SEMAGIX
POWER • THROUGH • RELEVANCE

Capturing and Applying Existing Knowledge to Semantic Applications
or ***Ontology-driven Information Systems in Action***

Invited Talk
"Sharing the Knowledge"
International CIDOC CRM Symposium
Washington DC, March 26 - 27, 2003

Amit Sheth
Semagix, Inc. and LSDIS Lab, University of Georgia

Syntax -> Semantics

SEMAGIX

Ontology-driven Information Systems are becoming reality

Software and practical tools to support key capabilities and requirements for such a system are now available:

- Ontology creation and maintenance
- Knowledge-based (and other techniques) supporting Automatic Classification
- Ontology-driven Semantic Metadata Extraction/Annotation and
  - Semantic normalization
- Utilizing semantic metadata and ontology
  - Semantic querying/browsing/analysis
  - Information and application integration

Achieved in the context of successful technology transfer from academic research (LSDIS lab, UGA's SCORE technology) into commercial product (Semagix's Freedom)

SEMAGIX

Ontology at the heart of the Semantic Web; Relationships at the heart of Semantics

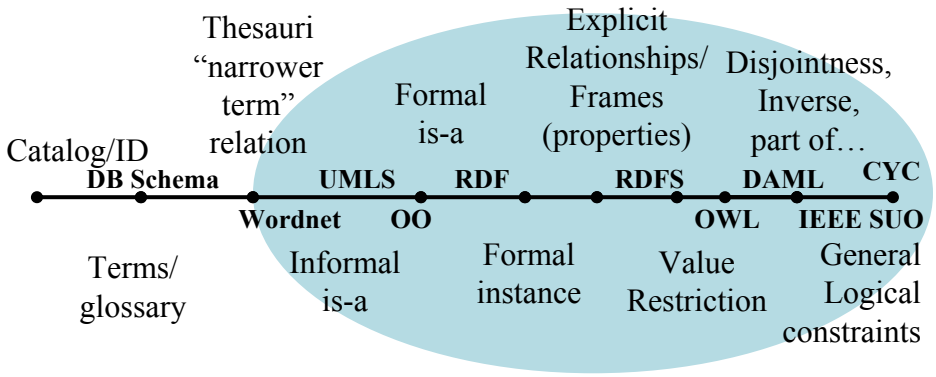
Ontology provides underpinning for semantic techniques in information systems.

- A model/representation of the real world (relevant concepts, entities, attributes, relationships, domain vocabulary and factual knowledge, all connected via a semantic network). Basic of agreement, applying knowledge
- Enabler for improved information systems functionalities and the Semantic Web:
  - Relevant information by (semantic) Search, Browsing
  - Actionable information by (semantic) information correlation and analysis
  - Interoperability and Integration
- Relationships – what makes ontologies richer (more semantic) than taxonomies … see "Relationships at the Heart of Semantic Web: Modeling, Discovering, Validating and Exploiting Complex Semantic Relationship

POWER • THROUGH • RELEVANCE

SEMAGIX

Increasingly More Semantic Representation



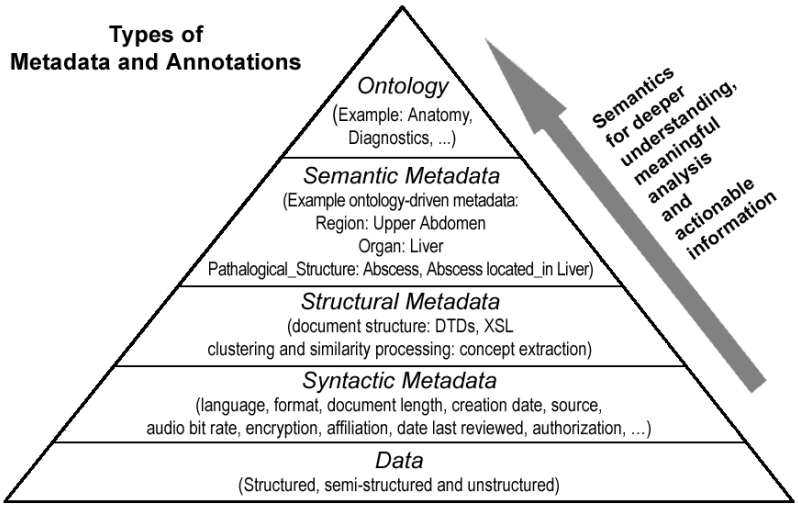
Simple Taxonomies ⟷ Expressive Ontologies

Better capability at higher complexity and computability

After McGuinness & Finin

SEMAGIX

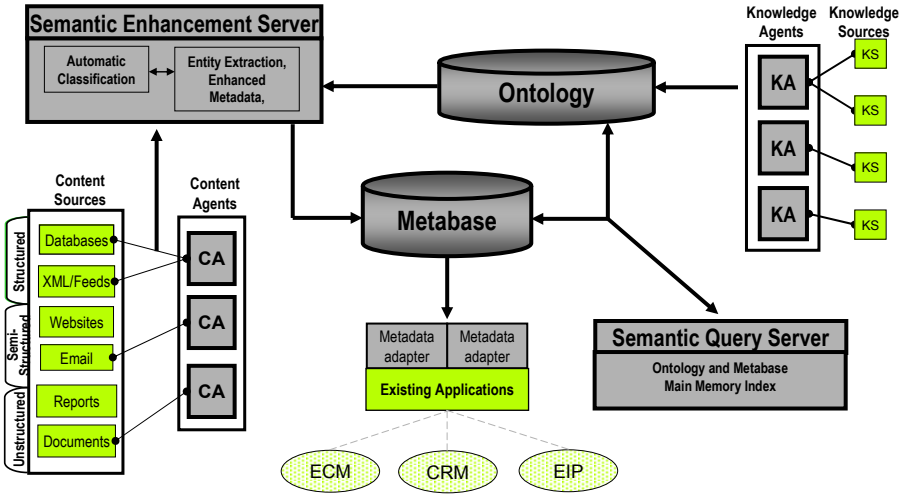
Metadata and Ontology: Primary Semantic Web enablers



SEMAGIX

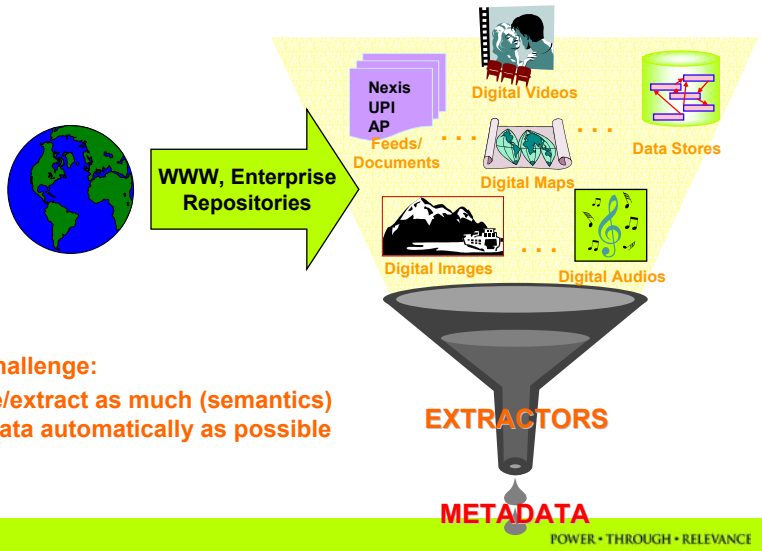
Semagix Freedom Architecture

(a platform for building ontology-driven information system)



SEMAGIX

Information Extraction and Metadata Creation

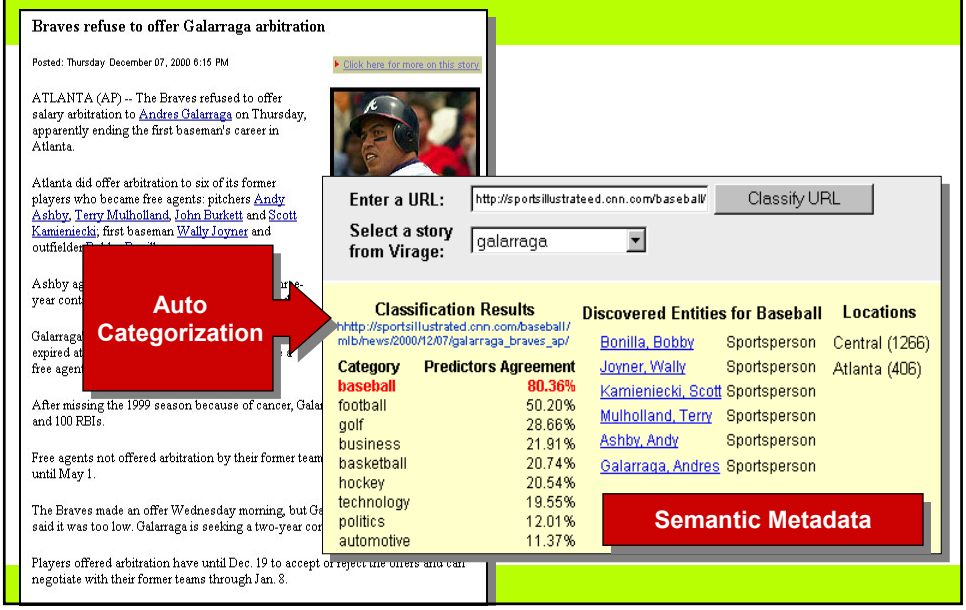


Key challenge:

Create/extract as much (semantics) metadata automatically as possible

Automatic Classification & Metadata Extraction (Web page)

SEMAGIX

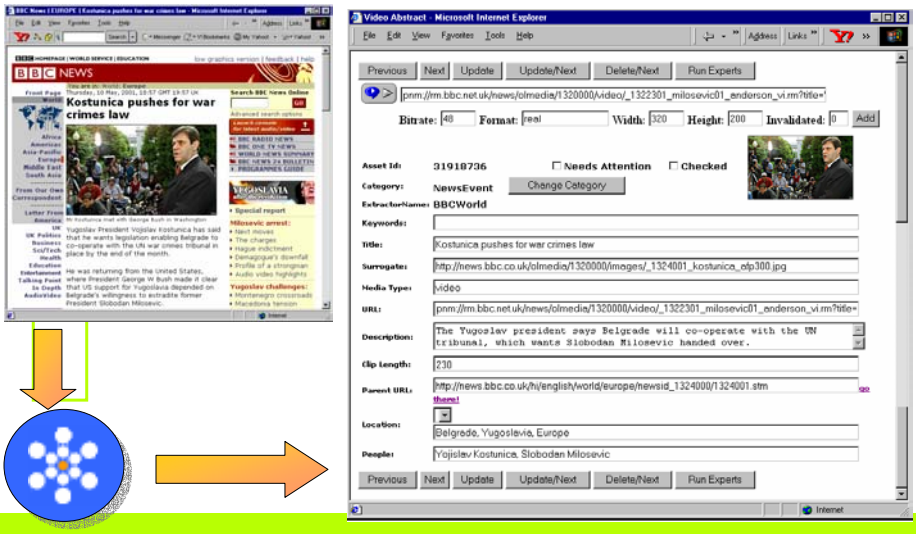


Ontology-directed Metadata Extraction (Semi-structured data)

SEMAGIX

Web Page

Enhanced Metadata Asset



Automatic Semantic Annotation of Text: Entity and Relationship Extraction

SEMAGIX

Blue-chip bonanza continues

company company company
Dow above 9,000 as HP, Home Depot lead advance; Microsoft upgrade helps techs.
date time
August 22, 2002: 11:44 AM EDT

phrase phrase
By Alexandra Twin, CNN/Money Staff Writer

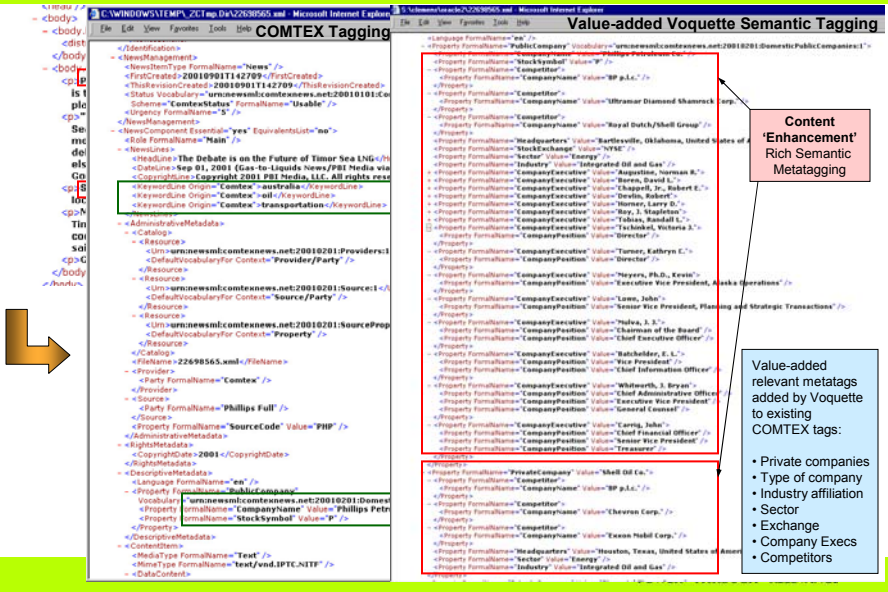
city company
New York (CNN/Money) - An upgrade of software leader Microsoft and strength in blue chips including
company company weekday
Hewlett-Packard and Home Depot were among the factors pushing stocks higher at midday Thursday,
financial index
with the Dow Jones industrial average spending time above the 9,000 level.
time financial index
Around 11:40 a.m. ET, the Dow Jones industrial average gained 65.06 to 9,022.09, continuing a more
date stock exchange
than 1,300-point resurgence since July 23. The Nasdaq composite gained 9.12 to 1,418.37.
financial index
The Standard & Poor's 500 index rose 9.61 to 958.97.
company stockSym $ $
Hewlett-Packard (HPQ: up $0.33 to $15.03, Research, Estimates) said a report shows its share of
the printer market grew in the second quarter, although another report showed that its share of the
continent region continent
computer server market declined in Europe, the Middle East and Africa.
company stockSym $ $
Home Depot (HD: up $1.07 to $33.75, Research, Estimates) was up for the third straight day after
topping fiscal second-quarter earnings estimates on Tuesday.
tech category company
Tech stocks managed a turnaround. Software continued to rise after Salomon Smith Barney upgraded
company stockSym $ $
No. 1 software maker Microsoft (MSFT: up $0.55 to $52.83, Research, Estimates) to "outperform"
$ $ company
from "neutral" and raised its price target to $59 from $56. Business software makers Oracle
stockSym $ $ company stockSym $ $
(ORCL: up $0.18 to $10.94, Research, Estimates), PeopleSoft (PSFT: up $1.17 to $20.67,
company stockSym $ $
Research, Estimates) and BEA Systems (BEAS: up $0.28 to $7.12, Research, Estimates)
all rose in tandem.

competes with

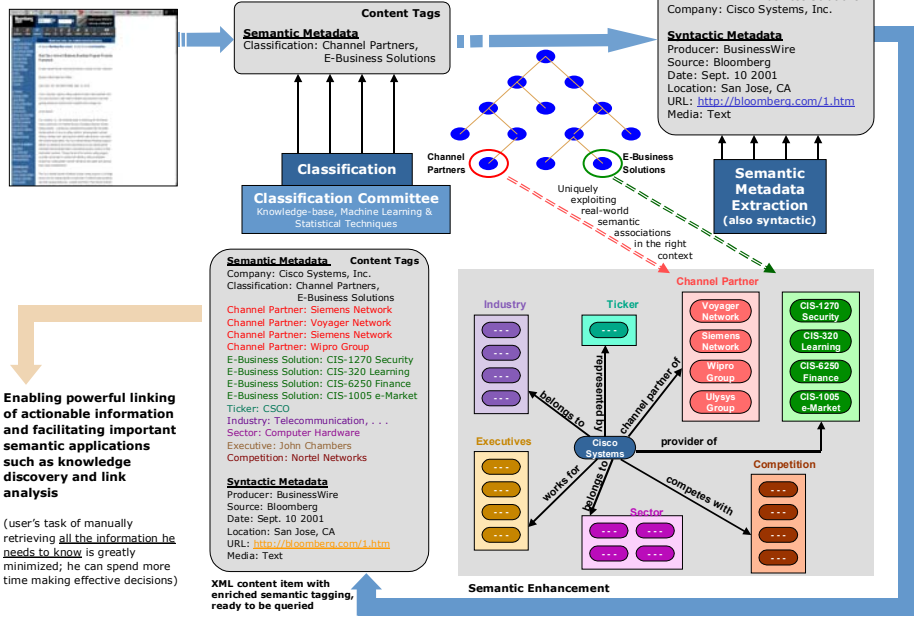
R • THROUGH • RELEVANCE

SEMAGIX

Automatic Semantic Annotation



Semantic Metadata Enhancement



SEMAGIX

The CIDOC CRM can be an excellent starting point for building the Semantic Web and ontology-driven information system for exchange, interoperability, integration of data/information and knowledge in the area of scientific and cultural heritage.

SEMAGIX

Types of Ontologies (or things close to ontology)

- Upper ontologies: modeling of time, space, process, etc
- Broad-based or general purpose ontology/nomenclatures: Cyc, CIRCA ontology (Applied Semantics), *WordNet*
- Domain-specific or Industry specific ontologies
  - News: politics, sports, business, entertainment
  - Financial Market
  - Terrorism
  - *(GO (a nomenclature), UMLS inspired ontology, …)*
- Application Specific and Task specific ontologies
  - Anti-money laundering
  - Equity Research

SEMAGIX

Practical Questions (for developing typical industry and application ontologies)

- Is there a typical ontology?
  - Three broad approaches:
    - social process/manual: many years, committees
    - automatic taxonomy generation (statistical clustering/NLP): limitation/problems on quality, dependence on corpus, naming
    - Descriptional component (schema) designed by domain experts; Assertional component (extension) by automated processes
- How do you develop ontology (methodology)?
- People (expertise), time, money
- Ontology maintenance

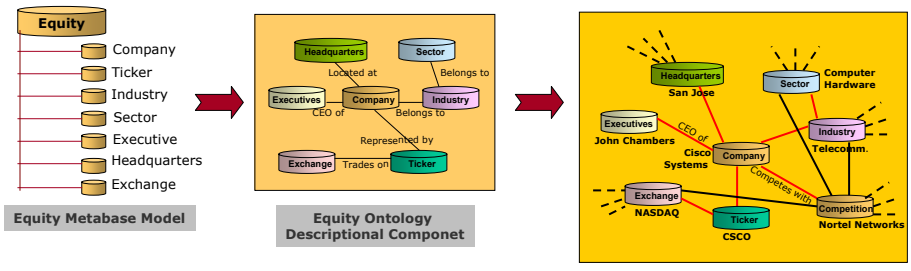
SEMAGIX

Practical Ontology Development Observation by Semagix

- Ontologies Semagix has designed:
  - Few classes to many tens (few hundreds) of classes and relationships (types); very small number of designers/knowledge experts; descriptional component (schema) designed with GUI
  - Hundreds of thousands to several millions entities and relationships (instances/assertions)
  - Tens of knowledge sources; populated by knowledge extractors
  - Primary scientific challenges faced: entity ambiguity resolution and data cleanup
  - Total effort: few person weeks

SEMAGIX

Ontology Example (Financial Equity domain)

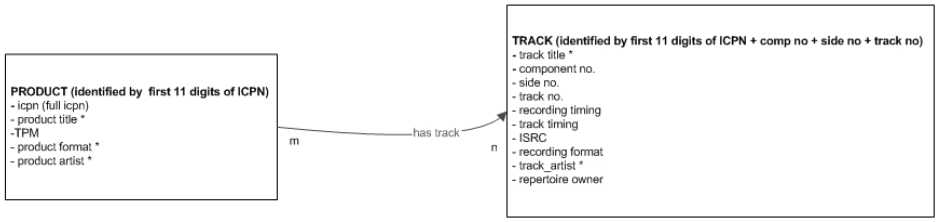


POWER • THROUGH • RELEVANCE

SEMAGIX

Ontology with simple schema

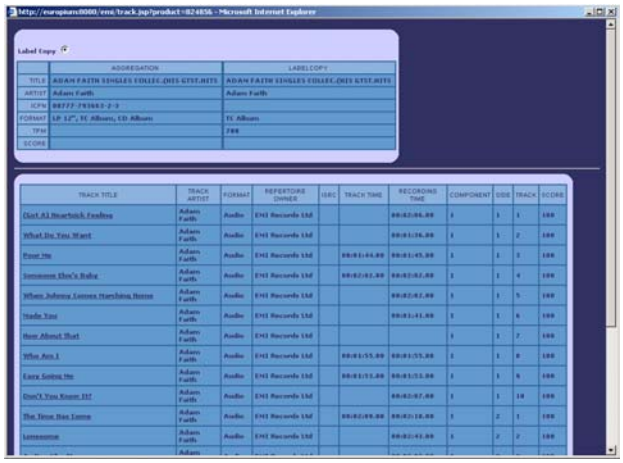
- **Ontology for a customer in Entertainment Industry**
- **Ontology Schema (Descriptional Component)**
  - Only 2 high-level entity classes: **Product** and **Track**
  - A few attributes for each entity class
  - Only 1 relationship between the 2 classes: "*has track*"
  - Many-to-many relationship between the two entity classes
    - A product can have multiple tracks
    - A track can belong to multiple products



SEMAGIX

Entertainment Ontology Schema (Assertional Component)

- About 400K entity instances in ontology
- About 3.8M attribute instances in ontology
- Entity instances and attribute instances extracted by Knowledge Agents from 5 disparate databases
- Databases contain little overlapping and mostly 'dirty' data (unfilled values, inconsistent data)



SEMAGIX

Technical Challenges Faced

- **Extremely 'dirty' data**
  - Inconsistent field values
  - Unfilled field values
  - Field values appearing to mean the same, but are different
- **Non-normalized Data**
  - Same field value referred to, in several different ways
- **Upper case vs. Lower case text analysis**
- **Modelling the ontology so that appropriate level (not too much, not too less) of information is modelled**
- **Optimizing the storage of the huge data**
  - How to load it into Freedom (currently distributed across 3 servers)
- **Scoring and pre-processing parameters changed frequently by customer, necessitating constant update of algorithm**
- **Efficiency measures**

SEMAGIX

Effort Involved

- **Ontology Schema Build-Out** (descriptional component)

  Essentially an iterative approach to refining the ontology schema based on periodic customer feedback

  - Very little technical effort (hours), but due to iterative decision making process with the multi-national customer, overall finalization of ontology took 3-4 weeks to complete

  Ontology Population (assertional component/knowledge base)

  - 5 Knowledge Agents, one for each database
  - Automated ontology population using Knowledge Agents took no longer than a day for all the Agents

POWER • THROUGH • RELEVANCE

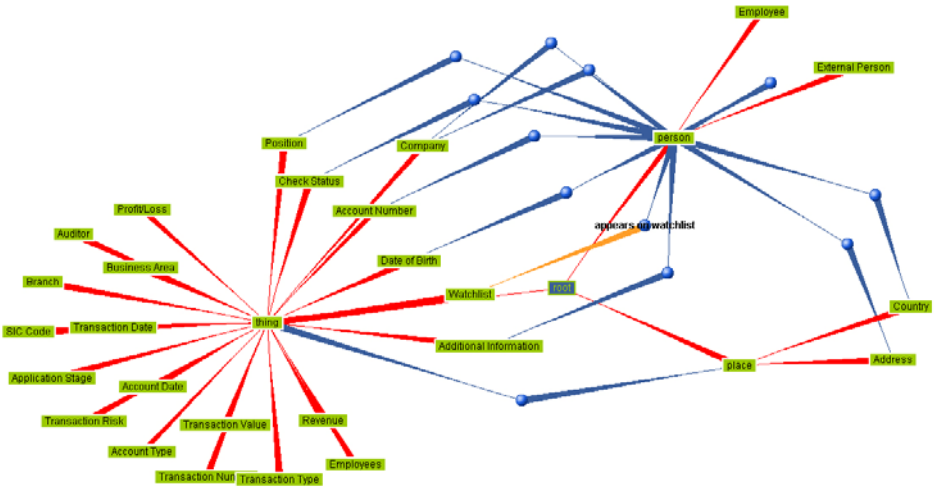
SEMAGIX

Example of Ontology with complex schema

- **Ontology for Anti-money Laundering (AML) application in Financial Industry**
- **Ontology Schema (Descriptional Component)**
  - About 40 entity classes
  - About 100 attribute types
  - About 50 relationship types between entity classes

SEMAGIX

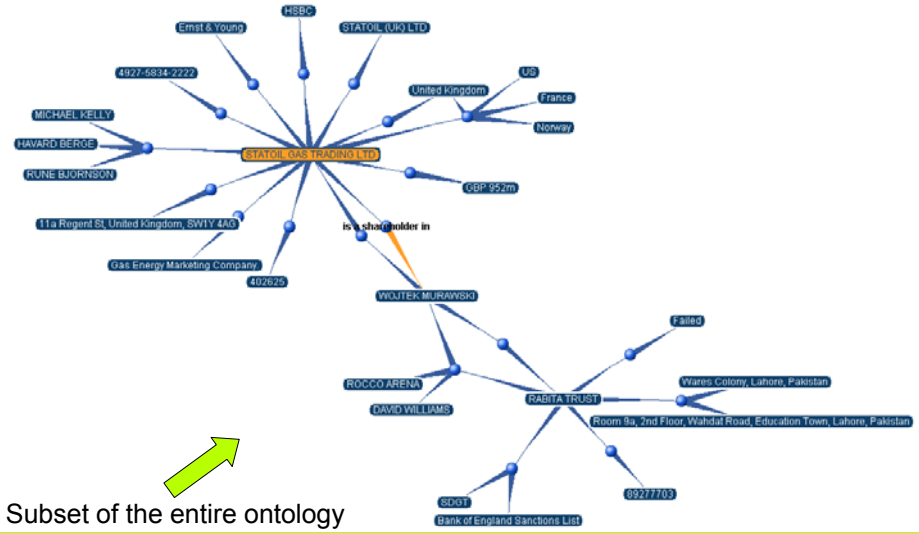
AML Ontology Schema (Descriptional Component)



POWER • THROUGH • RELEVANCE

SEMAGIX

AML Ontology Schema (Assertional Component)



Subset of the entire ontology

SEMAGIX

AML (Anti-Money Laundering) Ontology

Ontology Schema (Assertional Component)

- **About 1.5M entities, attributes and relationships**
- **4 different sources for knowledge extraction**
  - Dun and Bradstreet
  - Corporate 192
  - Companies House
  - Hoovers

Effort Involved

- **Ontology schema design: 3 days**
- **Automated Ontology population using Knowledge Agents: 2 days**

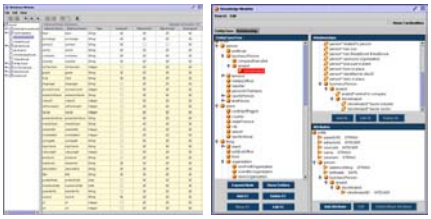
SEMAGIX

Technical Challenges Faced

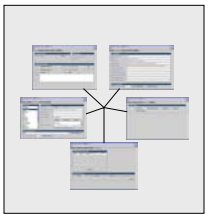
- Complex ambiguity resolution at entity extraction time
- Modelling the ontology so that appropriate level (not too much, not too less) of information is modelled
- Knowledge extraction from sources that needed extended cookie/HTTPS handling
- Programming ontology modelling through API
- Chalking out a balanced risk algorithm based on numerous parameters involved

SEMAGIX

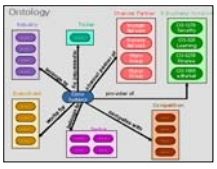
Ontology Creation and Maintenance Steps



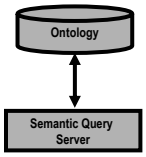
1. Ontology Model Creation



2. Knowledge Agent Creation



3. Automatic aggregation of Knowledge



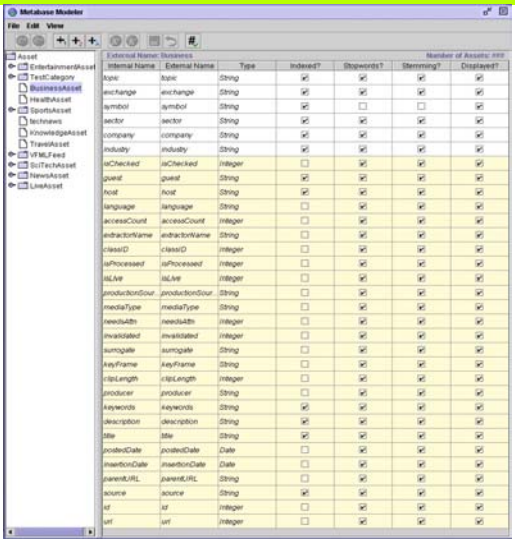
4. Querying the Ontology

POWER • THROUGH • RELEVANCE

SEMAGIX

Step 1: Ontology Model Creation

Create an Ontology Model using Semagix Freedom Toolkit GUIs

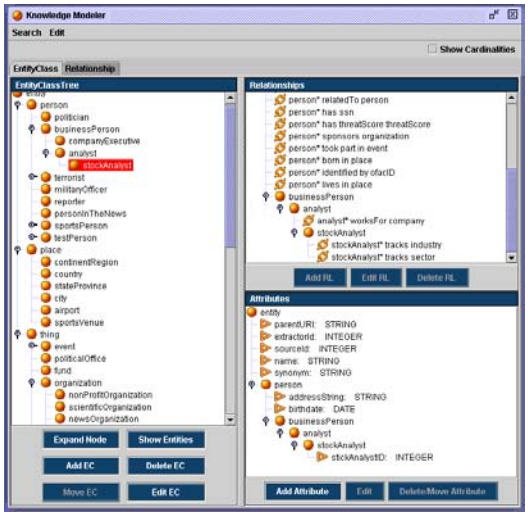


- This corresponds to the descriptioinal part (schema) of the Ontology
- Manually define Ontology structure (entity classes, relationship types, domain-specific and domain independent attributes)
- Configure parameters for attributes pertaining to indexing, lexical analysis, interface, etc.
- Existing industry-specific taxonomies like MESH (Medical), etc. can be reused or imported into the Ontology

SEMAGIX

Step 1: Ontology Model Creation

Create an Ontology Model using Semagix Freedom Toolkit GUIs (Cont.)



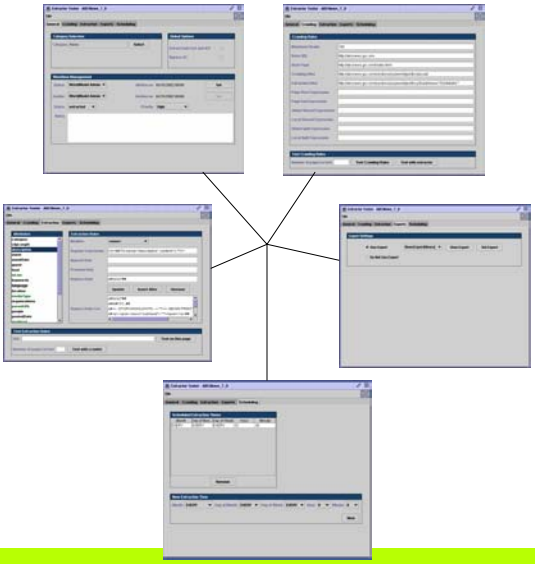
- This corresponds to the schema of the definitional part of the Ontology
- Manually define Ontology structure for knowledge (in terms of entities, entity attributes and relationships)
- Create entity class, organize them (e.g., in taxonomy)
  e.g. **Person**
  └ **BusinessPerson**
  └ **Analyst**
  └ **StockAnalyst . . .**
- Establish any number of meaningful (named) relationships between entity classes
  e.g. **Analyst works for Company**
  StockAnalyst tracks Sector
  BusinessPerson own shares in Company . . .
- Set any number of attributes for entity classes
  e.g. **Person**
  └ **Address <text>**
  └ **Birthdate <date>**
  StockAnalyst
  └ **StockAnalystID <integer>**

POWER • THROUGH • RELEVANCE

SEMAGIX

Step 2: Knowledge Agent Creation

Create and configure Knowledge Agents to populate the Ontology

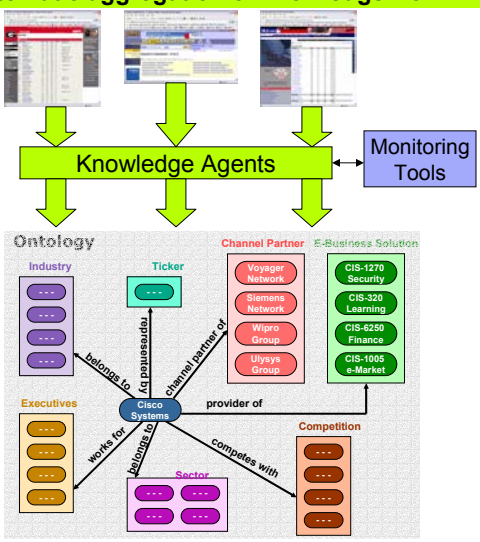


- Identify any number of trusted knowledge sources relevant to customer's domain from which to extract knowledge
  - Sources can be internal, external, secure/proprietary, public source, etc.
- Manually configure (one-time) the Knowledge Agent for a source by configuring
  - which relevant sections to crawl to
  - what knowledge to extract
  - what pre-defined intervals to extract knowledge at
- Knowledge Agent automatically) runs at the configured time-intervals and extracts entities and relationships from the source, to keep the Ontology up-to-date

SEMAGIX

Step 3: Automatic aggregation of knowledge

Automatic aggregation of knowledge from knowledge sources



- Automatic aggregation of knowledge at pre-defined intervals fo time
- Supplemented by easy-to-use monitoring tools
- Knowledge Agents extract and organize relevant knowledge into the Ontology, based on the Ontology Model
  - Tools for disambiguation and cleaning
- The Ontology is constantly growing and kept up-to-date

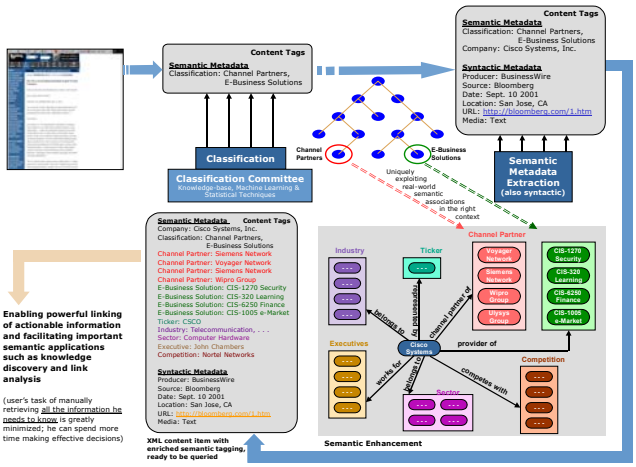
SEMAGIX

Semantic Enhancement Server

Semantic Enhancement Server: Semantic Enhancement Server classifies content into the appropriate topic/category (if not already pre-classified), and subsequently performs entity extraction and content enhancement with semantic metadata from the Semagix Freedom Ontology

How does it work?

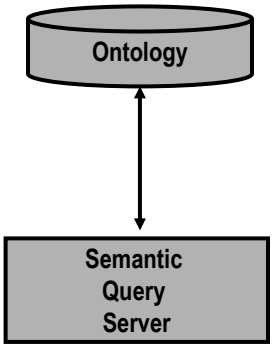
- Uses a hybrid of statistical, machine learning and knowledge-base techniques for classification
- Not only classifies, but also enhances semantic metadata with associated domain knowledge



SEMAGIX

Step 4: Querying the Ontology

Semantic Query Server can now query the Ontology



- Semantic Query Server can now perform in-memory complex querying on the Ontology and Metadata
  - Incremental indexing
  - Distributed indexing
  - High performance: 10M queries/hr; less than 10ms for typical search queries
  - 2 orders of magnitude faster than RDBMS for complex analytical queries
- Knowledge APIs provide a Java, JSP or an HTTP-based interface for querying the Ontology and Metadata

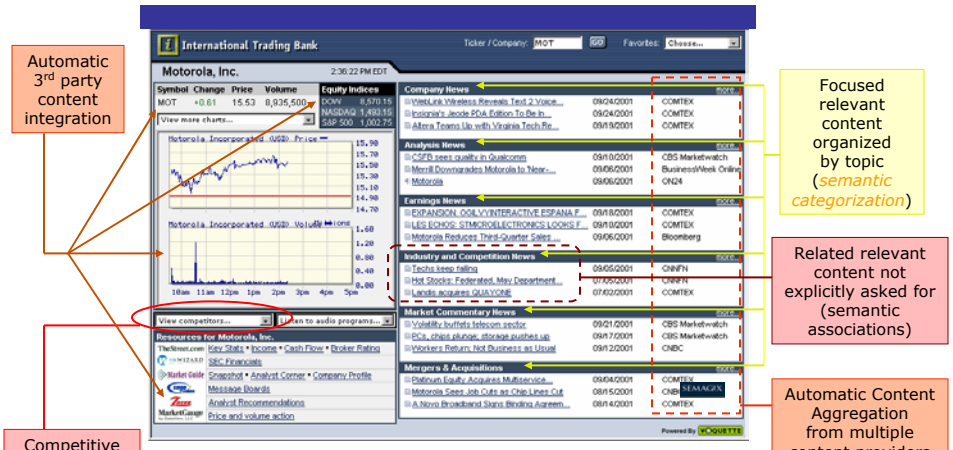
SEMAGIX

Ontology-based Semagix solutions

- **Equity Analysis Workbench**
  - Heterogeneous internal and extenral, push and pull content
  - Automatic Classification , Semantic Information Correlation, Semantic (domain-specific search)
- **CIRAS - Anti Money Laundering**:
  - **Business issue:** Optimisation of complex analysis from multiple sources
  - **Technology:** Integration of process specific business insight from structured and unstructured information sources
- **APITAS – Passenger threat assessment**
  - **Business issue :** Rapid identification of high risk scenarios from vast amounts of information
  - **Technology:** Managed high volume of information, speed of main memory indexed queries

SEMAGIX

Semantic Application Example – Analyst Workbench



SEMAGIX

CIRAS - Anti Money Laundering

(Know Your Customer – KYC)

POWER • THROUGH • RELEVANCE

SEMAGIX

Fundamental Issues – Current Processes

Existing service bureau offerings created for different purpose – credit scoring

- Majority of content supplied not applicable to KYC – **unnecessary cost**
- Rigid and static information require user interpretation – **elongation of process time**
- Not specific enough to comply with new legislation – **non-compliance**

Multiple manual checks against a variety of sources

- Difficulty to link different pieces of information – **reduced effectiveness**
- Checks are sequential and resource intensive - **Increase process time and cost**
- Duplication of content – **increased subscription cost**

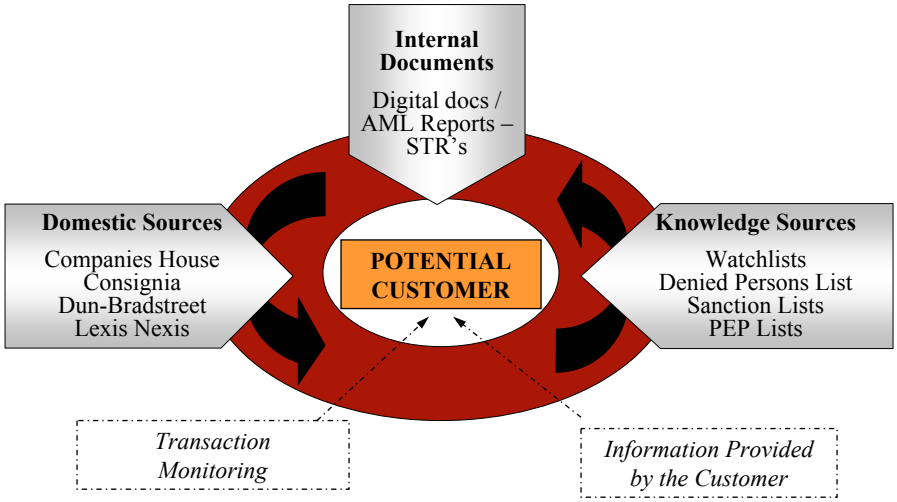
Inability to implement domain-specific 'best practises'

- Process knowledge resides with analysts – **variable quality of output**
- Difficulty to fine-tune processes to specific domain **– inflexible process**

Current processes are resource and time inefficient leading to inflexible and costly compliance

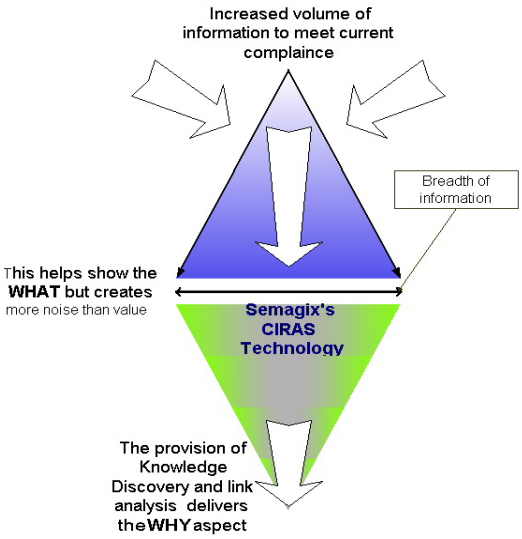
SEMAGIX

Constituent parts of 'reasonable grounds'



SEMAGIX

What vs. Why



SEMAGIX

What are the benefits

1. **Control** – compliance officers dictate the scale and scope of the checks made without incremental costs
2. **Protects integrity of the company** – reputation and confidence are maintained through effective systems and controls
    - Comply with new legislations and regulations **-** proceeds of crime act 2002 part 7, USA PATRIOT act
3. **Cost**
    - Lower total cost for compliance with current and future legislation
    - Lower content subscription and HR costs
4. **Increased quality and efficiency** of the compliance process
5. **Integration into existing processes** – open standards enables the technology to be integrated into current KYC processes
6. **Interoperability** – provides integration across disparate legacy systems facilitating 'retrospective reviews' of customer bases

POWER • THROUGH • RELEVANCE

SEMAGIX

CIRAS's Components

Customer Application Information:

Integration of structured information gathered during the account opening process

Risk Weighting

Relevant Knowledge

Relevant Content

Anti-Money Laundering Ontology

POWER • THROUGH • RELEVANCE

SEMAGIX

Semagix's Approach to KYC

This is achieved through:

1. Risk weighting based on the underlying information and pre-defined criteria
    - Watchlist check
    - Link Analysis
    - ID Verification
2. Verification of the identity of a customer's name and address against domestic knowledge and content sources, includes:
    - What is already known about the customer
    - 3rd Party integration if required
    - Details of content relevant to 'knowing the customer'

SEMAGIX

Actionable Information

STATOIL GAS TRADING LTD - Details

Score Component	Score	Reason
shareholder check	65	has a shareholder WOJTEK MURAWSKI who works for RABITA TRUST which appears on Bank of England Sanctions List
shareholder check	65	has a shareholder WOJTEK MURAWSKI who works for RABITA TRUST which appears on SDGT
Aggregate Score: 65		

Aggregated risk represented by a customer

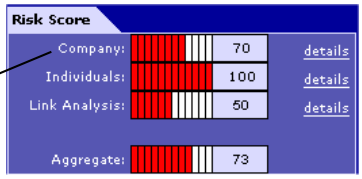
Summary of Capabilities

- Risk based approach to identification and verification
- Checks conducted against a wide variety of knowledge sources
- Integrates with existing processes
- Tailored for on-going and future requirements

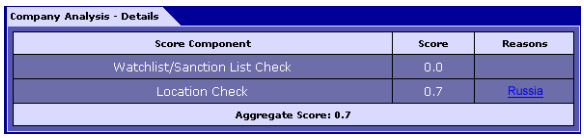
POWER • THROUGH • RELEVANCE

SEMAGIX

CIRAS's Components



1. Company Analysis



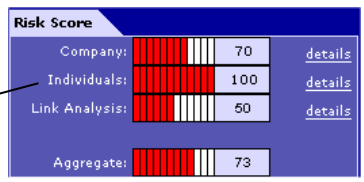
Company Analysis - Details

Score Component	Score	Reasons
Watchlist/Sanction List Check	0.0	
Location Check	0.7	Russia
Aggregate Score: 0.7		

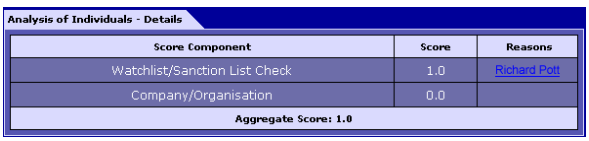
- Cross references international and domestic watchlists
- Tailored to the operational environment
- Scheduled (every day) updates of the changes to lists

SEMAGIX

CIRAS's Components



2. ID Verification



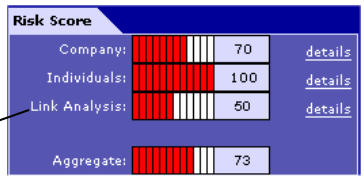
Analysis of Individuals - Details

Score Component	Score	Reasons
Watchlist/Sanction List Check	1.0	Richard Pott
Company/Organisation	0.0	
Aggregate Score: 1.0		

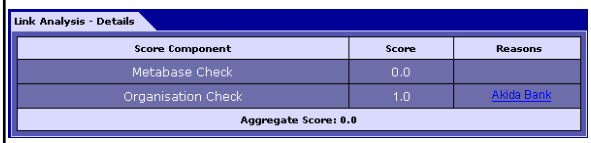
- Provides an indication as to the risk posed by individuals associated with the company
- Allows navigation into possible causes of 'false positive's

SEMAGIX

CIRAS's Components



3. Link Analysis Check



- Identification and verification of relationships customer holds with other entities (organisations, people etc)
- Flags high-risk transaction flows
- References internal reports held

SEMAGIX

CIRAS's Components

Provision of 'knowledge' already held about a prospect and provides the ability to navigate through each 'instance' to verify information

1. Normalisation of information to understand multiple formats of an identity
2. Key Employees
3. Company Details
4. Associated Companies

SEMAGIX

CIRAS's Components

External content, from multiple sources, in any format relevant to 'knowing the customer'

Internal content, previous KYC checks undertaken, STR reports filed and transaction monitoring alerts relevant to the customer in question

SEMAGIX

Current applications of the technology

- **CIRAS - Anti Money Laundering**
- **Passenger Threat Assessment System**

External demo page

About Semagix

Semagix, through a patented *semantic* approach to Enterprise Information Integration (EII), allows enterprises to integrate and extract insights from their structured and unstructured information assets in order to conceive and develop smarter business processes and applications

SEMAGIX

POWER • THROUGH • RELEVANCE