

SEMAGIX
POWER • THROUGH • RELEVANCE

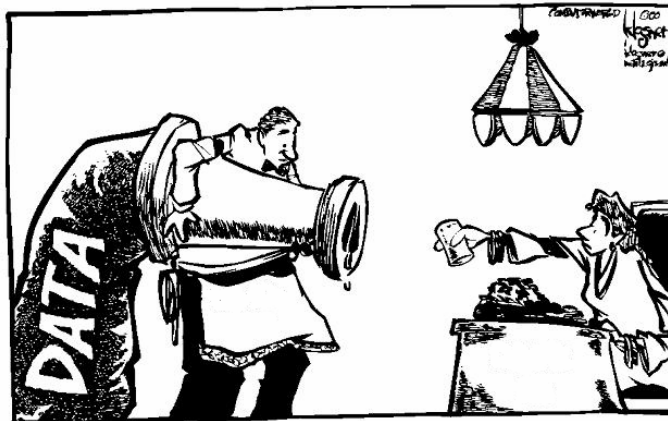
**Capturing and Applying Existing Knowledge to Semantic Applications
or *Ontology-driven Information Systems in Action***

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

SEMAGIX

Syntax -> Semantics



POWER • THROUGH • RELEVANCE

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

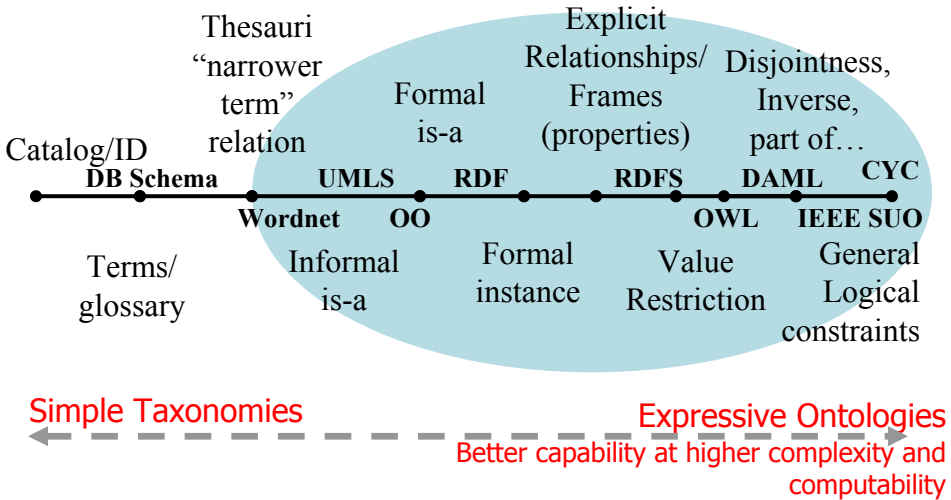
Achieved in the context of successful technology transfer from academic research (LSDIS lab, UGA's SCORE technology) into commercial product (Semagix's Freedom)

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

Ontology provides underpinning for semantic techniques in information systems.

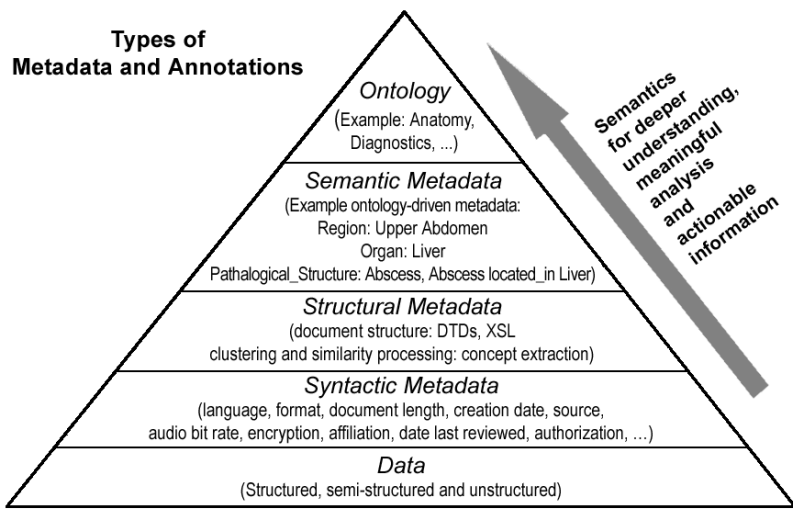
- ◆ A model/representation of the real world (relevant concepts, entities, attributes, relationships, domain vocabulary and factual knowledge, all connected via a semantic network). Basic of agreement, applying knowledge
- ◆ Enabler for improved information systems functionalities and the Semantic Web:
 - ◆ 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”](#)

Increasingly More Semantic Representation



After McGuinness & Finin POWER • THROUGH • RELEVANCE

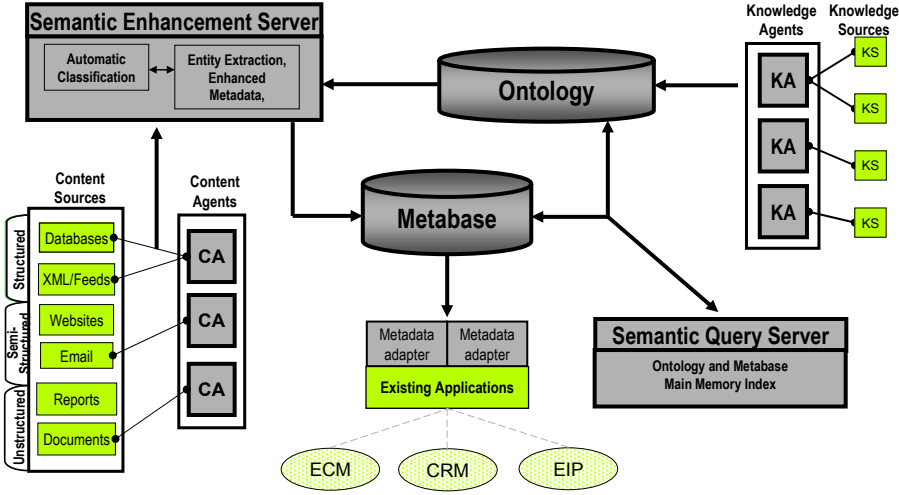
Metadata and Ontology: Primary Semantic Web enablers



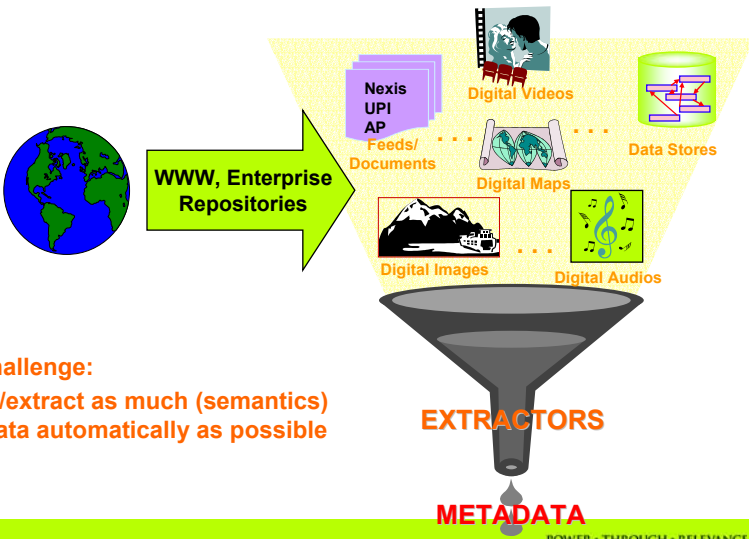
POWER • THROUGH • RELEVANCE

Semagix Freedom Architecture

(a platform for building ontology-driven information system)



Information Extraction and Metadata Creation



Key challenge:
Create/extract as much (semantics) metadata automatically as possible

Automatic Classification & Metadata Extraction (Web page)

SEMAGIX

Braves refuse to offer Galarraga arbitration

Posted: Thursday December 07, 2000 6:16 PM

ATLANTA (AP) -- The Braves refused to offer salary arbitration to [Andres Galarraga](#) on Thursday, apparently ending the first baseman's career in Atlanta.

[Click here for more on this story](#)



Atlanta did offer arbitration to six of its former players who became free agents: pitchers [Andy Ashby](#), [Terry Mulholland](#), [John Burkett](#) and [Scott Kamieniecki](#); first baseman [Wally Joyner](#); and outfielder [Pete Burchett](#).

Ashby signed a year contract with the Braves.

Galarraga expired at free agent.

After missing the 1999 season because of cancer, Galarraga had 100 RBIs.

Free agents not offered arbitration by their former team until May 1.

The Braves made an offer Wednesday morning, but Galarraga said it was too low. Galarraga is seeking a two-year contract.

Players offered arbitration have until Dec. 19 to accept or reject the offers and can negotiate with their former teams through Jan. 8.

Auto Categorization

Enter a URL:

Select a story from Virage:

Classification Results		Discovered Entities for Baseball	Locations
Category	Predictors Agreement		
baseball	80.36%	Bonilla, Bobby Sportsperson	Central (1266)
football	50.20%	Jowner, Wally Sportsperson	Atlanta (406)
golf	28.66%	Kamieniecki, Scott Sportsperson	
business	21.91%	Mulholland, Terry Sportsperson	
basketball	20.74%	Ashby, Andy Sportsperson	
hockey	20.54%	Galarraga, Andres Sportsperson	
technology	19.55%		
politics	12.01%		
automotive	11.37%		

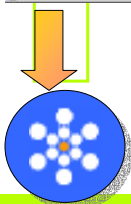
Semantic Metadata

Ontology-directed Metadata Extraction (Semi-structured data)

SEMAGIX

Web Page

Enhanced Metadata Asset



Video Abstract - Microsoft Internet Explorer

Previous Next Update Update/Next Delete/Next Run Experts

Bitrate: 148 Format: real Width: 320 Height: 200 Invalidated: 0 Add

Asset Id: 31918736 Needs Attention Checked

Category: NewsEvent

ExtractorName: BBCWorld

Keywords:

Title: Kostunica pushes for war crimes law

Surrogates: http://news.bbc.co.uk/1/mediastore/1/32000/1324001_kostunica_otp300.jpg

Media Types: video

URL: http://news.bbc.co.uk/1/mediastore/1/32000/1324001_kostunica_otp300.jpg

Description: The Yugoslav president says Belgrade will co-operate with the UN tribunal, which wants Slobodan Milosevic handed over.

Clip Length: 2:30

Parent URL: http://news.bbc.co.uk/1/english/world/europe/newsid_1324001/1324001.stm

Locations: Belgrade, Yugoslavia, Europe

People: Yujoslav Kostunica, Slobodan Milosevic

Previous Next Update Update/Next Delete/Next Run Experts

Automatic Semantic Annotation of Text: Entity and Relationship Extraction

SEMAGIX

Blue-chip bonanza continues

Dow above 9,000 as HP, Home Depot lead advance; Microsoft upgrade helps techs.

August 22, 2002: 11:44 AM EDT

By Alexandra Twin, CNN/Money Staff Writer

New York (CNN/Money) - An upgrade of software leader Microsoft and strength in blue chips including Hewlett-Packard and Home Depot were among the factors pushing stocks higher at midday Thursday, with the Dow Jones industrial average spending time above the 9,000 level.

Around 11:40 a.m. ET, the Dow Jones industrial average gained 65.06 to 9,022.09, continuing a more than 1,300-point resurgence since July 23. The Nasdaq composite gained 9.12 to 1,418.37.

The Standard & Poor's 500 index rose 9.61 to 958.97.

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 computer server market declined in Europe, the Middle East and Africa.

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 stocks managed a turnaround. Software continued to rise after Salomon Smith Barney upgraded No. 1 software maker Microsoft (MSFT: up \$0.55 to \$52.83, Research, Estimates) to "outperform" from "neutral" and raised its price target to \$59 from \$56. Business software makers Oracle (ORCL: up \$0.18 to \$10.94, Research, Estimates), PeopleSoft (PSFT: up \$1.17 to \$20.67, Research, Estimates) and BEA Systems (BEAS: up \$0.28 to \$7.12, Research, Estimates) all rose in tandem.

R • THROUGH • RELEVANCE

Automatic Semantic Annotation

SEMAGIX

COMTEX Tagging

```

<?xml version="1.0" encoding="UTF-8" ?>
<document type="News" id="200109011142709" firstCreated="200109011142709"
  status="Vocabulary" urn="sem:comctxnews.net:20010101:ComctxNews"
  xmlns="http://www.comctxnews.net/20010101/ComctxNews"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.comctxnews.net/20010101/ComctxNews
    http://www.comctxnews.net/20010101/ComctxNews.xsd" />
<newsManagement>
  <newComponentEssentials>yes</newComponentEssentials>
  <role>Main</role>
  <headline>The Debate is on the Future of Timor Sea LNG</headline>
  <subline>Rep. Bill Leno (D-Calif.) says he supports the bill</subline>
  <keywords>
    <keyword>Context</keyword>
    <keyword>Context</keyword>
    <keyword>Context</keyword>
  </keywords>
  <administrativeMetadata>
    <catalog>
      <urn:sem:comctxnews.net:20010201:ProviderParty />
    </catalog>
    <resource>
      <urn:sem:comctxnews.net:20010201:SourceParty />
    </resource>
    <resource>
      <urn:sem:comctxnews.net:20010201:SourceProperty />
    </resource>
    <catalog>
      <id>22498565.xml</id>
      <partyFormName>Phillips Full />
    </catalog>
    <source>
      <partyFormName>Phillips Full />
    </source>
    <administrativeMetadata>
      <propertyFormName>SourceCode Value="PHIP" />
    </administrativeMetadata>
    <rightsMetadata>
      <copyrightDate>2001</copyrightDate>
    </rightsMetadata>
    <descriptionMetadata>
      <languageFormName>en />
      <propertyFormName>PublicCompany />
      <vocabulary urn="sem:comctxnews.net:20010201:DomesticPublicCompanies" />
      <propertyFormName>CompanyName Value="Phillips Petroleum" />
      <propertyFormName>StockSymbol Value="P" />
    </descriptionMetadata>
    <content>
      <text />
    </content>
  </newsManagement>
  <dataContent />
</document>

```

Value-added Voquette Semantic Tagging

```

<Property FormName="PublicCompany" Value="sem:comctxnews.net:20010201:DomesticPublicCompanies" />
<Property FormName="StockSymbol" Value="P" />
<Property FormName="CompanyName" Value="BP a.s." />
<Property FormName="Competitor" />
<Property FormName="Competitor" Value="Ultranor Diamond Shamrock" />
<Property FormName="Competitor" Value="Royal Dutch/Shell Group" />
<Property FormName="Headquarters" Value="Bartleville, Oklahoma, United States of America" />
<Property FormName="Sector" Value="Energy" />
<Property FormName="Industry" Value="Integrated Oil and Gas" />
<Property FormName="CompanyExecutive" Value="Angell, Norman B." />
<Property FormName="CompanyExecutive" Value="Breen, David C." />
<Property FormName="CompanyExecutive" Value="Chappell, Jr., Robert E." />
<Property FormName="CompanyExecutive" Value="Dreier, Robert" />
<Property FormName="CompanyExecutive" Value="Homer, Larry D." />
<Property FormName="CompanyExecutive" Value="Ray, J. Magdalen" />
<Property FormName="CompanyExecutive" Value="Tobias, Randall L." />
<Property FormName="CompanyExecutive" Value="Tschobol, Victoria A." />
<Property FormName="CompanyExecutive" Value="Director" />
<Property FormName="CompanyExecutive" Value="Turner, Kathryn E." />
<Property FormName="CompanyExecutive" Value="Director" />
<Property FormName="CompanyExecutive" Value="Neyers, Ph.D., Kevin" />
<Property FormName="CompanyExecutive" Value="Executive Vice President, Alaska Operations" />
<Property FormName="CompanyExecutive" Value="Lowe, John" />
<Property FormName="CompanyExecutive" Value="Senior Vice President, Planning and Strategic Transactions" />
<Property FormName="CompanyExecutive" Value="Nava, J. A." />
<Property FormName="CompanyExecutive" Value="Chairman of the Board" />
<Property FormName="CompanyExecutive" Value="Chief Executive Officer" />
<Property FormName="CompanyExecutive" Value="Batholde, E. L." />
<Property FormName="CompanyExecutive" Value="Vice President" />
<Property FormName="CompanyExecutive" Value="Chief Information Officer" />
<Property FormName="CompanyExecutive" Value="Whitworth, J. Bryan" />
<Property FormName="CompanyExecutive" Value="Chief Administrative Officer" />
<Property FormName="CompanyExecutive" Value="Executive Vice President" />
<Property FormName="CompanyExecutive" Value="General Counsel" />
<Property FormName="CompanyExecutive" Value="Carrig, John" />
<Property FormName="CompanyExecutive" Value="Chief Financial Officer" />
<Property FormName="CompanyExecutive" Value="Senior Vice President" />
<Property FormName="CompanyExecutive" Value="Treasurer" />
</pre>

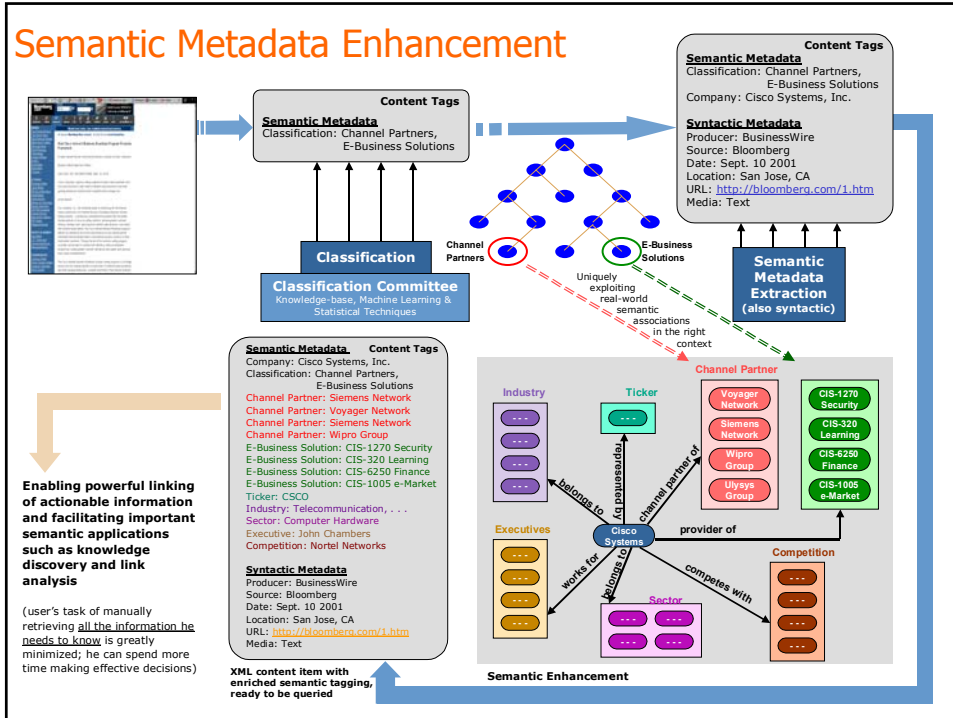
```

Content 'Enhancement' Rich Semantic Metatagging

Value-added relevant metatags added by Voquette to existing COMTEX tags:

- Private companies
- Type of company
- Industry affiliation
- Sector
- Exchange
- Company Execs
- Competitors

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.

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

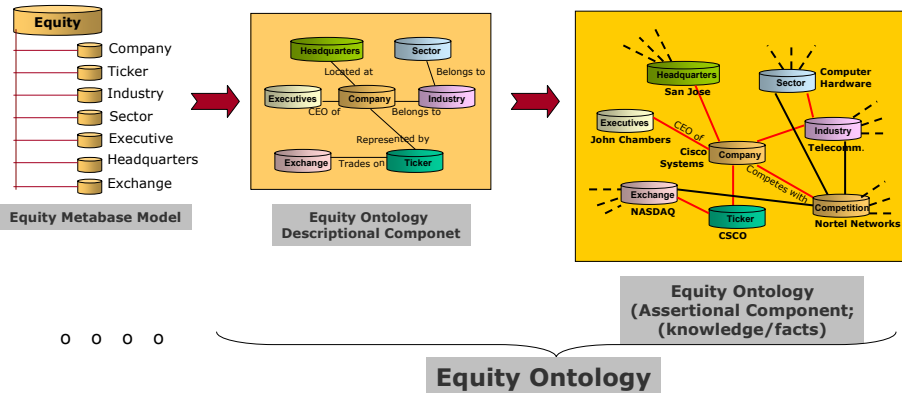
Practical Questions (for developing typical industry and application ontologies)

- ◆ 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
 - ◆ Descriptive component (schema) designed by domain experts; Assertional component (extension) by automated processes
- ◆ How do you develop ontology (methodology)?
- ◆ People (expertise), time, money
- ◆ Ontology maintenance

Practical Ontology Development Observation by Semagix

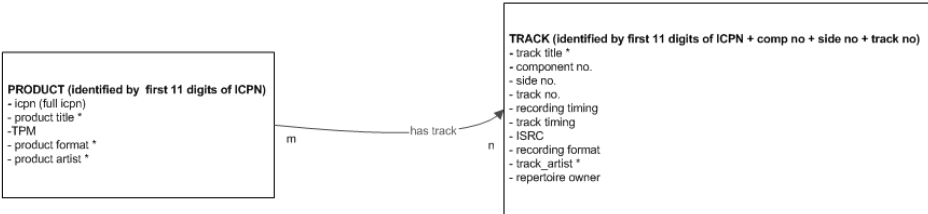
- ◆ Ontologies Semagix has designed:
 - ◆ Few classes to many tens (few hundreds) of classes and relationships (types); very small number of designers/knowledge experts; descriptive component (schema) designed with GUI
 - ◆ Hundreds of thousands to several millions entities and relationships (instances/assertions)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

Ontology Example (Financial Equity domain)



Ontology with simple schema

- ◆ **Ontology for a customer in Entertainment Industry**
- ◆ **Ontology Schema (Descriptive 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



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)

TRACK TITLE	TRACK ARTIST	FORMATT	REPORTING COUNTRY	ISRC	TRACK TIME	RECORDING TIME	COMPONENT	SIDE	TRACK	SCORE
What You Want	Adam Faith	Audio	EMI Records Ltd		00:02:06.00		1	1	1	100
I'm A Man	Adam Faith	Audio	EMI Records Ltd	00:01:04.00	00:01:04.00		1	1	1	100
Shine Bright Like A Star	Adam Faith	Audio	EMI Records Ltd	00:02:02.00	00:02:02.00		1	1	1	100
When Johnny Comes Marching Home	Adam Faith	Audio	EMI Records Ltd		00:02:07.00		1	1	1	100
Trade Winds	Adam Faith	Audio	EMI Records Ltd		00:02:01.00		1	1	1	100
How About That	Adam Faith	Audio	EMI Records Ltd				1	1	1	100
Who Am I	Adam Faith	Audio	EMI Records Ltd	00:01:05.00	00:01:05.00		1	1	1	100
Love, Love Me Do	Adam Faith	Audio	EMI Records Ltd	00:01:11.00	00:01:11.00		1	1	1	100
Don't You Know Just	Adam Faith	Audio	EMI Records Ltd		00:02:07.00		1	1	1	100
Hey, Hey, Hey, Hey	Adam Faith	Audio	EMI Records Ltd	00:02:06.00	00:02:06.00		1	2	1	100
Let's Dance	Adam Faith	Audio	EMI Records Ltd		00:02:03.00		1	2	1	100

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**

Effort Involved

- ◆ **Ontology Schema Build-Out** (descriptive component)

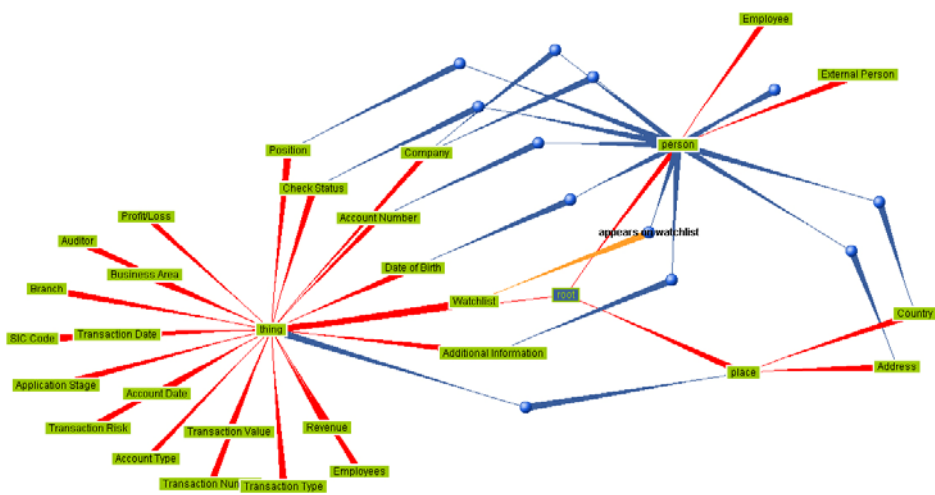
Essentially an iterative approach to refining the ontology schema based on periodic customer feedback

 - ◆ 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

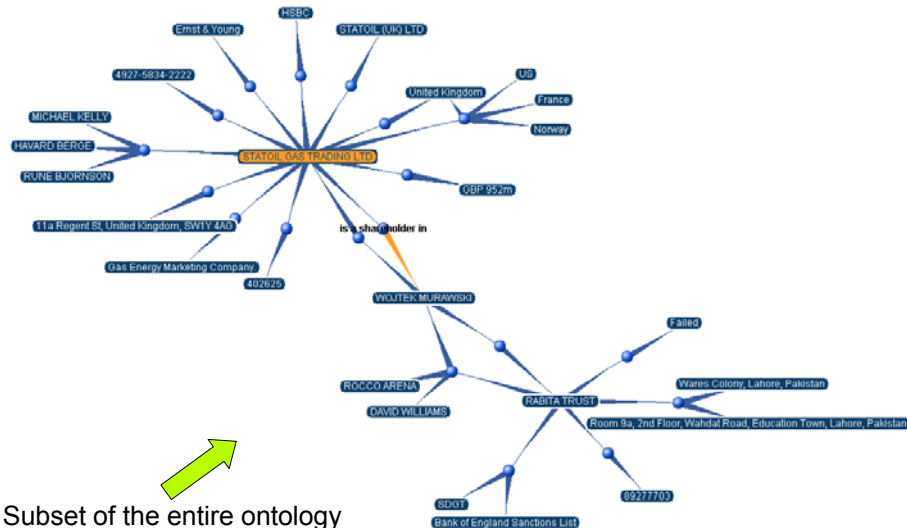
Example of Ontology with complex schema

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

AML Ontology Schema (Descriptive Component)



AML Ontology Schema (Assertional Component)



Subset of the entire ontology

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

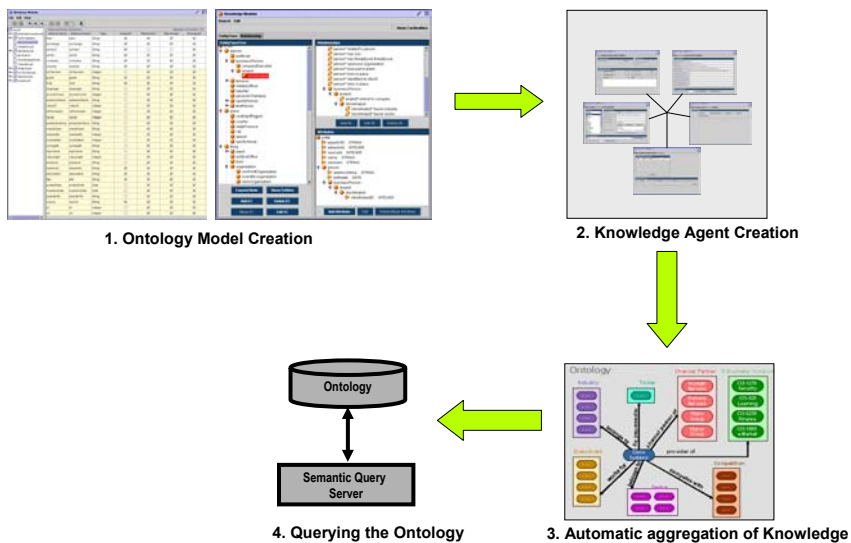
Effort Involved

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

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

Ontology Creation and Maintenance Steps



Step 1: Ontology Model Creation

Create an Ontology Model using Semagix Freedom Toolkit GUIs

Asset	Internal Name	External Name	Type	Indexed?	Stopwords?	Stemming?	Disputed?	Number of Assets (min)
Asset								
TextCategory	topic	topic	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
BusinessAsset	exchange	exchange	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
HealthAsset	symbol	symbol	String	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
SportsAsset	sector	sector	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
StockNews	company	company	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
KnowledgeAsset	industry	industry	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
TravelAsset	checked	checked	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
WikipediaAsset	guest	guest	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
NewsAsset	host	host	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
LinkAsset	language	language	String	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	accessCount	accessCount	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	extractorName	extractorName	String	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	classID	classID	Integer	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	isProcessed	isProcessed	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	id	id	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	productID	productID	String	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	mediaType	mediaType	String	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	needsID	needsID	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	isVidated	isVidated	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	sumgate	sumgate	String	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	keyFrame	keyFrame	String	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	urlLength	urlLength	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	producer	producer	String	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	keywords	keywords	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	description	description	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	title	title	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	postedDate	postedDate	Date	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	insertionDate	insertionDate	Date	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	parentURL	parentURL	String	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	source	source	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	id	id	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
	url	url	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	

- This corresponds to the descriptive 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

Step 1: Ontology Model Creation

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

EntityClass Tree

- person
 - politician
 - businessPerson
 - companyExecutive
 - analyst
 - StockAnalyst
 - seniorist
 - militaryOfficer
 - reporter
 - personInTheNews
 - sportsPerson
 - basePerson
- place
 - continentRegion
 - country
 - stateProvince
 - city
 - airport
 - sportsVenue
- event
 - politicalOffice
 - fund
 - organization
 - nonProfitOrganization
 - scientificOrganization
 - newsOrganization

Relationships

- person* relatedTo person
- person* has a/an
- person* has threatScore threatScore
- person* sponsors organization
- person* took part in event
- person* born in place
- person* identified by ofacID
- person* lives in place
- businessPerson
 - analyst
 - analyst* works for company
 - stockAnalyst
 - stockAnalyst* tracks industry
 - stockAnalyst* tracks sector

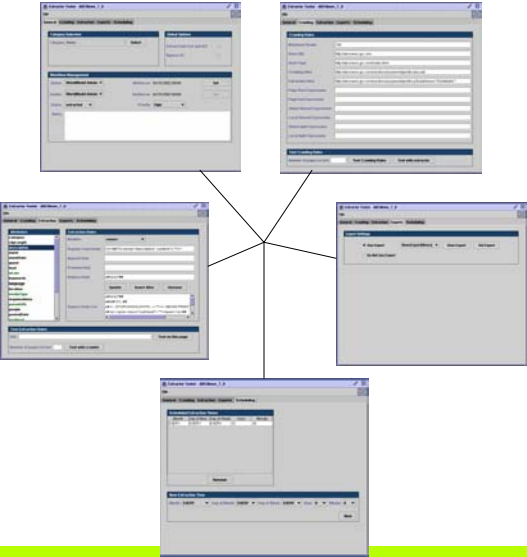
Attributes

- entity
 - parentURL: STRING
 - extractorID: INTEGER
 - sourceID: INTEGER
 - name: STRING
 - synonym: STRING
 - person
 - addressString: STRING
 - birthdate: DATE
 - analyst
 - stockAnalystID: INTEGER

- 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>**

Step 2: Knowledge Agent Creation

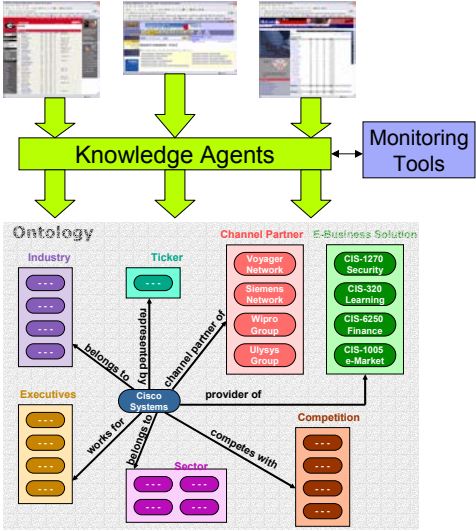
Create and configure Knowledge Agents to populate the Ontology



- 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

Step 3: Automatic aggregation of knowledge

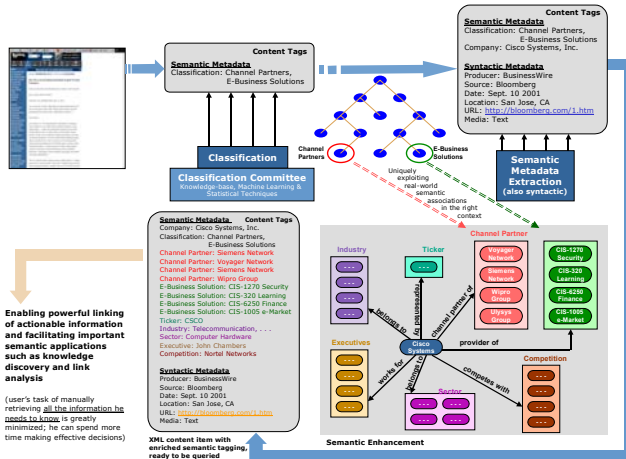
Automatic aggregation of knowledge from knowledge sources



- 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

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



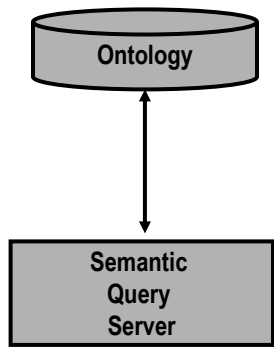
- How does it work?**
- Uses a hybrid of statistical, machine learning and knowledge-base techniques for classification
 - Not only classifies, but also enhances semantic metadata with associated domain knowledge

Enabling powerful linking of actionable information and facilitating important semantic applications such as knowledge discovery and link analysis

(user's task of manually retrieving all the information he needs to know is greatly minimized; he can spend more time making effective decisions)

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

Ontology-based Semagix solutions

- ◆ **Equity Analysis Workbench**
 - ◆ Heterogeneous internal and external, 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

POWER • THROUGH • RELEVANCE

Semantic Application Example – Analyst Workbench

The screenshot displays a financial analyst's workbench for Motorola, Inc. (MOT). The interface includes a top navigation bar with the company name and ticker symbol, a central data table with columns for Symbol, Change, Price, and Volume, and a right-hand pane with multiple news and analysis feeds. The feeds are categorized by topic, such as 'Company News', 'Earnings News', and 'Market Commentary News'. Each news item includes a headline, date, and source. Annotations with arrows point to specific elements: 'Automatic 3rd party content integration' points to the top navigation bar; 'Competitive research inferred automatically' points to a 'View competitors...' link; 'Focused relevant content organized by topic (semantic categorization)' points to the news feed categories; 'Related relevant content not explicitly asked for (semantic associations)' points to a 'Show competitors...' link; and 'Automatic Content Aggregation from multiple content providers and feeds' points to the bottom of the news feed pane.

POWER • THROUGH • RELEVANCE

CIRAS - Anti Money Laundering

(Know Your Customer – KYC)

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

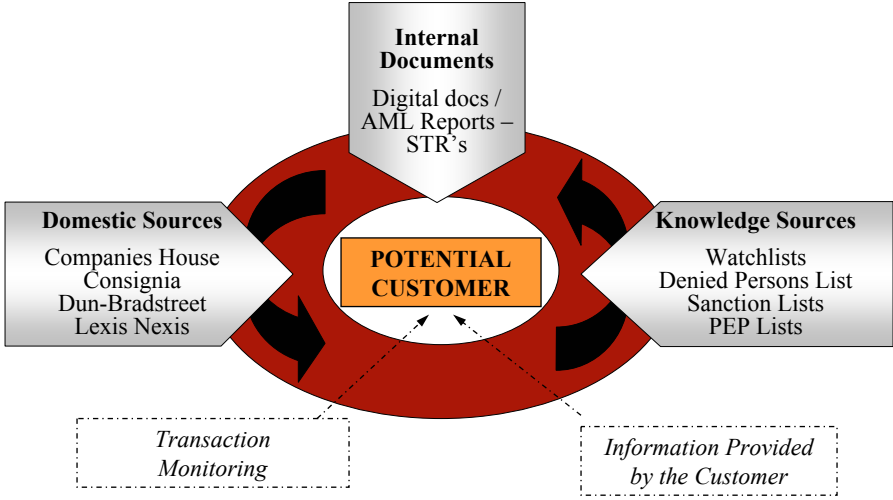
- ◆ Difficulty to link different pieces of information – **reduced effectiveness**
- ◆ Checks are sequential and resource intensive - **Increase process time and cost**
- ◆ Duplication of content – **increased subscription cost**

Inability to implement domain-specific 'best practises'

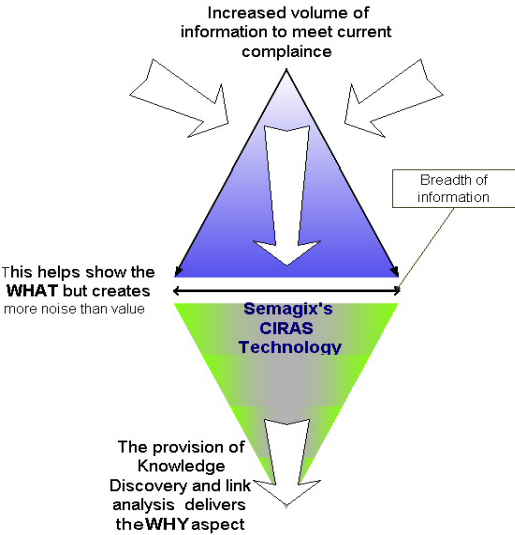
- ◆ Process knowledge resides with analysts – **variable quality of output**
- ◆ Difficulty to fine-tune processes to specific domain – **inflexible process**

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

Constituent parts of 'reasonable grounds'



What vs. Why



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

CIRAS's Components

The screenshot displays the SEMAGIX CIRAS interface with several key components highlighted:

- Customer Application Information:** Integration of structured information gathered during the account opening process. This includes fields for Company Name (Bayer AG), Nature of Business, Company Address, Incorporated in, Company Representative, Conducts Business in, and Representative's Title. A Risk Score section shows metrics for Company (90), Subscore (100), Link Analysis (90), and Appraisal (75).
- Relevant Knowledge:** A list of entities related to Bayer AG, including Bayer Corporation, Bayer Crop Science, Bayer Faser GmbH, Wacker Chemie, Klont Köln, Richard Don, Udo Oels, Wacker Chemie, and Work Levenskassen Levenskassen.
- Relevant Content:** A list of documents such as 'Credit and financial validation', 'Lexus Nexus Validation', and 'Lexus Nexus Validation'.
- Anti-Money Laundering Ontology:** A box indicating the system's focus on AML.
- Risk Weighting:** A box indicating the system's risk assessment capabilities.

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'

Actionable Information

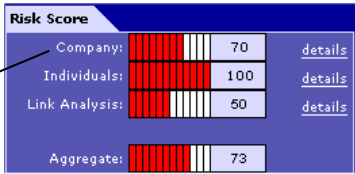
STATOIL GAS TRADING LTD - Details		
Score Component	Score	Reason
shareholder check	65	has a shareholder WOITEK MURAWSKI who works for RABITA TRUST which appears on Bank of England Sanctions List
shareholder check	65	has a shareholder WOITEK MURAWSKI who works for RABITA TRUST which appears on SDGI
Aggregate 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

CIRAS's Components

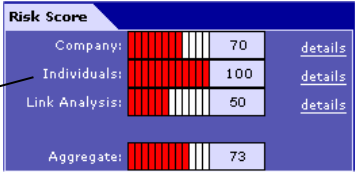


1. Company Analysis

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

CIRAS's Components

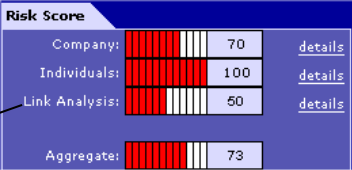


2. ID Verification

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

CIRAS's Components



3. Link Analysis Check

Score Component	Score	Reasons
Metabase Check	0.0	
Organisation Check	1.0	Akida Bank
Aggregate Score: 0.0		

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

CIRAS's Components

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

Company Knowledge

STATOIL GAS TRADING LTD [Company] Visualiser

Synonyms:
Statoil Gas Trading

Relationships:

- HAVARD BERGE** works for STATOIL GAS TRADING LTD
- RUNE BJORNSON** works for STATOIL GAS TRADING LTD
- MICHAEL KELLY** works for STATOIL GAS TRADING LTD
- WOJTEK MURAWSKI** is a shareholder in STATOIL GAS TRADING LTD
- STATOIL GAS TRADING LTD is audited by **Ernst & Young**
- STATOIL GAS TRADING LTD operates in **Gas Energy Marketing Company**
- STATOIL GAS TRADING LTD has address of **11a Regent St, United Kingdom, SW1Y 4AG**
- STATOIL GAS TRADING LTD has revenues of **GBP 952m**
- STATOIL GAS TRADING LTD is a subsidiary of **STATOIL IUK LTD**
- STATOIL GAS TRADING LTD conducts business in **Norway**
- STATOIL GAS TRADING LTD conducts business in **United Kingdom**

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

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

The screenshot displays a user interface with a blue header 'Relevant Documents'. Below it, there are two sections: 'EXTERNAL DOCUMENTS' and 'AUDIT TRAIL'. Under 'EXTERNAL DOCUMENTS', there are two entries: 'Statoil signs Iran gas deal' with a brief description and source 'news.bbc.co.uk', and 'Dun & Bradstreet Report' with a description and source 'http://neon/'. Under 'AUDIT TRAIL', there is one entry: 'Know Your Customer Check' with sub-headers 'Retrospective Check', 'Application Date', and 'Request Outcome'.

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

SEMAG!X
POWER • THROUGH • RELEVANCE

