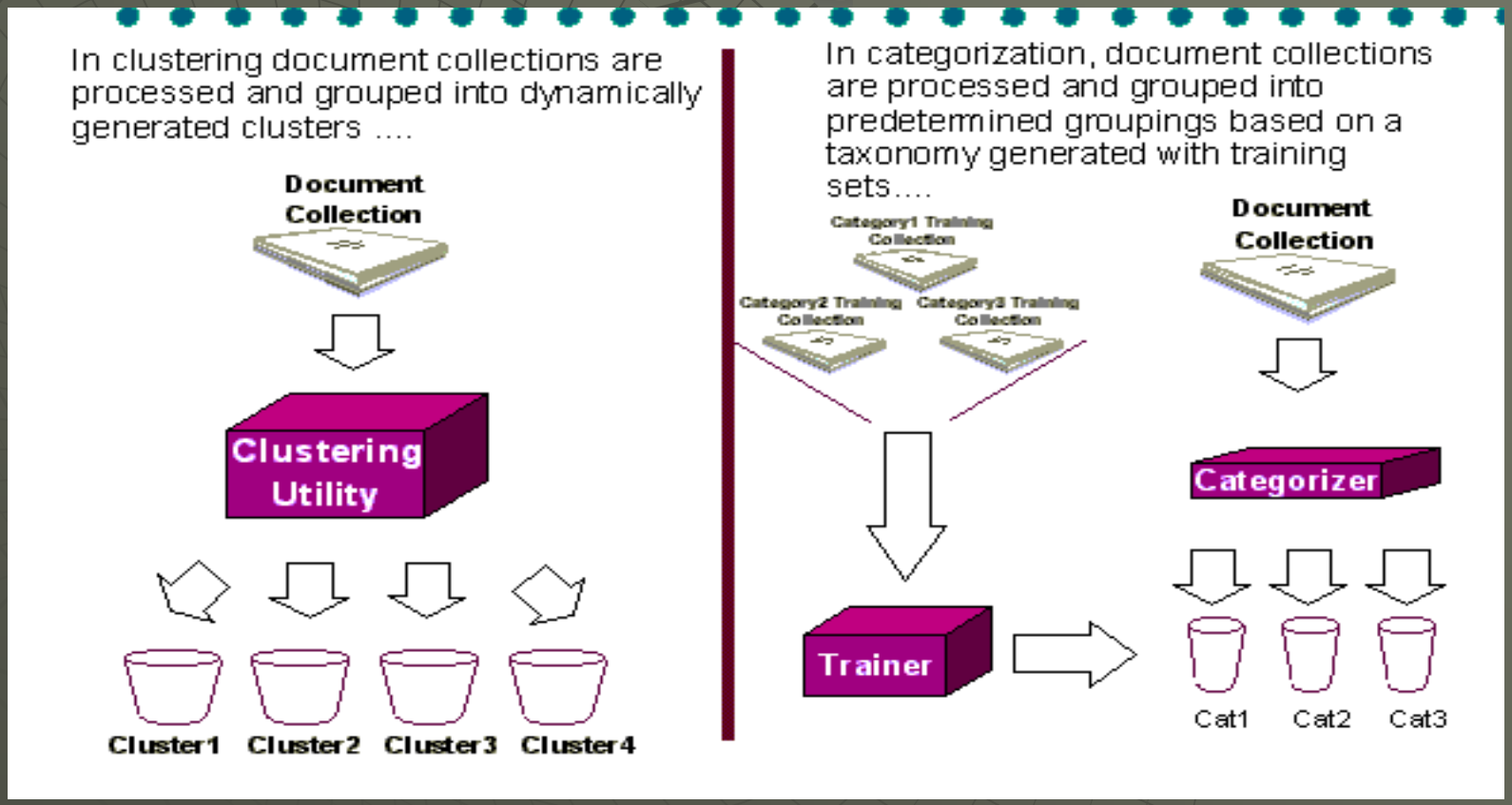
Semantic Analysis Basics

- Once you have made some sense of the sentence, reconstruct entities for information extraction (compose)
  - Identify names and other fixed form expressions – people, organizations, conferences
  - Identify basic noun groups, verb groups, presentations, other grammatical elements
  - Construct complex noun groups and verb groups
  - Identify event structures
  - Identify common elements and associate

Clustering vs. Categorization

- Clustering

Categorization



Leveraging the Structure

- Each subtopic is a knowledge domain (hierarchical taxonomy)
- Each subtopic has an extensive concept level definition (1,000 – 5,000+ concepts)
- Concepts are controlled vocabularies in their raw form (flat taxonomy)
- Concepts with relationships (extensive per new Z39.19 standard) comprise semantic network (network taxonomy)
- Categorization tools work with topic structure & concept definitions to categorize and index content
- The following screen illustrates how that same structure is embedded into Teragram profile to support categorization

Topics.tk2 - Teragram TK240

File Edit View Build Project Category Concept Testing Server Help

Topics
English
Categorizer
Top
Topics

Subtopics

644279 - Macroeconomics & Economic Gro
644280 - Social Development
644281 - Culture & Development
644282 - Law & Justice
644283 - Governance
644284 - Communities & Human Settlemer
644285 - Private Sector Development
644286 - Industry
644287 - Water Supply & Sanitation
644288 - Environment
644289 - Science & Technology Innovation
644290 - Agriculture
672784 - Agriculture and Farming Syster
672785 - Agribusiness
672786 - Agricultural Extension
672787 - Agricultural Producer Organizat
672788 - Dairies & Dairying
672789 - Crops & Crop Management Sy
672790 - Fertilizers
672791 - Livestock & Animal Husbandry
672792 - Pest Management
672793 - Agricultural Irrigation
672794 - Agricultural Research
672795 - Fisheries & Aquaculture
672797 - Agricultural Knowledge & Inforr
672798 - Agricultural Economics
672854 - Forestry

Taxonomy | Dependencies

```
(OR,(OR,"Abattoirs"),(OR,"AAT"),(OR,"aborted cows"),(OR,"absentee farming"),
(OR,"Animal health projects"),(OR,"animal health requirements"),(OR,"animal h
(OR,"animal tenure rights"),(OR,"Animal textile fibers"),(OR,"animal traction
(OR,"breeds farmers"),(OR,"Broiler chickens"),(OR,"Broilers poultry"),(OR,"Br
fat"),(OR,"crude fiber"),(OR,"crude fibre"),(OR,"crude form"),(OR,"crude prot
(OR,"family-farming sector"),(OR,"farm"),(OR,"Farm animals"),(OR,"farm budget
alternatives"),(OR,"fodder availability"),(OR,"fodder banks"),(OR,"fodder cos
(OR,"Grazing intensity"),(OR,"grazing land"),(OR,"Grazing lands"),(OR,"grazin
pig production"),(OR,"Intensive pork"),(OR,"intensive poultry production unit
(OR,"livestock groups"),(OR,"Livestock guts"),(OR,"livestock head per farm"),
(OR,"livestock sector supply"),(OR,"livestock sectors"),(OR,"livestock servic
products"),(OR,"meat import"),(OR,"meat importing countries"),(OR,"meat indus
(OR,"pastoral ecology"),(OR,"pastoral economies"),(OR,"pastoral ecosystems"),
animals"),(OR,"protein-energy malnutrition"),(OR,"protein-rich food supplies"
herders"),(OR,"Sheep industry"),(OR,"Sheep meat"),(OR,"sheep population"),(OR
production"),(OR,"sustainable pastoralism"),(OR,"Swine"),(OR,"Swine equipment
(OR,"Veterinary hygiene"),(OR,"veterinary inputs"),(OR,"veterinary inspection
production"),(OR,"abundant vegetation"),(OR,"additional feed grain requiremen
(OR,"animal production efficiency"),(OR,"animal production enterprises"),(OR,
(OR,"cattle breeding"),(OR,"cattle cattle"),(OR,"cattle development"),(OR,"ca
gain"),(OR,"dairy cattle"),(OR,"dairy cattle producers"),(OR,"dairy cooperat
(OR,"feed imports"),(OR,"feed improvement"),(OR,"feed intake"),(OR,"feed inta
(OR,"grazing lease markets"),(OR,"grazing livestock"),(OR,"grazing management
(OR,"livestock development specialists"),(OR,"Livestock Development Strategie
(OR,"livestock traders"),(OR,"livestock transport"),(OR,"livestock units"),(
fertilization"),(OR,"organic fertilizer"),(OR,"organic inputs"),(OR,"organic
Pastoralists"),(OR,"poor livestock development strategies"),(OR,"poor livesto
(OR,"sown forage crops"),(OR,"sown forage production"),(OR,"sown forages"),(
(OR,"veterinary clinics"),(OR,"veterinary coverage"),(OR,"veterinary delivery
```

Domain concepts or controlled vocabulary

Syntax Check | Indent | Text View | Tree View | Load Text... | Expand Forms | Server Query... | Ln 30

Rules | Testing | Data | Document

Ready | NUM

start | > TK240 license - Lot... | Untitled - MetaDataM... | Microsoft PowerPoint ... | Topics.tk2 - Teragra... | EN | 3:25 PM Monday 06/27/2005

Topics.tk2 - Teragram TK240

File Edit View Build Project Category Concept Testing Server Help

- Topics
  - English
    - Categorizer
      - Top
        - Topics
          - 644279 - Macroeconomics & Economic Gro
          - 644280 - Social Development
          - 644281 - Culture & Development
          - 644282 - Law & Justice
          - 644283 - Governance
          - 644284 - Communities & Human Settlemer
          - 644285 - Private Sector Development
          - 644286 - Industry
          - 644287 - Water Supply & Sanitation
          - 644288 - Environment
          - 644289 - Science & Technology Innovation
          - 644290 - Agriculture
            - 672784 - Agriculture and Farming Syster
            - 672785 - Agribusiness
            - 672786 - Agricultural Extension
            - 672787 - Agricultural Producer Organizat
            - 672788 - Dairies & Dairying
            - 672789 - Crops & Crop Management Sy
            - 672790 - Fertilizers
            - 672791 - Livestock & Animal Husbandry
            - 672792 - Pest Management
            - 672793 - Agricultural Irrigation
            - 672794 - Agricultural Research
            - 672795 - Fisheries & Aquaculture
            - 672797 - Agricultural Knowledge & Inforr
            - 672798 - Agricultural Economics
            - 672854 - Forestry

Taxonomy | Dependencies

- OR
  - OR
    - "Abattoirs"
  - OR
    - "AAT"
  - OR
    - "aborted cows"
  - OR
    - "absentee farming"
  - OR
    - "absentee herd owners"
  - OR
    - "absentee herd ownership"
  - OR
  - OR
    - "African animal trypa
  - OR
    - "Aftosa fever"
  - OR
    - "Agribusiness servic
  - OR
    - "Agricultural animals
  - OR
    - "agricultural brucello
  - OR

Add Statement
Paste Macro
Add Operator
Change Operator
Delete Node

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

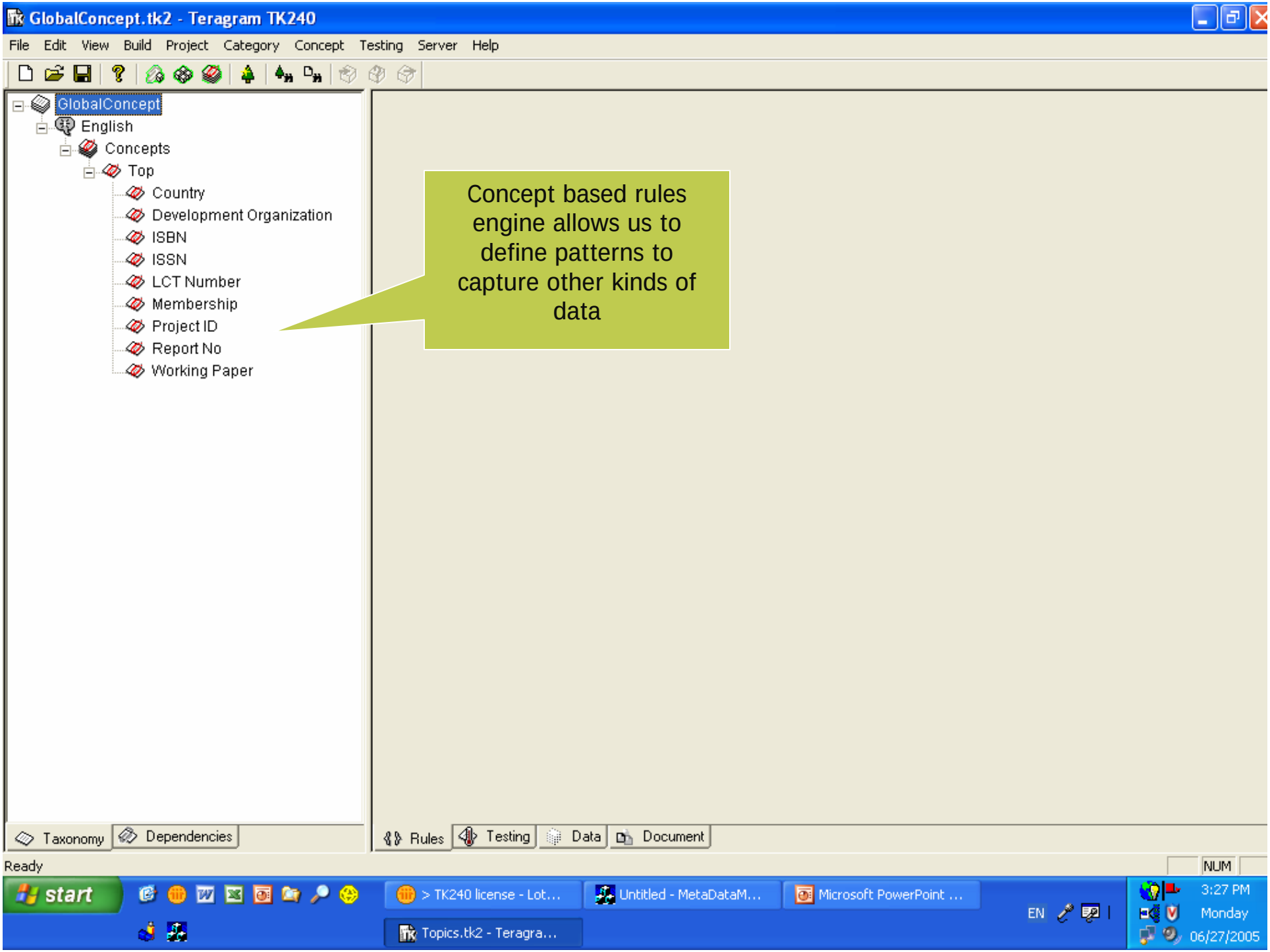
Extensive operators allow us to write grammatical rules to manage typical semantic problems

Expand Forms | Server Query... | Ln 30

Rules | Testing | D

Ready | NUM

start | > TK240 license - Lot... | Untitled - MetaDataM... | Microsoft PowerPoint ... | Topics.tk2 - Teragra... | EN | 3:26 PM Monday 06/27/2005


GlobalConcept.tk2 - Teragram TK240
File Edit View Build Project Category Concept Testing Server Help
GlobalConcept
English
Concepts
Top
Country
Development Organization
ISBN
ISSN
LCT Number
Membership
Project ID
Report No
Working Paper
Concept based rules engine allows us to define patterns to capture other kinds of data
Taxonomy
Dependencies
Rules
Testing
Data
Document
Ready
NUM
start
> TK240 license - Lot...
Untitled - MetaDataM...
Microsoft PowerPoint ...
Topics.tk2 - Teragra...
EN
3:27 PM
Monday
06/27/2005

GlobalConcept.tk2 - Teragram TK240

File Edit View Build Project Category Concept Testing Server Help

- GlobalConcept
  - English
    - Concepts
      - Top
        - Country
        - Development Organization
        - ISBN
        - ISSN
        - LCT Number
        - Membership
        - Project ID
        - Report No
        - Working Paper

```
Antigua/and/Barbuda, Antigua and Barbuda
Antigua/y/Barbuda, Antigua and Barbuda
Antigua/e/Barbuda, Antigua and Barbuda
Antigua/-/et/-/Barbuda, Antigua and Barbuda
Antigua/und/Barbuda, Antigua and Barbuda
Antigua/&/Barbuda, Antigua and Barbuda
Andorra, Andorra
Anguilla, Anguilla
Argentina, Argentina
Argentine/Republic, Argentina
Argentinië, Argentina
République/argentine, Argentina
Rio/de/la/Plata, Argentina
Armenia, Armenia
Hayastan, Armenia
Hayastani/Hanrapetut/'/yun, Armenia
Huyasdan, Armenia
Armenija, Armenia
Republic/of/Armenia, Armenia
Armianskaïa, Armenia
Armyanskaya, Armenia
Arménie, Armenia
République/d/'/Arménie, Armenia
Arm/'/anskaja/Sovetskaja/Socialisticeskaja/Respublika, Armenia
Armenian/SSR, Armenia
Armenian/Soviet/Socialist/Republic, Armenia
Armeniya/SSR, Armenia
Sovyetakan/Hayastan, Armenia
Russian/Armenia, Armenia
Soviet/Armenia, Armenia
Aruba, Aruba
Australia, Australia
Commonwealth/of/Australia, Australia
Australie, Australia
```

Example of use of Authority Control to capture country names but extract 'authorized' version of country name

Example of use of a gazetteer + concept extraction + rules engine to support semantic interoperability

Syntax Check | Classifier | Grammar | Filename ... | Load Text...

Taxonomy | Dependencies | Definition | Testing | Data | Document

Ready | NUM

start | > TK240 license - Lot... | Untitled - MetaDataM... | Microsoft PowerPoint ... | Topics.tk2 - Teragra... | EN | 3:28 PM Monday 06/27/2005

GlobalConcept.tk2 - Teragram TK240

File Edit View Build Project Category Concept Testing Server Help

- GlobalConcept
  - English
    - Concepts
      - Top
        - Country
        - Development Organization
        - ISBN
        - ISSN
        - LCT Number
        - Membership
        - Project ID
        - Report No
        - Working Paper

```
__REGEX__
TF[0-9]{6},TF022714
TF\-[0-9]{5},TF-26673
Credit\s+[0-9]+[A-Z]+,Credit
Loan\s+[0-9]+[A-Z]+,Loan
Trust\s+Funds\s+[0-9]+[A-Z]+,Trust Funds
Credit\s+No.:\s+[0-9]+[A-Z]+,Credit No.:
Loan\s+No.:\s+[0-9]+[A-Z]+,Loan No.:
Trust\s+Funds\s+No\.:\s+[0-9]+[A-Z]+,Trust Funds No.:
Credit\s+no\.\s+[0-9]+\s+\-\s+[A-Z][A-Z],Credit no.
Credit\s+no\.\s+[0-9]+\s+\-\s+[A-Z][A-Z][A-Z],Credit no.
Loan\s+no\.\s+[0-9]+\s+\-\s+[A-Z][A-Z],Loan no.
Loan\s+no\.\s+[0-9]+\s+\-\s+[A-Z][A-Z][A-Z],Loan no.
Trust\s+Funds\s+no\.\s+[0-9]+\s+\-\s+[A-Z][A-Z],Trust Funds no.
Trust\s+Funds\s+no\.\s+[0-9]+\s+\-\s+[A-Z][A-Z][A-Z],Trust Funds no.
Credit\s+[0-9]+\-[A-Z][A-Z][A-Z]?,Credit
Loan\s+[0-9]+\-[A-Z][A-Z][A-Z]?,Loan
Trust\s+Funds\s+[0-9]+\-[A-Z][A-Z][A-Z]?,Trust Funds
```

Use of concept extraction + rules engine to capture Loan #, Credit #, Project ID#

Syntax Check | Classifier | Grammar | Filename ... | Load Text...

Taxonomy | Dependencies | Definition | Testing | Data | Document

Ready | NUM

start | > TK240 license - Lot... | Untitled - MetaDataM... | Microsoft PowerPoint ... | Topics.tk2 - Teragra... | EN | 3:28 PM Monday 06/27/2005

Romania - Danube Delta Biodiversity Project

Archives Accession No:	A2005-001	**Archives Box No:**	85
Bank Group Institution:	IBRD	**Country:**	Romania;
Date Stored:	2005/07/15	**Document Date:**	2005/06/30
Document Type:	Project Performance Assessment Report	**Document Version:**	Final
Geographic Region:	Eastern Europe; Europe; World	**Language:**	English
Lending Instrument:	Specific Investment Loan	**Major Sector:**	Law and justice and public administration; Agriculture, fishing, and forestry; Health and other social services
Product Line Code:	Global Environment Project	**Profiler:**	Muhammad,Phyllis
Rel. Proj ID:	RO-Danube Delta Biodiversity Gef Project-271505 -- -- P008689;	**Project Status:**	LEND
Region:	Europe and Central Asia	**Rep Title:**	Romania - Danube Delta Biodiversity Project
Report Number:	32684	**Sector:**	Central government administration; General agriculture, fishing and forestry sector; Other social services
Security Classification:	Public	**Task Manager:**	Petrescu,Doina
Trust Fund Name:	TF028614-GET-PPA FOR ROMANIA DANUBE DELTA BIO.; TF028660-GET-ROMANIA DANUBE DELTA BIODIVERSITY	**Unique Entity ID:**	000160016_20050715142011
Unit Owning:	Environ &Social Sustain Dev Unit (ECSSD)	**Volume No:**	1 of 1

Abstract: Project ratings for the Danube Delta Biodiversity Project for Romania are as follows: Project outcome is moderately satisfactory; sustainability is likely; institutional development impact is modest; Bank performance is satisfactory; and Borrower performance is satisfactory. There are four lessons: 1) Biodiversity conservation cannot be carried out in isolation. It has to be integrated within the economic interests of local and regional communities. Resentment is created when financing of nature conservation appears to have preference over unmet local needs, be it employment or delivery of basic services. Failure to integrate local interests in the conservation and management strategy of a biosphere reserve can endanger its longer-term sustainability. 2) Conservation areas will be sustainable only if there is good management and sufficient funding. The Global Environment Facility (GEF) project designers must help establish sound management and governance arrangements that include local stakeholders and promote income-generating activities. 3) When establishing biodiversity reserves, facilitate networking of the reserve staff with the national and international nongovernmental organizations and promote recognition by international conventions. 4) The GEF should move towards a country-focused strategic approach to complement its thematically-driven development framework. By doing so GEF would develop synergy from a more coherent policy framework, thus improving effectiveness and reducing transaction costs.

Keywords:
agriculture, aquatic resources, BIODIVERSITY, biodiversity conservation, Biological Diversity, Biosphere, birds, buffer zones, community development, Conservation, Conservation areas, dams, delta ecosystems, discharge, drainage, economic activity, economic

Abstract: Project ratings for the Danube Delta Biodiversity Project for Romania are as follows: Project outcome is moderately satisfactory; sustainability is likely; institutional development impact is modest; Bank performance is satisfactory; and Borrower performance is satisfactory. There are four lessons: 1) Biodiversity conservation cannot be carried out in isolation. It has to be integrated within the economic interests of local and regional communities. Resentment is created when financing of nature conservation appears to have preference over unmet local needs, be it employment or delivery of basic services. Failure to integrate local interests in the conservation and management strategy of a biosphere reserve can endanger its longer-term sustainability. 2) Conservation areas will be sustainable only if there is good management and sufficient funding. The Global Environment Facility (GEF) project designers must help establish sound management and governance arrangements that include local stakeholders and promote income-generating activities. 3) When establishing biodiversity reserves, facilitate networking of the reserve staff with the national and international nongovernmental organizations and promote recognition by international conventions. 4) The GEF should move towards a country-focused strategic approach to complement its thematically-driven development framework. By doing so GEF would develop synergy from a more coherent policy framework, thus improving effectiveness and reducing transaction costs.

Keywords:
agriculture, aquatic resources, BIODIVERSITY, biodiversity conservation, Biological Diversity, Biosphere, birds, buffer zones, community development, Conservation, Conservation areas, dams, delta ecosystems, discharge, drainage, economic activity, economic development, economic growth, economic value, ecosystem, ecosystem restoration, ecosystems, employment, endangered species, energy efficiency, Environment Protection, environmental conservation, environmental degradation, Environmental Protection, Environmental Protection Agency, erosion, exchange rate, Exploitation, fauna, fish, fisheries, fishing, flora, Forests, freshwater, Geographic Information, geographic information systems, Global Environment, Global Environment Facility, habitat, Habitats, health services, high unemployment, housing, Income, international conventions, international organizations, intervention, inventories, isolation, lakes, land management, land use, laws, legislation, local authorities, migrants, migration, mitigation, natural resource management, natural resources, nature conservation, nature Reserves, pollution, ponds, productivity, protected areas, public awareness, public sector, quotas, RAMSAR, reclamation, recreation, resource use, sand, scientific research, silviculture, social costs, strategic planning, Sustainable Development, sustainable resource management, sustainable resource use, sustainable use, transaction costs, vegetation, waste, waste disposal, water pollution, water resources, Waterfowl, wetland ecosystems, wetland management, wetlands, wetlands management, Wildlife

Official Documents

Official version of document (may contain signatures, etc)

File Type		Description	File Size (mb)
	PDF 41 pages	Official Version	4.29 (approx.)
	Text	Text version*	

Order This Document

Overview of Process & Tools

Activity	*Approach*	*Tools*
Create new facet	Human review & consultation, data structures, governance	Oracle DBMS, in future Metadata Repository tools (ISO 11179); Oracle representation of data classes
Create new class	Human review & harmonization of existing information structures; tool based discovery of new structures through clustering & extraction	Teragram dynamic concept extraction using grammars, categorization, clustering; Oracle representation of data classes
Create new concept	Create training sets working with experts, identify & integrate existing vocabularies	Teragram concept extraction, Oracle representation of values
Create new relationship	Human relationship creation, augmented by technological discovery	Teragram clustering engine, MultiTes Thesaurus Management System, Oracle copy of thesaurus relationships
Create new metadata	Enterprise Profile Development with human review in some cases, no review in others; Metadata in the language of the document/content	Teragram enterprise profile leveraging concept extraction, categorization, and summarizaiton

Enterprise Profile Creation and Maintenance

Update & Change Requests

UCM Service Requests

Data Governance Process for Topics, Business Function, Country, Region, Keywords, People, Organizations, Project ID

e-CDS Reference Sources for Country, Region, Topics Business Function, Keywords, Project ID, People, Organization

Enterprise Profile Development & Maintenance

TK240 Client

Teragram Team

Enterprise Metadata Profile

Concept Extraction Technology
- ✓Country
- ✓Organization Name
- ✓People Name
- ✓Series Name/Collection Title
- ✓Author/Creator
- ✓Title
- ✓Publisher
- ✓Standard Statistical Variable
- ✓Version/Edition

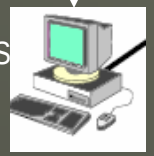
Categorization Technology
- ✓Topic Categorization
- ✓Business Function Categorization
- ✓Region Categorization
- ✓Sector Categorization
- ✓Theme Categorization

Rule-Based Capture
- ✓Project ID
- ✓Trust Fund #
- ✓Loan #
- ✓Credit #
- ✓Series #
- ✓Publication Date
- ✓Language

Summarization

ISP

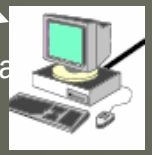
IRIS

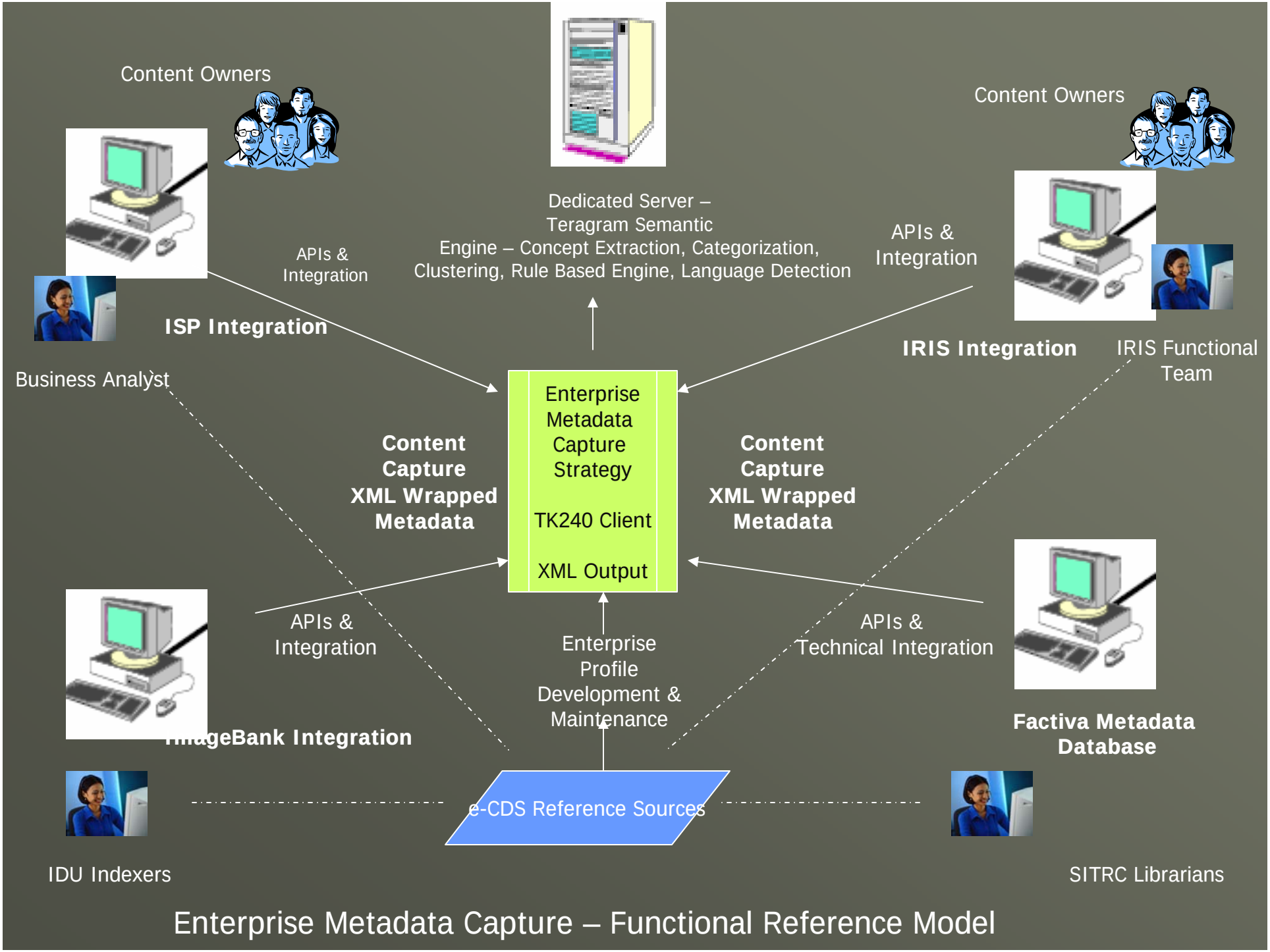


ImageBa

JOLIS E-Journa

Factiva





Enterprise Metadata Capture – Functional Reference Model

Caution Regarding Tools

- Not all tools will do what we describing here
- You need to have an underlying semantic engine which can perform semantic analysis
- You need to have a semantic engine in multiple languages – semantics vary by language
- You need to have access to the programs through a user-friendly interface so you can adapt them to your environment without having to have programming knowledge
- You need to have several different kinds of technologies to do what I'm describing here
- Not all the tools on the market today support this work

Good News

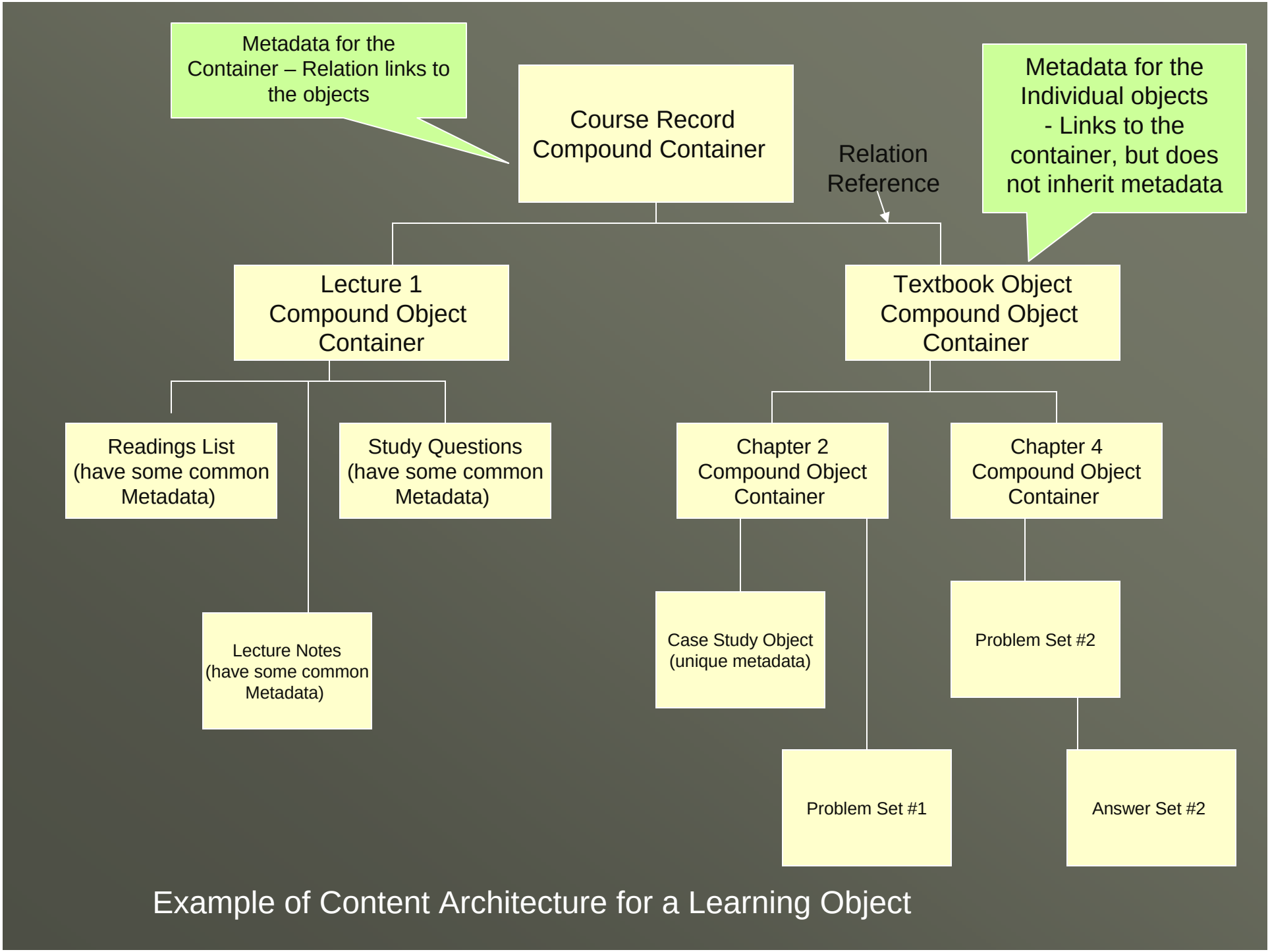
- More and more of the tools are beginning to support this kind of semantic functionality
- More feedback we provide to the vendors, the better the tools will be
- More we try to push the envelope for our clients and users, the more and better functionality we will be able to deliver to them without overwhelming our scarce resources

Impacts & Outcomes

- Information Access impacts
  - Increased precision of search
  - Better control over recall
  - Searching like we talk
  - Exact match searching – known item searching will work better
  - Metadata based searching now begins to resemble full-text searching but with all the advantages of structure & context, and a significant reduction in the amount of noise
- Productivity Improvements
  - Can now assign deep metadata to all kinds of content
  - Remove the human review aspect from the metadata capture
  - Reduce unit times where human review is still used
- Information Quality impacts
  - All metadata carries the information architecture with it
  - Apply quality metrics at the metadata level to eliminate need to build 'fuzzy search architectures' – these rarely scale or improve in performance
  - Use the technologies to identify and fix problems with our data

More Granular Searching

- We had it under control, they change the whole game on us
- Now users want to be able to target the parts of content that they want to search against – not just words against the whole content object
- Content is changing – we now need to pay attention to content architectures
- Context is changing – we now need to know how the information is being used and by whom for what purpose
- We need to begin applying everything we've just talked about at the content 'part' level



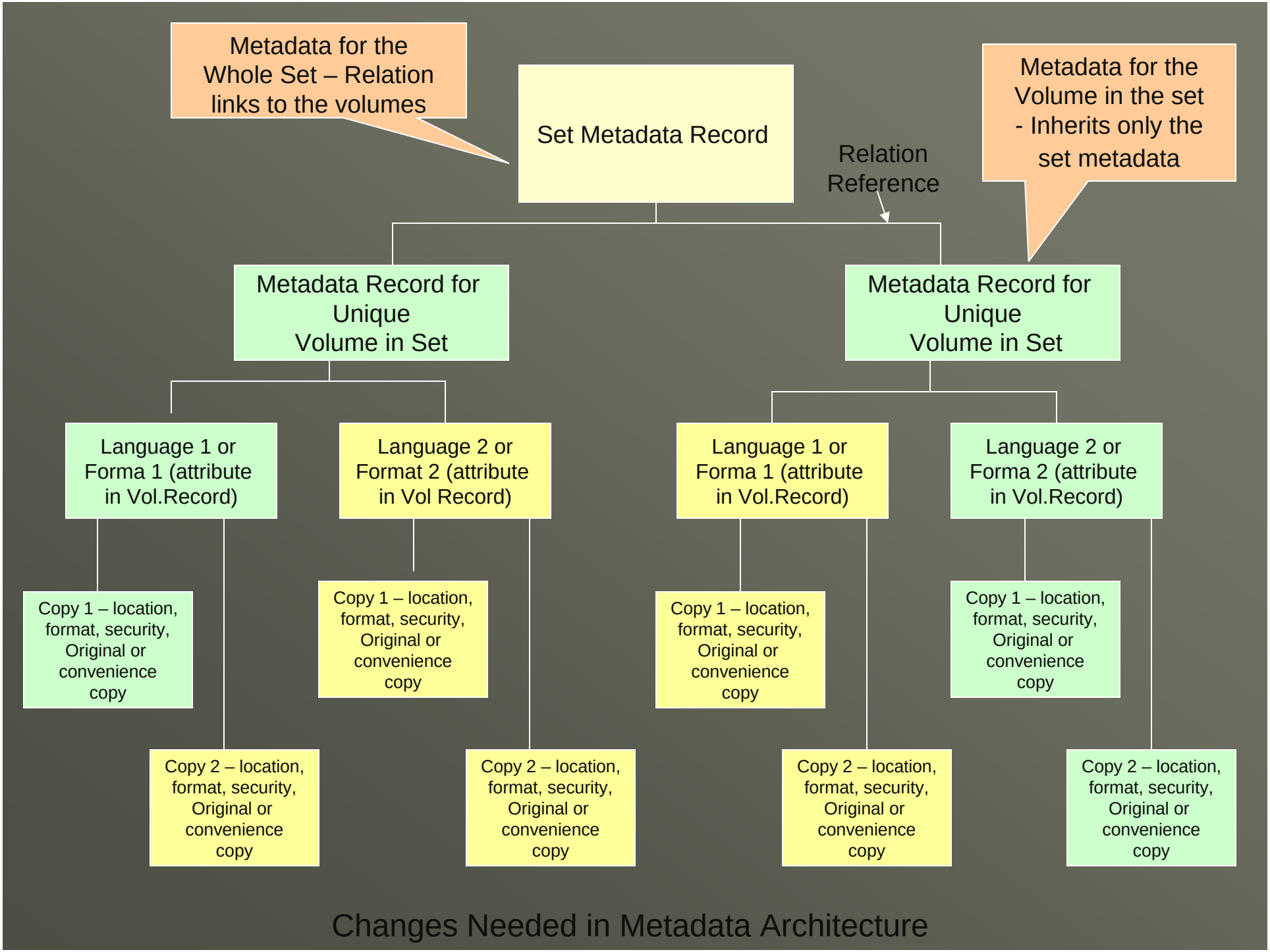
Example of Content Architecture for a Learning Object

Content Architecture

- Each object in the container has its own original metadata
- There are links from some objects to others in the container, but the links are defined by the context in which they are used not based on the source of the original content
- Metadata for the container must describe the compound object, with links to the individual components
- Compound object outside of the context only serves to point users to the location of another instance of the object
- Metadata from other sources can be repurposed to build the container

Metadata Architecture Challenges

- Metadata architecture is not sufficiently agile to handle multiple copies, multiple versions, multiple languages in an elegant and efficient way
- Metadata architecture supports redundancies now
- Metadata architecture not sufficiently agile to identify & distinguish record and convenience copies



Changes Needed in Metadata Architecture

Thank You.

Questions & Discussions