3-26-2003

Ontology Driven Information Systems in Action (Capturing and Applying Existing Knowledge to Semantic Applications)

Amit P. Sheth
Wright State University - Main Campus, amit@sc.edu

Repository Citation

Sheth, A. P. (2003). Ontology Driven Information Systems in Action (Capturing and Applying Existing Knowledge to Semantic Applications). .
https://corescholar.libraries.wright.edu/knoesis/36

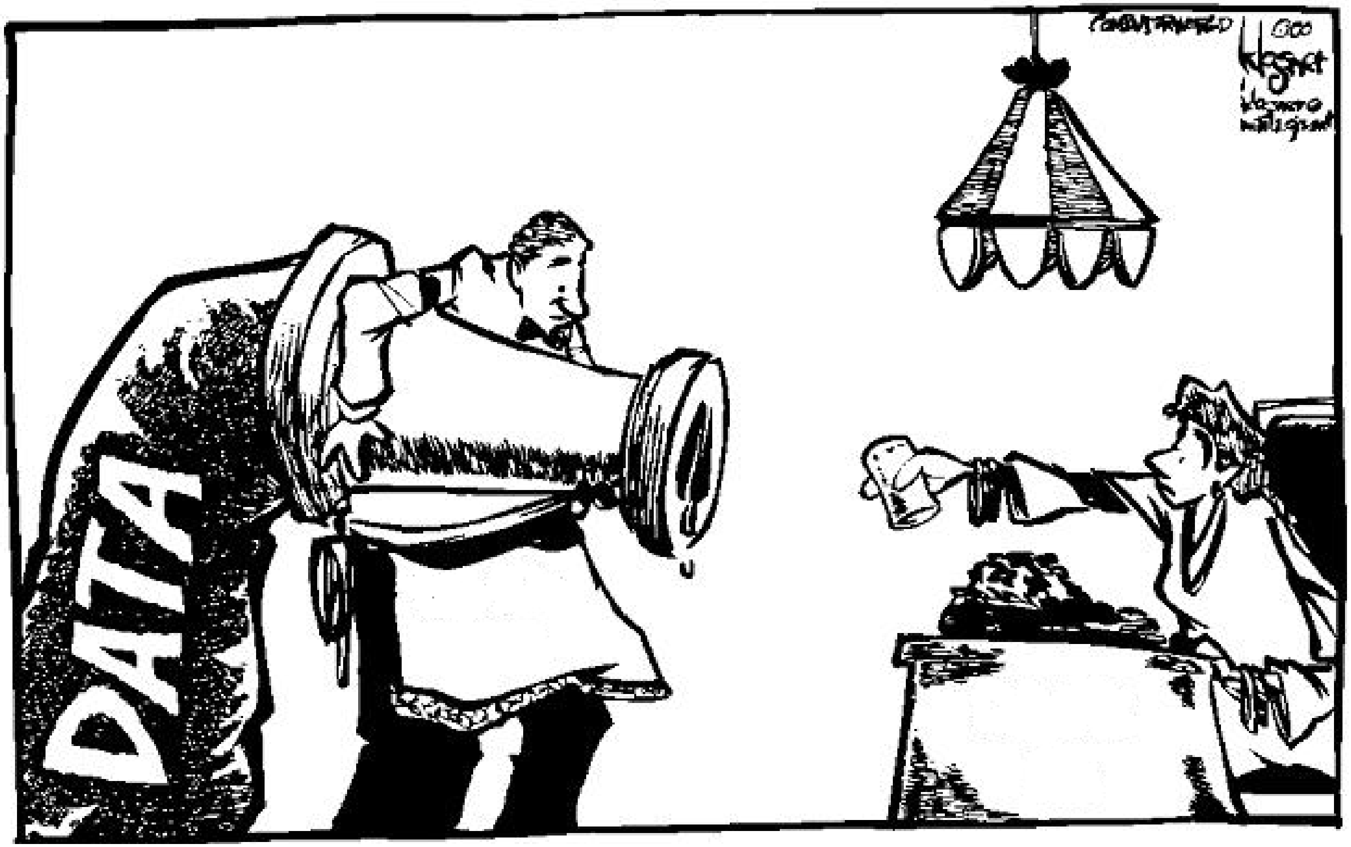
SEMAG!X
POWER • THROUGH • RELEVANCE

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



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

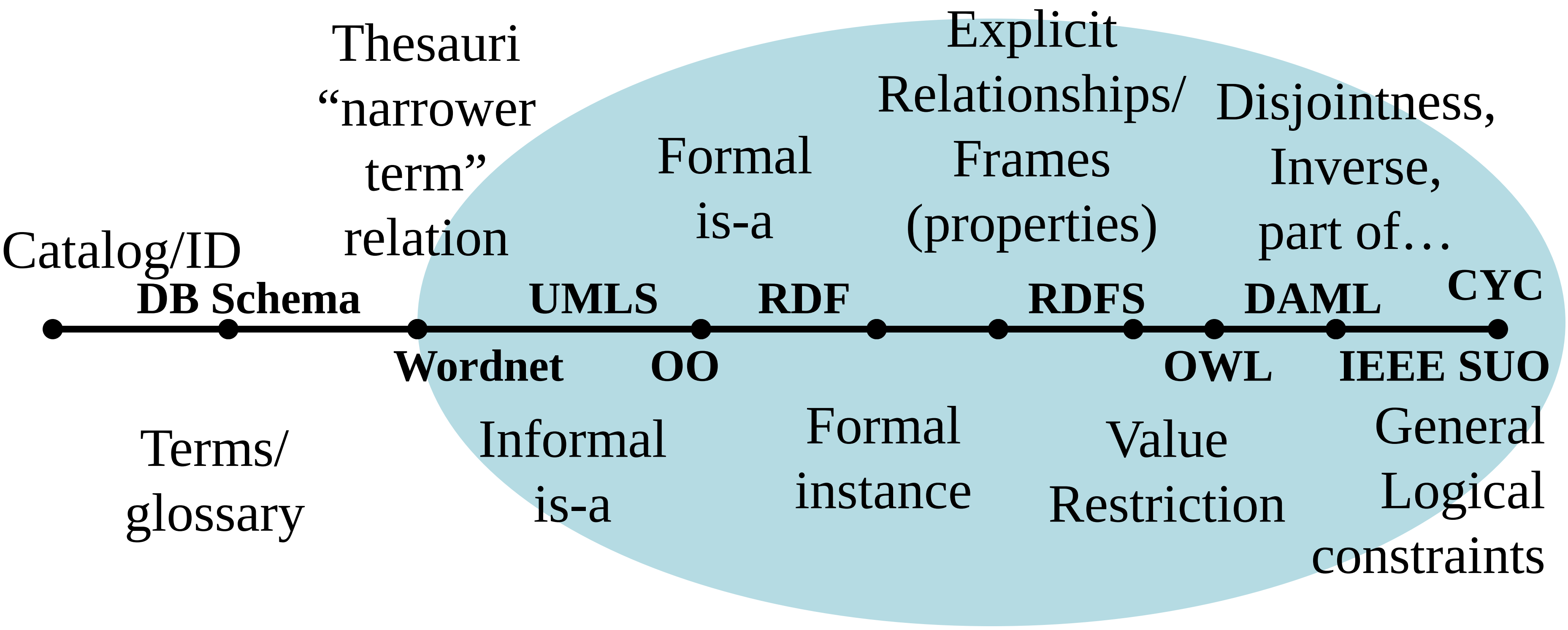
Achieved in the context of successful technology transfer from academic research (LSDIS lab, UGA's SCORE technology) into commercial product (Semagix's Freedom)

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



Catalog/ID

Thesauri “narrower term” relation

Formal is-a

Explicit Relationships/ Frames (properties)

Disjointness, Inverse, part of…

DB Schema

UMLS

RDF

RDFS

DAML

CYC

Wordnet

OO

OWL

IEEE SUO

Terms/ glossary

Informal is-a

Formal instance

Value Restriction

General Logical constraints

Simple Taxonomies

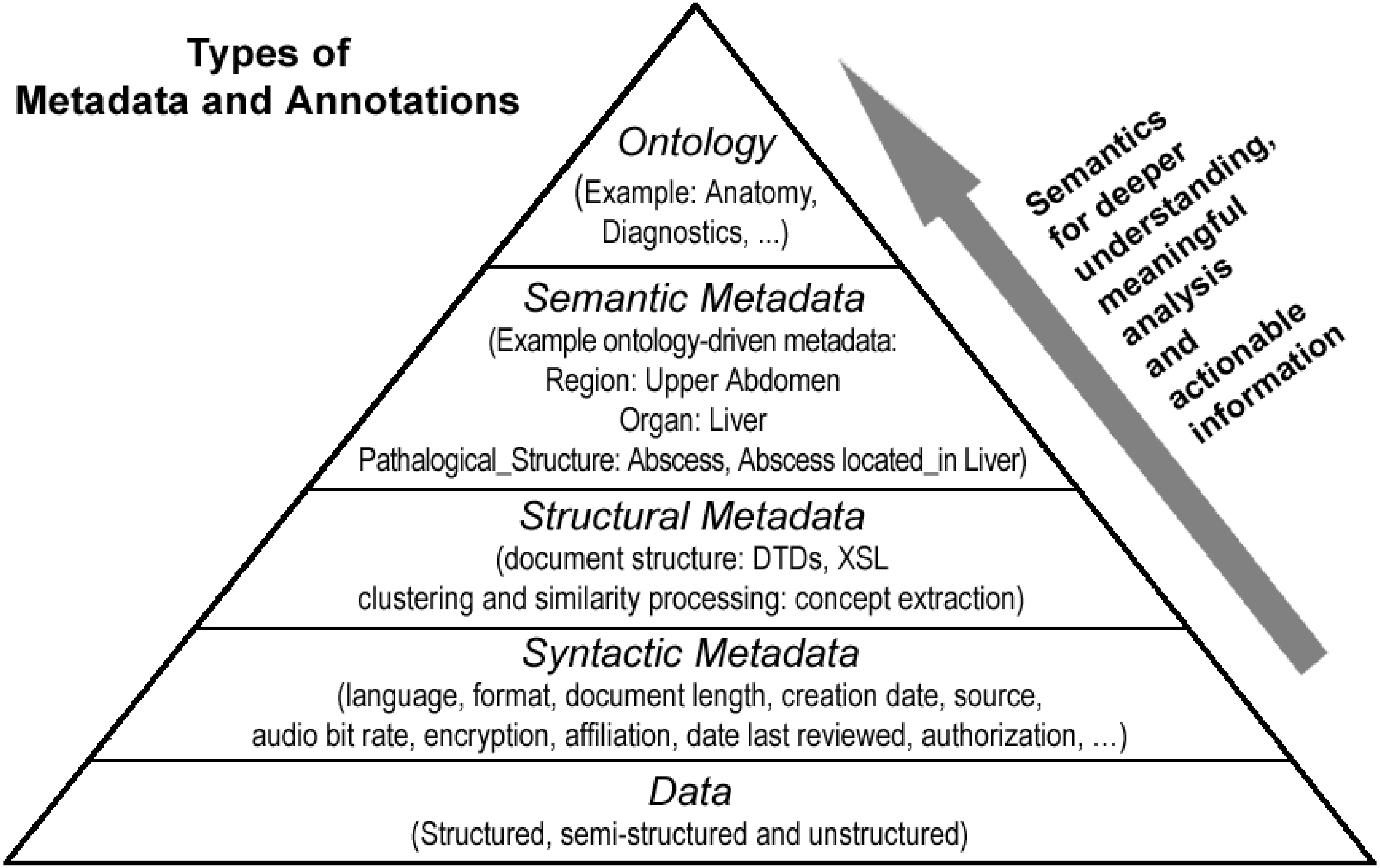
Expressive Ontologies

Better capability at higher complexity and computability

After McGuinness & Finin

Metadata and Ontology:
Primary Semantic Web enablers

Types of
Metadata and Annotations



Ontology
(Example: Anatomy,
Diagnostics, ...)

Semantic Metadata
(Example ontology-driven metadata:
Region: Upper Abdomen
Organ: Liver
Pathalogical_Structure: Abscess, Abscess located_in Liver)

Structural Metadata
(document structure: DTDs, XSL
clustering and similarity processing: concept extraction)

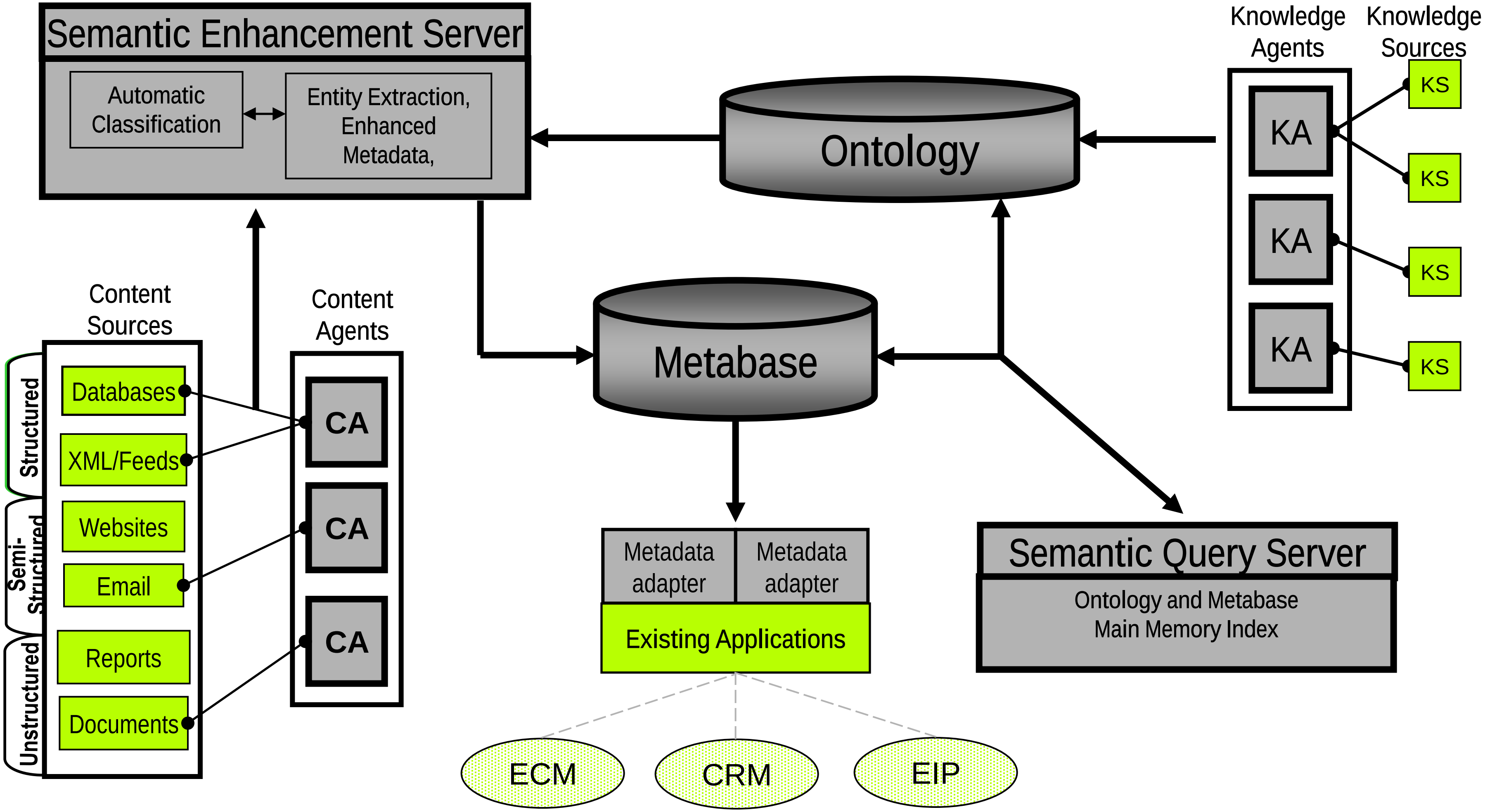
Syntactic Metadata
(language, format, document length, creation date, source,
audio bit rate, encryption, affiliation, date last reviewed, authorization, ...)

Data
(Structured, semi-structured and unstructured)

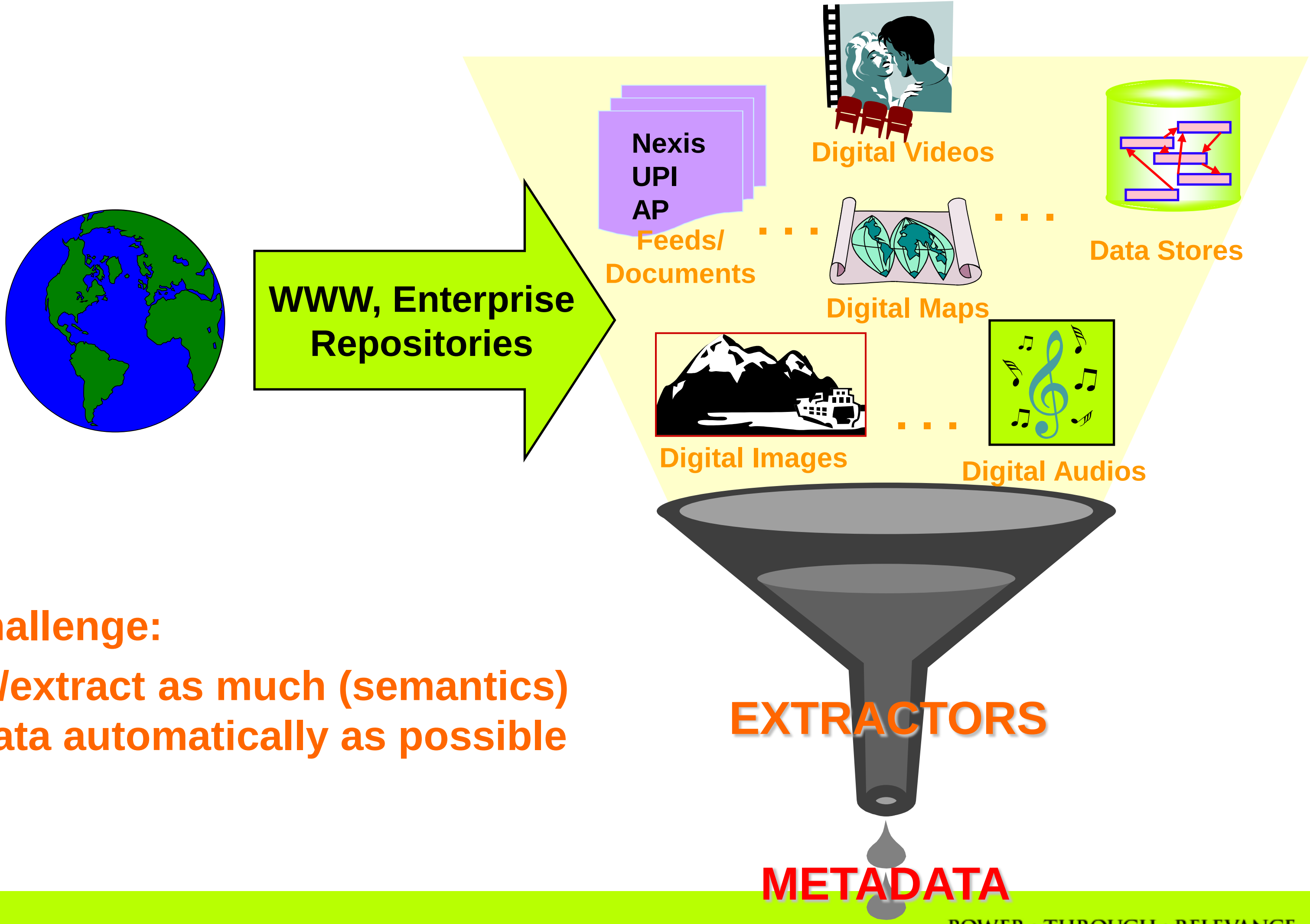
Semantics for deeper understanding, meaningful analysis and actionable information

Semagix Freedom Architecture

(a platform for building ontology-driven information system)



Information Extraction and Metadata Creation



WWW, Enterprise Repositories

Nexis
UPI
AP
Feeds/ Documents

Digital Videos

Digital Maps

Data Stores

Digital Images

Digital Audios

. . .

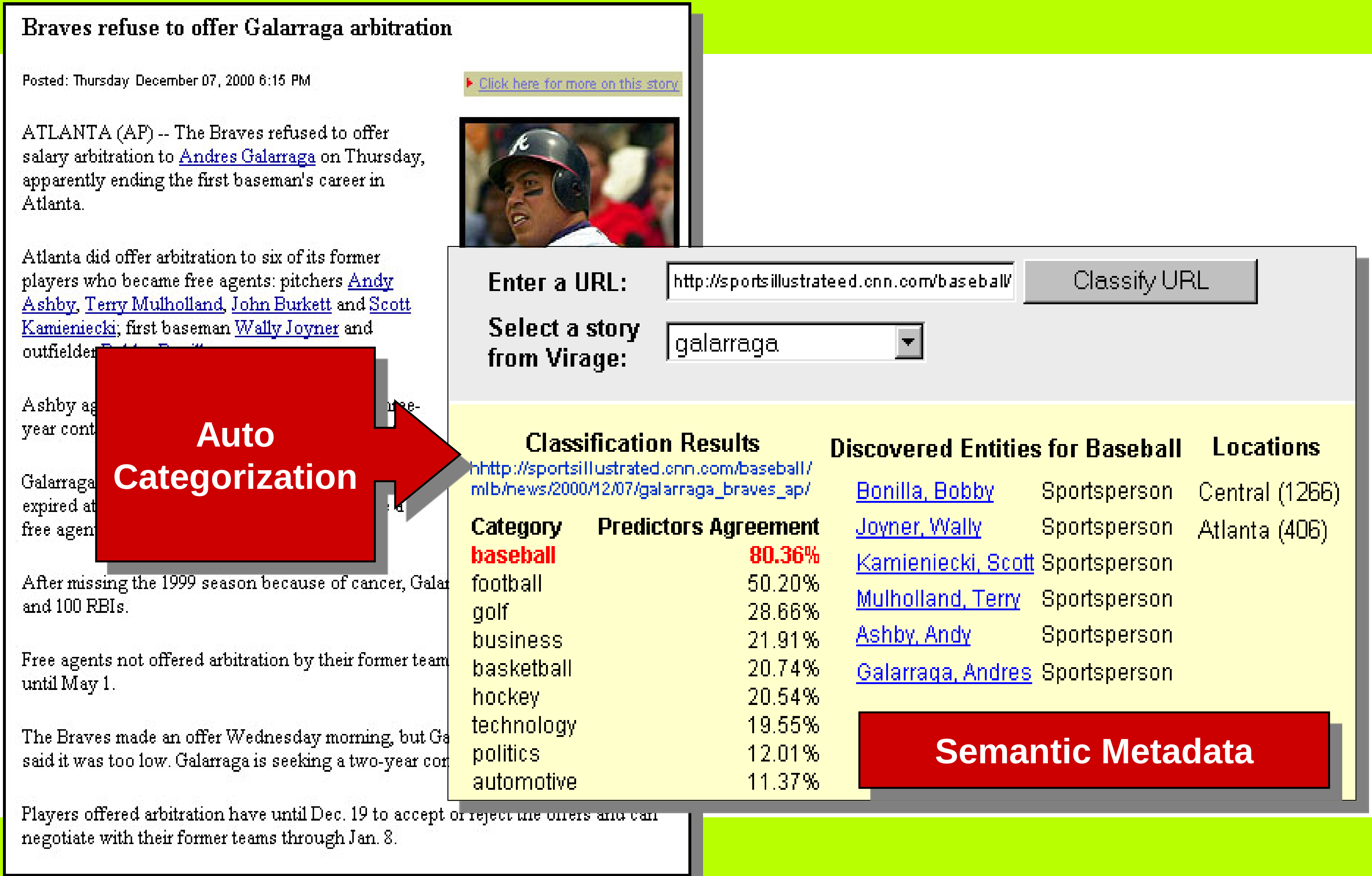
EXTRACTORS

METADATA

Key challenge:

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

Automatic Classification & Metadata Extraction (Web page)



Braves refuse to offer Galarraga arbitration

Posted: Thursday December 07, 2000 6:15 PM

Click here for more on this story

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

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

Ashby ag ree-
year cont

Galarraga
expired at
free agen

After missing the 1999 season because of cancer, Galar
and 100 RBIs.

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

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

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: http://sportsillustrateed.cnn.com/basebal/

Classify URL

Select a story from Virage: galarraga

Classification Results

hhttp://sportsillustrated.cnn.com/baseball/mlb/news/2000/12/07/galarraga_braves_ap/

Category	Predictors Agreement
baseball	**80.36%**
football	50.20%
golf	28.66%
business	21.91%
basketball	20.74%
hockey	20.54%
technology	19.55%
politics	12.01%
automotive	11.37%

Discovered Entities for Baseball

Bonilla, Bobby	Sportsperson
Joyner, Wally	Sportsperson
Kamieniecki, Scott	Sportsperson
Mulholland, Terry	Sportsperson
Ashby, Andy	Sportsperson
Galarraga, Andres	Sportsperson

Locations

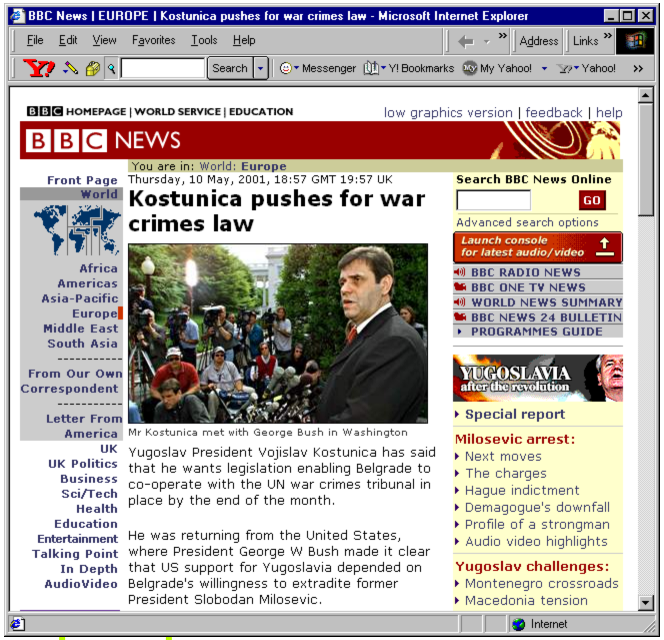
Central (1266)
Atlanta (406)

Semantic Metadata

Ontology-directed Metadata Extraction (Semi-structured data)

SEMAG!X

Web Page



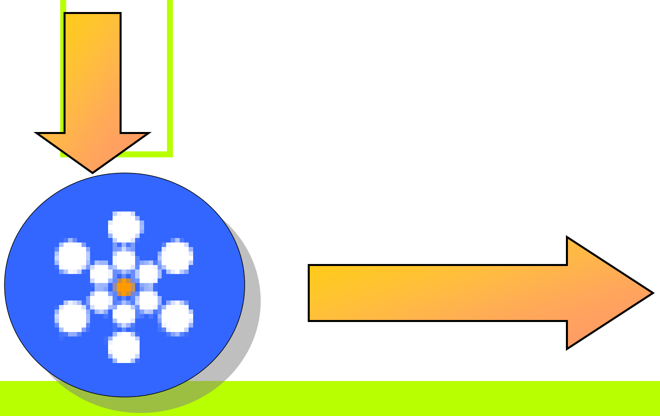
Thursday, 10 May, 2001, 18:57 GMT 19:57 UK

Kostunica pushes for war crimes law

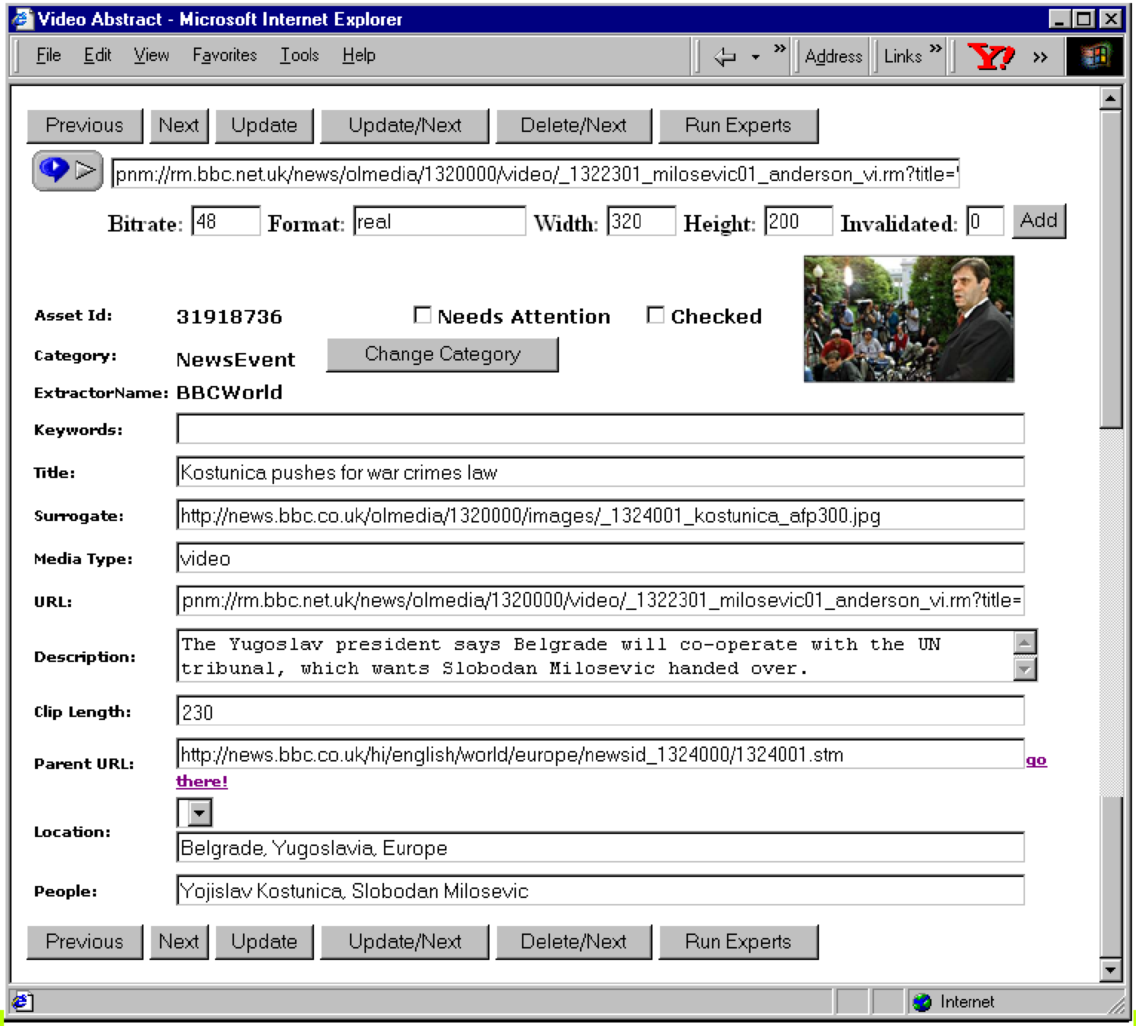
Mr Kostunica met with George Bush in Washington

Yugoslav President Vojislav Kostunica has said that he wants legislation enabling Belgrade to co-operate with the UN war crimes tribunal in place by the end of the month.

He was returning from the United States, where President George W Bush made it clear that US support for Yugoslavia depended on Belgrade's willingness to extradite former President Slobodan Milosevic.



Enhanced Metadata Asset



pnm://rm.bbc.net.uk/news/olmedia/1320000/video/_1322301_milosevic01_anderson_vi.rm?title='

Bitrate: 48 **Format:** real **Width:** 320 **Height:** 200 **Invalidated:** 0

Asset Id: 31918736 ☐ Needs Attention ☐ Checked

Category: NewsEvent

ExtractorName: BBCWorld

Keywords:

Title: Kostunica pushes for war crimes law

Surrogate: http://news.bbc.co.uk/olmedia/1320000/images/_1324001_kostunica_afp300.jpg

Media Type: video

URL: pnm://rm.bbc.net.uk/news/olmedia/1320000/video/_1322301_milosevic01_anderson_vi.rm?title=

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

Clip Length: 230

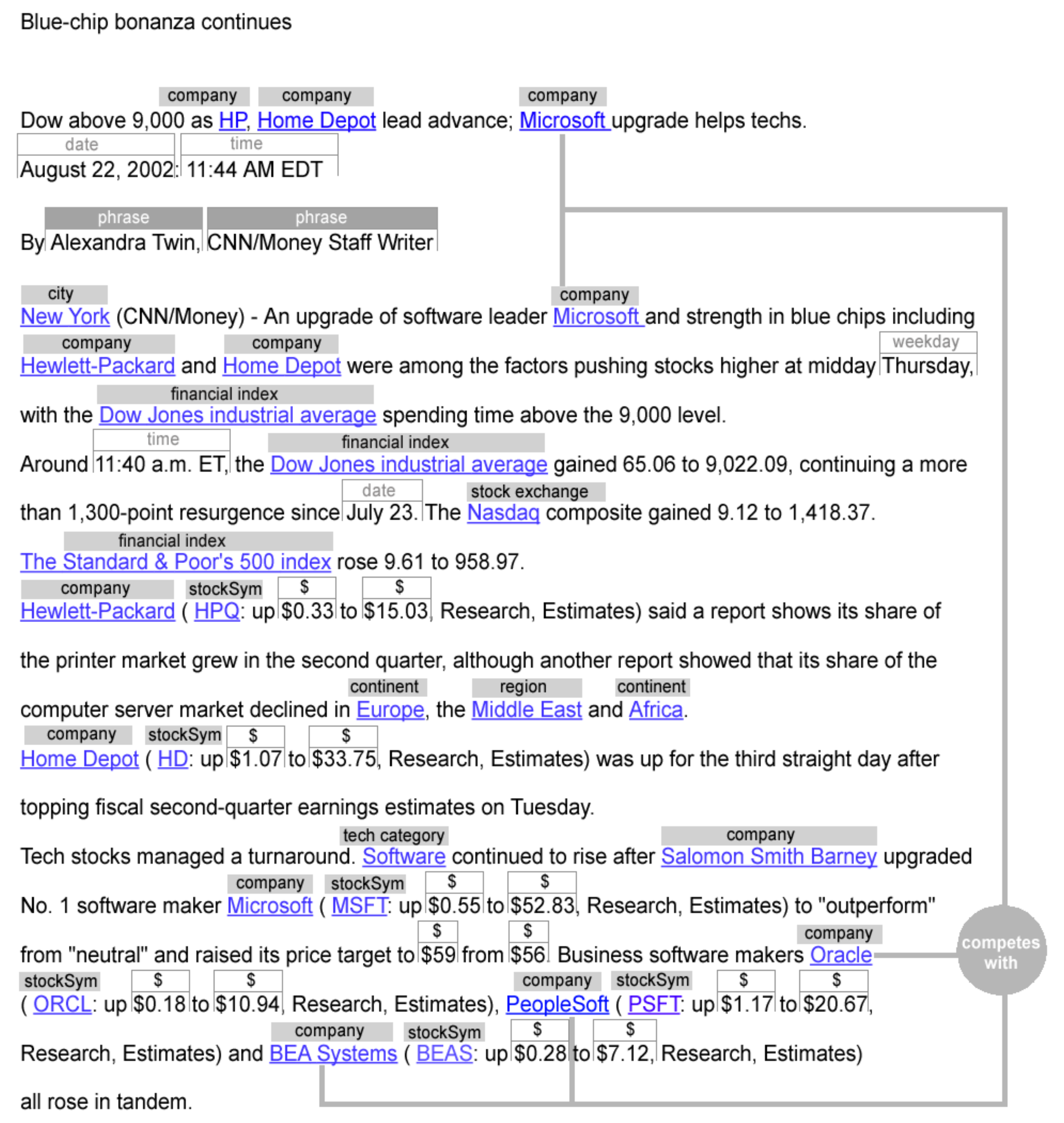
Parent URL: http://news.bbc.co.uk/hi/english/world/europe/newsid_1324000/1324001.stm

Location: Belgrade, Yugoslavia, Europe

People: Yojislav Kostunica, Slobodan Milosevic

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.

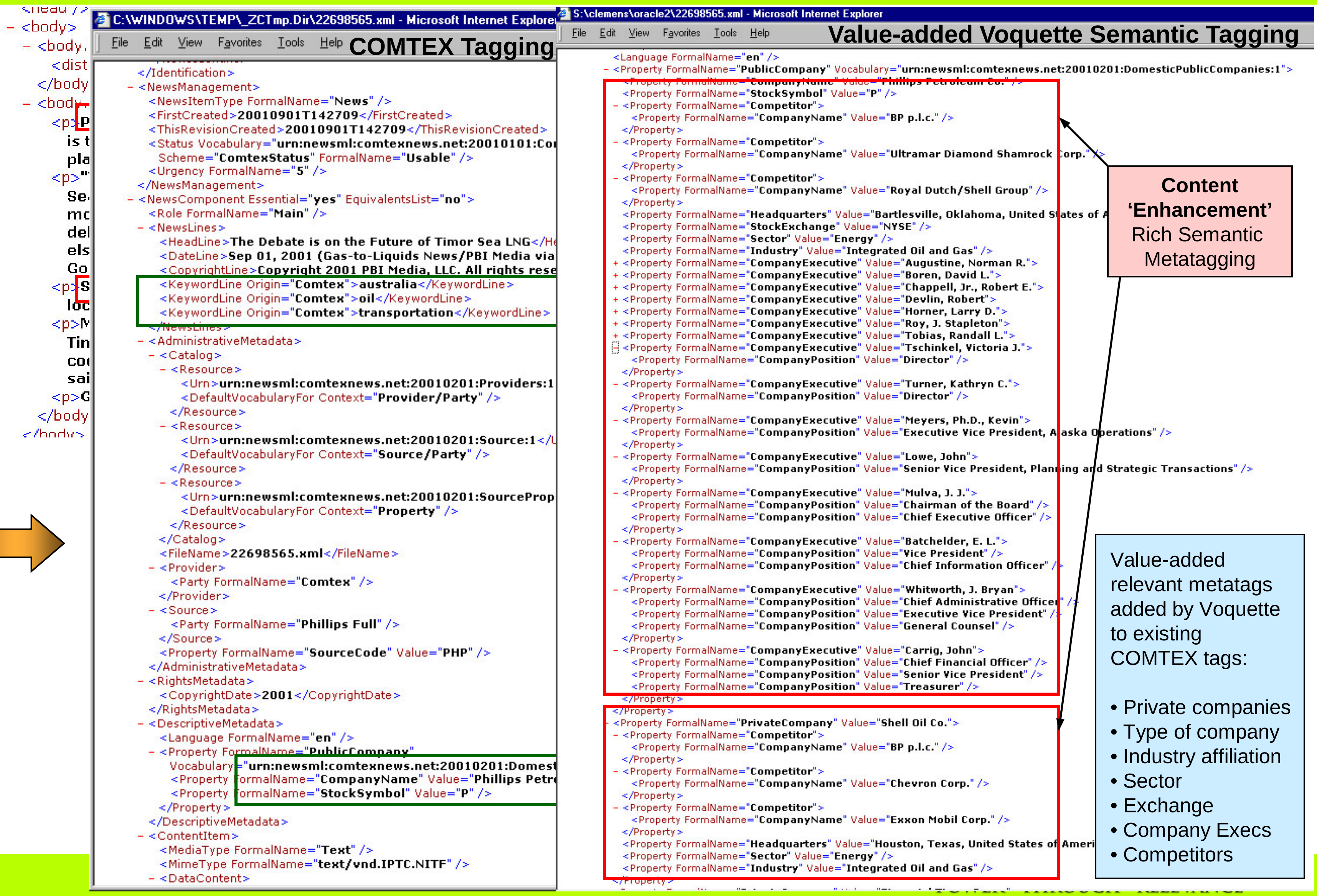
Around 11:40 a.m. ET, the Dow Jones industrial average gained 65.06 to 9,022.09, continuing a more than 1,300-point resurgence since July 23. The Nasdaq composite gained 9.12 to 1,418.37. The Standard & Poor's 500 index rose 9.61 to 958.97.

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.

Automatic Semantic Annotation



```
<head />
- <body>
  - <body.
      <dist
    </body
  - <body
      <p>P
        is t
        pla
      <p>"
        Se
        mc
        del
        els
        Go
      <p>S
        loc
      <p>M
        Tin
        co
        sai
      <p>G
    </body
  </body>
```

C:\WINDOWS\TEMP_ZCTmp.Dir\22698565.xml - Microsoft Internet Explorer

COMTEX Tagging

```
  </Identification>
- <NewsManagement>
    <NewsItemType FormalName="News" />
    <FirstCreated>20010901T142709</FirstCreated>
    <ThisRevisionCreated>20010901T142709</ThisRevisionCreated>
    <Status Vocabulary="urn:newsml:comtexnews.net:20010101:Co
      Scheme="ComtexStatus" FormalName="Usable" />
    <Urgency FormalName="5" />
  </NewsManagement>
- <NewsComponent Essential="yes" EquivalentsList="no">
    <Role FormalName="Main" />
  - <NewsLines>
      <HeadLine>The Debate is on the Future of Timor Sea LNG</He
      <DateLine>Sep 01, 2001 (Gas-to-Liquids News/PBI Media via
      <CopyrightLine>Copyright 2001 PBI Media, LLC. All rights rese
      <KeywordLine Origin="Comtex">australia</KeywordLine>
      <KeywordLine Origin="Comtex">oil</KeywordLine>
      <KeywordLine Origin="Comtex">transportation</KeywordLine>
    </NewsLines>
  - <AdministrativeMetadata>
    - <Catalog>
      - <Resource>
          <Urn>urn:newsml:comtexnews.net:20010201:Providers:1
          <DefaultVocabularyFor Context="Provider/Party" />
        </Resource>
      - <Resource>
          <Urn>urn:newsml:comtexnews.net:20010201:Source:1</U
          <DefaultVocabularyFor Context="Source/Party" />
        </Resource>
      - <Resource>
          <Urn>urn:newsml:comtexnews.net:20010201:SourceProp
          <DefaultVocabularyFor Context="Property" />
        </Resource>
      </Catalog>
      <FileName>22698565.xml</FileName>
    - <Provider>
        <Party FormalName="Comtex" />
      </Provider>
    - <Source>
        <Party FormalName="Phillips Full" />
      </Source>
      <Property FormalName="SourceCode" Value="PHP" />
    </AdministrativeMetadata>
  - <RightsMetadata>
      <CopyrightDate>2001</CopyrightDate>
    </RightsMetadata>
  - <DescriptiveMetadata>
      <Language FormalName="en" />
    - <Property FormalName="PublicCompany"
        Vocabulary="urn:newsml:comtexnews.net:20010201:Domest
        <Property FormalName="CompanyName" Value="Phillips Petr
        <Property FormalName="StockSymbol" Value="P" />
      </Property>
    </DescriptiveMetadata>
  - <ContentItem>
      <MediaType FormalName="Text" />
      <MimeType FormalName="text/vnd.IPTC.NITF" />
    - <DataContent>
```

S:\clemens\oracle2\22698565.xml - Microsoft Internet Explorer

Value-added Voquette Semantic Tagging

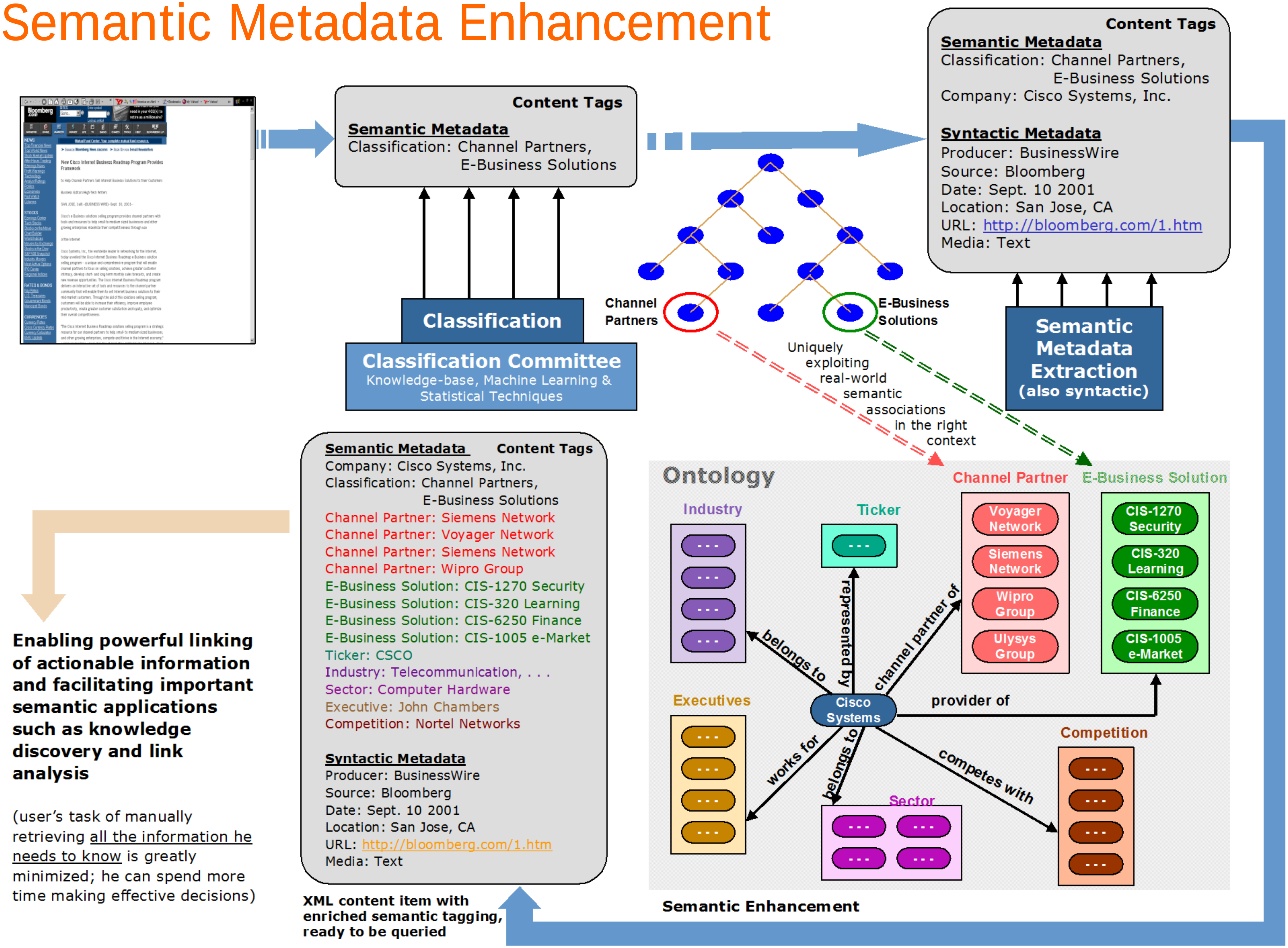
```
  <Language FormalName="en" />
- <Property FormalName="PublicCompany" Vocabulary="urn:newsml:comtexnews.net:20010201:DomesticPublicCompanies:1">
    <Property FormalName="CompanyName" Value="Phillips Petroleum Co." />
    <Property FormalName="StockSymbol" Value="P" />
  - <Property FormalName="Competitor">
      <Property FormalName="CompanyName" Value="BP p.l.c." />
    </Property>
  - <Property FormalName="Competitor">
      <Property FormalName="CompanyName" Value="Ultramar Diamond Shamrock Corp." />
    </Property>
  - <Property FormalName="Competitor">
      <Property FormalName="CompanyName" Value="Royal Dutch/Shell Group" />
    </Property>
    <Property FormalName="Headquarters" Value="Bartlesville, Oklahoma, United States of A
    <Property FormalName="StockExchange" Value="NYSE" />
    <Property FormalName="Sector" Value="Energy" />
    <Property FormalName="Industry" Value="Integrated Oil and Gas" />
  + <Property FormalName="CompanyExecutive" Value="Augustine, Norman R.">
  + <Property FormalName="CompanyExecutive" Value="Boren, David L.">
  + <Property FormalName="CompanyExecutive" Value="Chappell, Jr., Robert E.">
  + <Property FormalName="CompanyExecutive" Value="Devlin, Robert">
  + <Property FormalName="CompanyExecutive" Value="Horner, Larry D.">
  + <Property FormalName="CompanyExecutive" Value="Roy, J. Stapleton">
  + <Property FormalName="CompanyExecutive" Value="Tobias, Randall L.">
  - <Property FormalName="CompanyExecutive" Value="Tschinkel, Victoria J.">
      <Property FormalName="CompanyPosition" Value="Director" />
    </Property>
  - <Property FormalName="CompanyExecutive" Value="Turner, Kathryn C.">
      <Property FormalName="CompanyPosition" Value="Director" />
    </Property>
  - <Property FormalName="CompanyExecutive" Value="Meyers, Ph.D., Kevin">
      <Property FormalName="CompanyPosition" Value="Executive Vice President, Alaska Operations" />
    </Property>
  - <Property FormalName="CompanyExecutive" Value="Lowe, John">
      <Property FormalName="CompanyPosition" Value="Senior Vice President, Planning and Strategic Transactions" />
    </Property>
  - <Property FormalName="CompanyExecutive" Value="Mulva, J. J.">
      <Property FormalName="CompanyPosition" Value="Chairman of the Board" />
      <Property FormalName="CompanyPosition" Value="Chief Executive Officer" />
    </Property>
  - <Property FormalName="CompanyExecutive" Value="Batchelder, E. L.">
      <Property FormalName="CompanyPosition" Value="Vice President" />
      <Property FormalName="CompanyPosition" Value="Chief Information Officer" />
    </Property>
  - <Property FormalName="CompanyExecutive" Value="Whitworth, J. Bryan">
      <Property FormalName="CompanyPosition" Value="Chief Administrative Officer" />
      <Property FormalName="CompanyPosition" Value="Executive Vice President" />
      <Property FormalName="CompanyPosition" Value="General Counsel" />
    </Property>
  - <Property FormalName="CompanyExecutive" Value="Carrig, John">
      <Property FormalName="CompanyPosition" Value="Chief Financial Officer" />
      <Property FormalName="CompanyPosition" Value="Senior Vice President" />
      <Property FormalName="CompanyPosition" Value="Treasurer" />
    </Property>
  </Property>
- <Property FormalName="PrivateCompany" Value="Shell Oil Co.">
  - <Property FormalName="Competitor">
      <Property FormalName="CompanyName" Value="BP p.l.c." />
    </Property>
  - <Property FormalName="Competitor">
      <Property FormalName="CompanyName" Value="Chevron Corp." />
    </Property>
  - <Property FormalName="Competitor">
      <Property FormalName="CompanyName" Value="Exxon Mobil Corp." />
    </Property>
    <Property FormalName="Headquarters" Value="Houston, Texas, United States of Ameri
    <Property FormalName="Sector" Value="Energy" />
    <Property FormalName="Industry" Value="Integrated Oil and Gas" />
  </Property>
```

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



Content Tags

Semantic Metadata
Classification: Channel Partners,
E-Business Solutions

Classification

Classification Committee
Knowledge-base, Machine Learning & Statistical Techniques

Content Tags

Semantic Metadata
Classification: Channel Partners,
E-Business Solutions
Company: Cisco Systems, Inc.

Syntactic Metadata
Producer: BusinessWire
Source: Bloomberg
Date: Sept. 10 2001
Location: San Jose, CA
URL: http://bloomberg.com/1.htm
Media: Text

Semantic Metadata Extraction (also syntactic)

Semantic Metadata **Content Tags**
Company: Cisco Systems, Inc.
Classification: Channel Partners,
E-Business Solutions
Channel Partner: Siemens Network
Channel Partner: Voyager Network
Channel Partner: Siemens Network
Channel Partner: Wipro Group
E-Business Solution: CIS-1270 Security
E-Business Solution: CIS-320 Learning
E-Business Solution: CIS-6250 Finance
E-Business Solution: CIS-1005 e-Market
Ticker: CSCO
Industry: Telecommunication, . . .
Sector: Computer Hardware
Executive: John Chambers
Competition: Nortel Networks

Syntactic Metadata
Producer: BusinessWire
Source: Bloomberg
Date: Sept. 10 2001
Location: San Jose, CA
URL: http://bloomberg.com/1.htm
Media: Text

XML content item with enriched semantic tagging, ready to be queried

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)

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

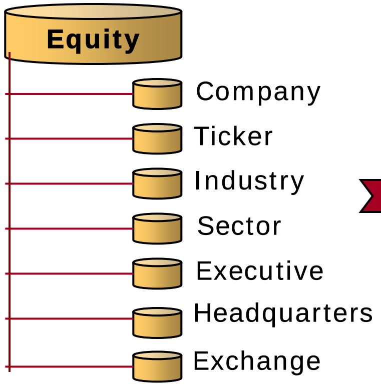
Practical Questions (for developing typical industry and application ontologies)

- 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
        - 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

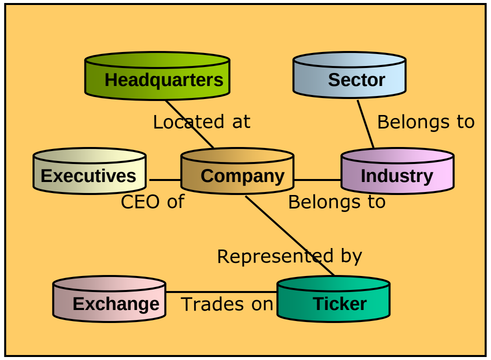
Practical Ontology Development Observation by Semagix

- 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

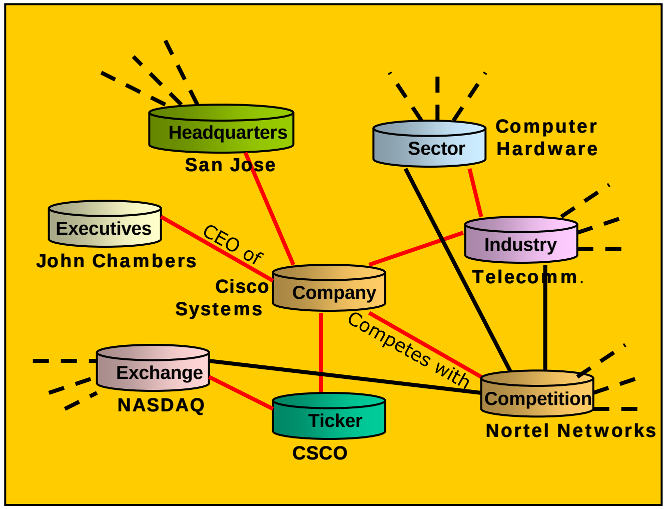
Ontology Example (Financial Equity domain)



Equity Metabase Model



Equity Ontology Descriptional Componet



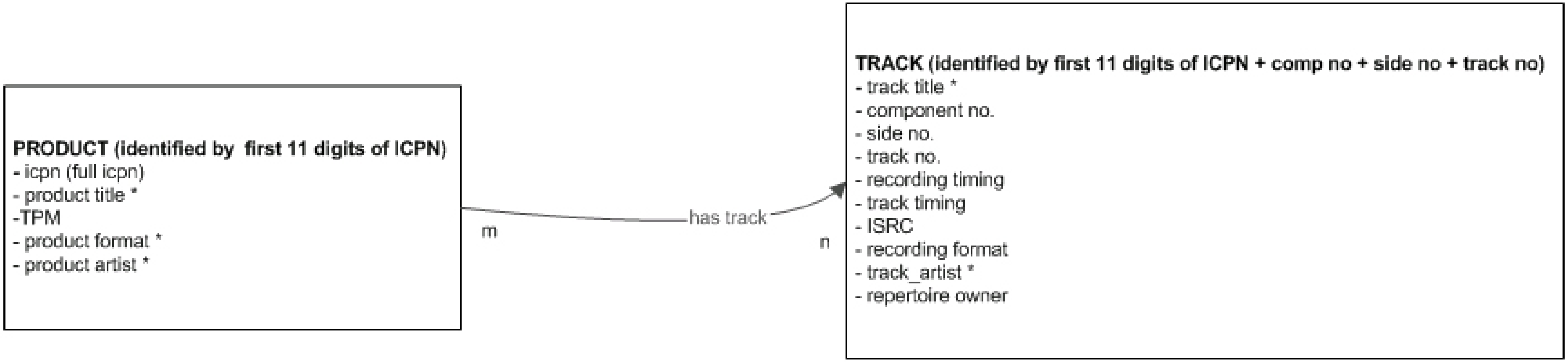
Equity Ontology (Assertional Component; (knowledge / facts)

o o o o

Equity Ontology

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: "*<u>has track</u>*"
  - Many-to-many relationship between the two entity classes
    - A product can have multiple tracks
    - A track can belong to multiple products

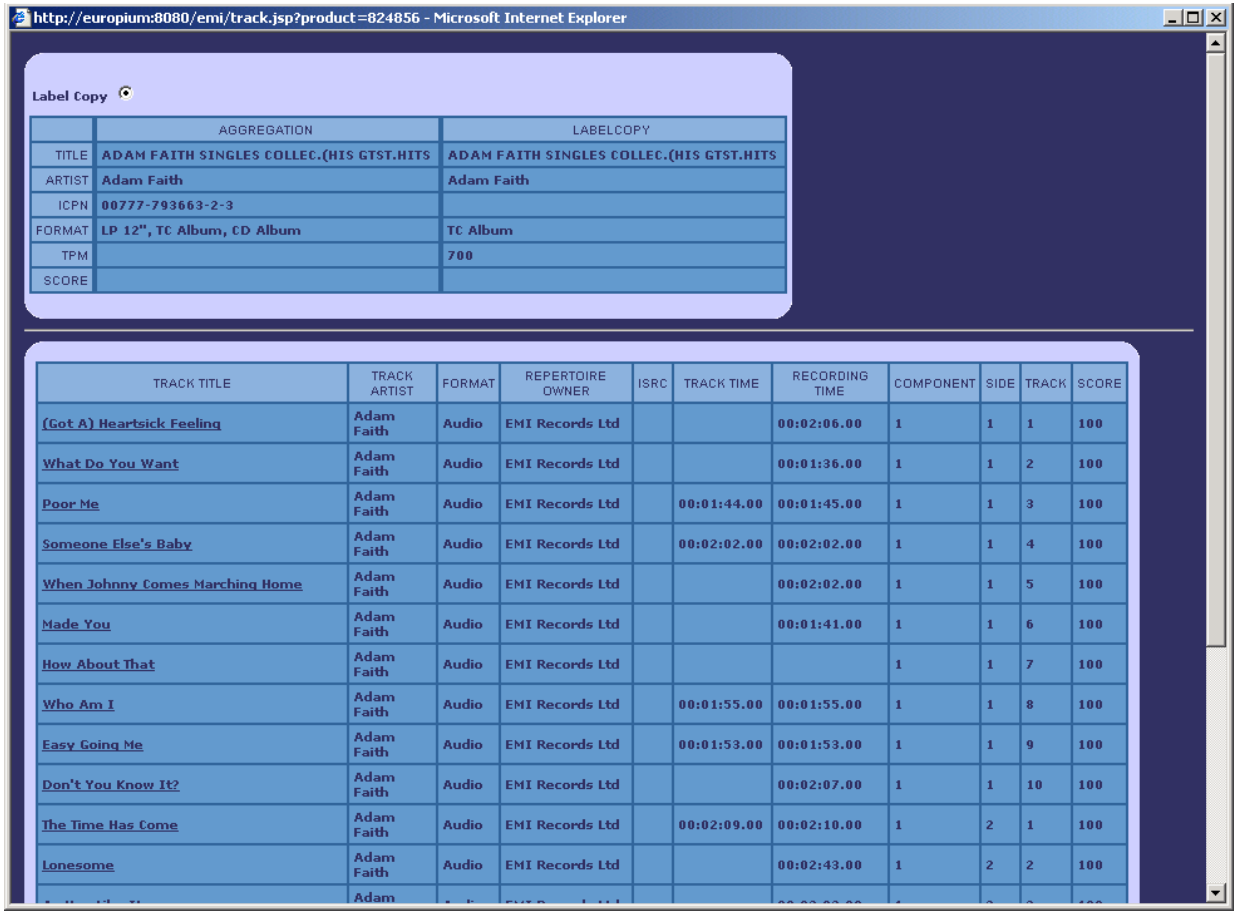


PRODUCT (identified by first 11 digits of ICPN)
- icpn (full icpn)
- product title *
-TPM
- product format *
- product artist *

TRACK (identified by first 11 digits of ICPN + comp no + side no + track no)
- track title *
- component no.
- side no.
- track no.
- recording timing
- track timing
- ISRC
- recording format
- track_artist *
- repertoire owner

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)


http://europium:8080/emi/track.jsp?product=824856 - Microsoft Internet Explorer

Label Copy ●

	AGGREGATION	LABELCOPY
TITLE	ADAM FAITH SINGLES COLLEC.(HIS GTST.HITS	ADAM FAITH SINGLES COLLEC.(HIS GTST.HITS
ARTIST	Adam Faith	Adam Faith
ICPN	00777-793663-2-3	
FORMAT	LP 12", TC Album, CD Album	TC Album
TPM		700
SCORE		

TRACK TITLE	TRACK ARTIST	FORMAT	REPERTOIRE OWNER	ISRC	TRACK TIME	RECORDING TIME	COMPONENT	SIDE	TRACK	SCORE
(Got A) Heartsick Feeling	Adam Faith	Audio	EMI Records Ltd			00:02:06.00	1	1	1	100
What Do You Want	Adam Faith	Audio	EMI Records Ltd			00:01:36.00	1	1	2	100
Poor Me	Adam Faith	Audio	EMI Records Ltd		00:01:44.00	00:01:45.00	1	1	3	100
Someone Else's Baby	Adam Faith	Audio	EMI Records Ltd		00:02:02.00	00:02:02.00	1	1	4	100
When Johnny Comes Marching Home	Adam Faith	Audio	EMI Records Ltd			00:02:02.00	1	1	5	100
Made You	Adam Faith	Audio	EMI Records Ltd			00:01:41.00	1	1	6	100
How About That	Adam Faith	Audio	EMI Records Ltd				1	1	7	100
Who Am I	Adam Faith	Audio	EMI Records Ltd		00:01:55.00	00:01:55.00	1	1	8	100
Easy Going Me	Adam Faith	Audio	EMI Records Ltd		00:01:53.00	00:01:53.00	1	1	9	100
Don't You Know It?	Adam Faith	Audio	EMI Records Ltd			00:02:07.00	1	1	10	100
The Time Has Come	Adam Faith	Audio	EMI Records Ltd		00:02:09.00	00:02:10.00	1	2	1	100
Lonesome	Adam Faith	Audio	EMI Records Ltd			00:02:43.00	1	2	2	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** (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)

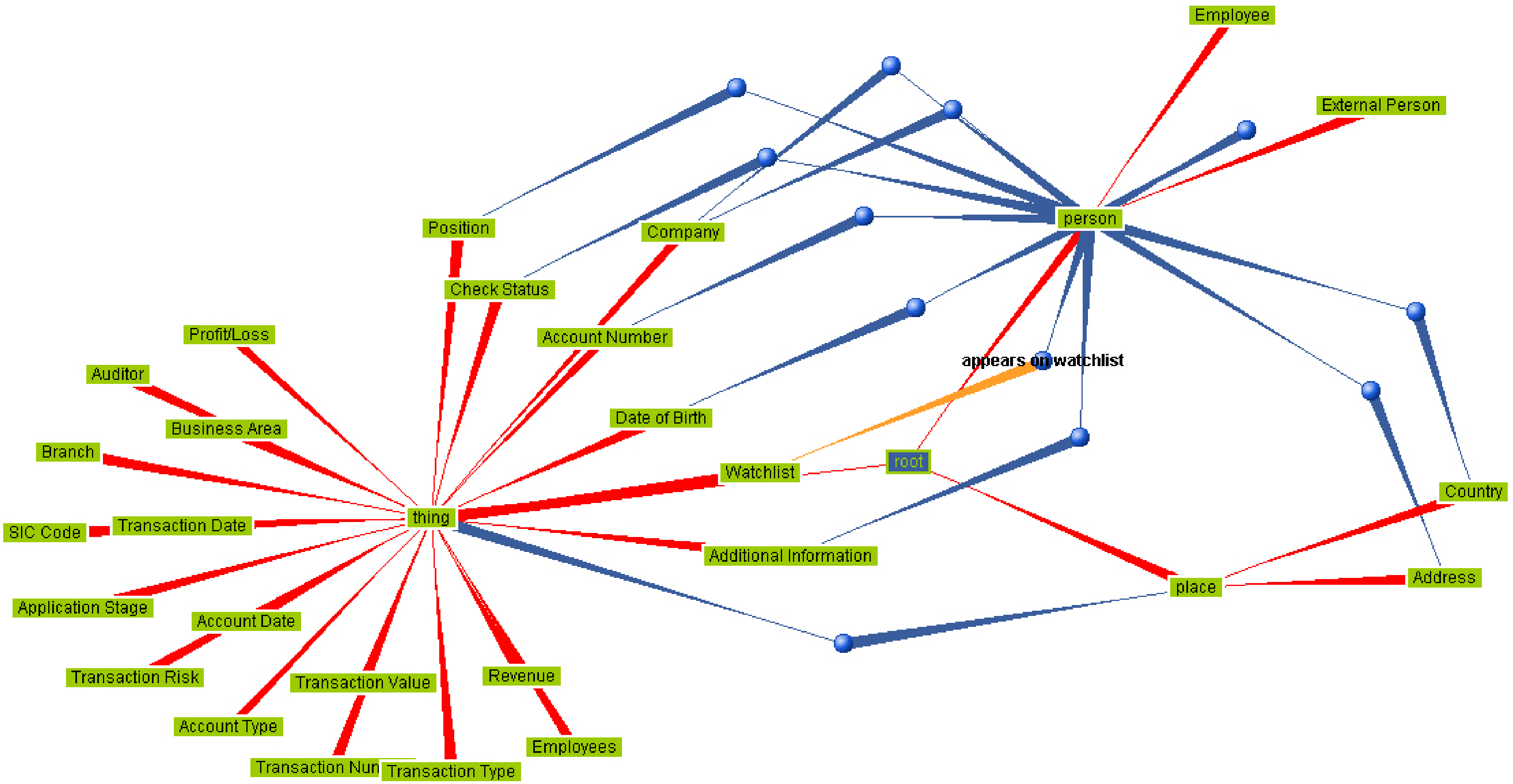
  - 5 Knowledge Agents, one for each database
  - Automated ontology population using Knowledge Agents took <u>no longer than a day</u> for all the Agents

Example of Ontology with complex schema

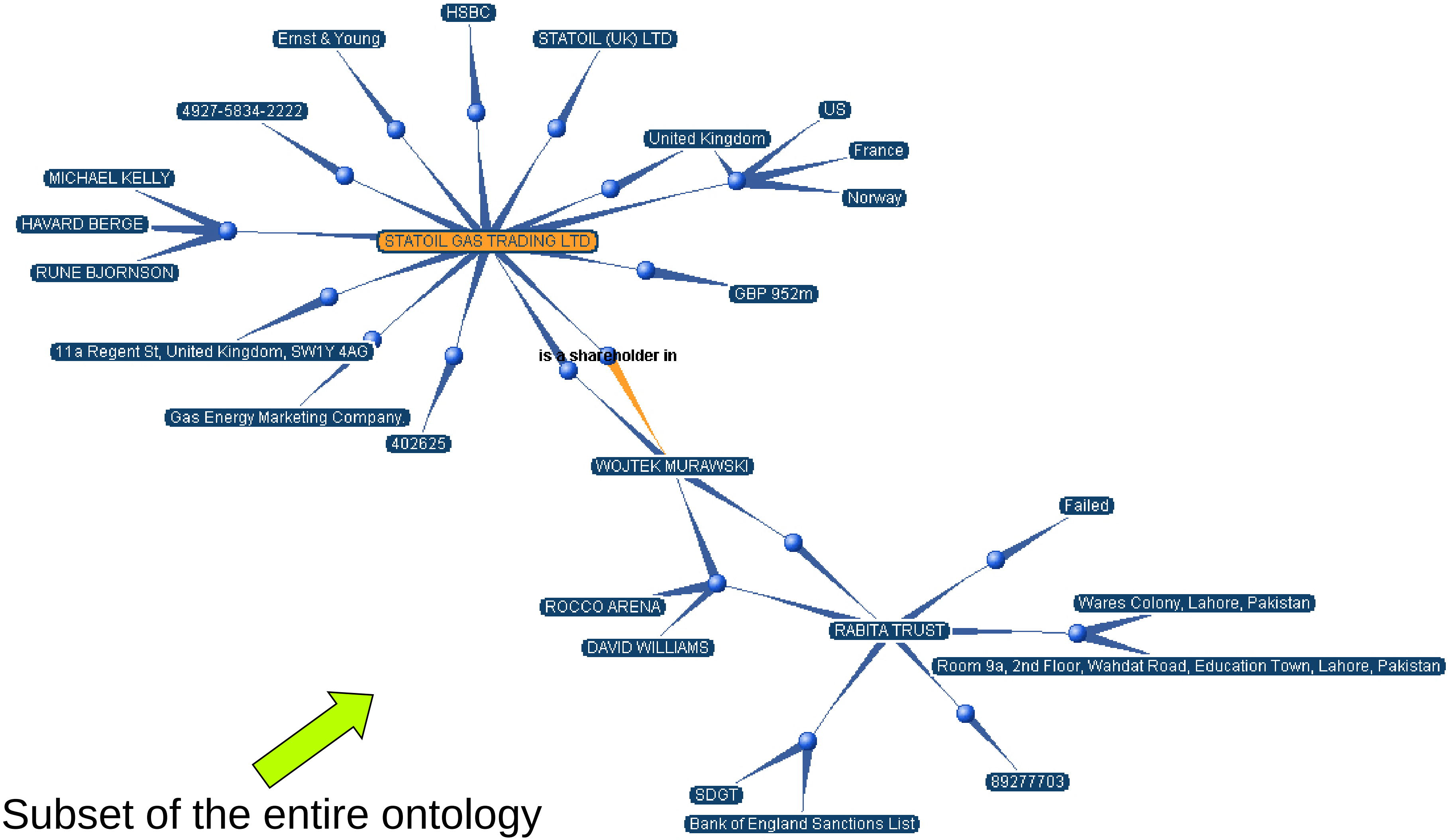
- **Ontology for Anti-money Laundering (AML) application in Financial Industry**
- **Ontology Schema (Descriptional Component)**
  - About 40 entity classes
  - About 100 attribute types
  - About 50 relationship types between entity classes

SEMAGIX

AML Ontology Schema (Descriptional Component)



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

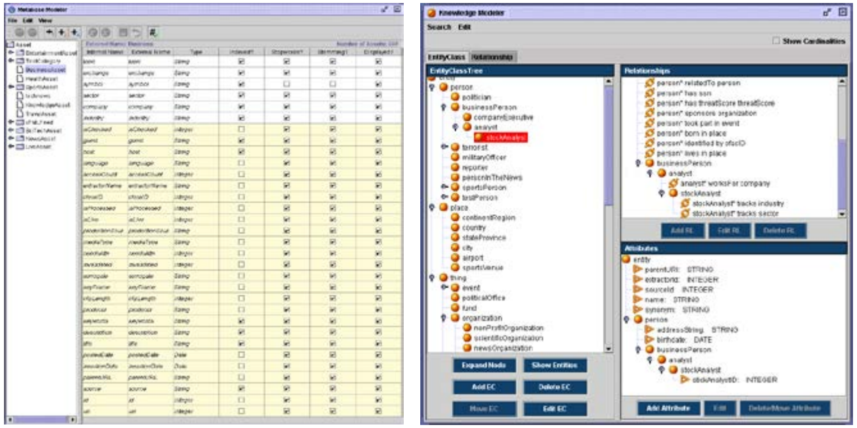
Effort Involved

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

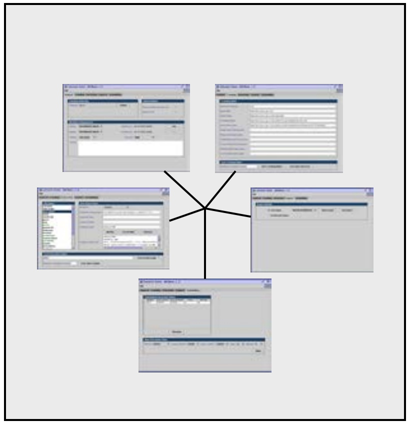
Technical Challenges Faced

- **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**

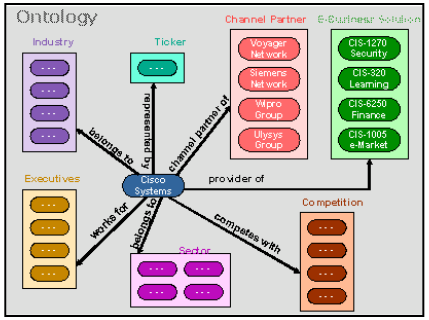
Ontology Creation and Maintenance Steps



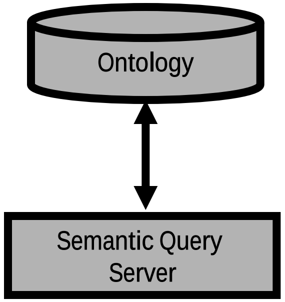
1. Ontology Model Creation



2. Knowledge Agent Creation



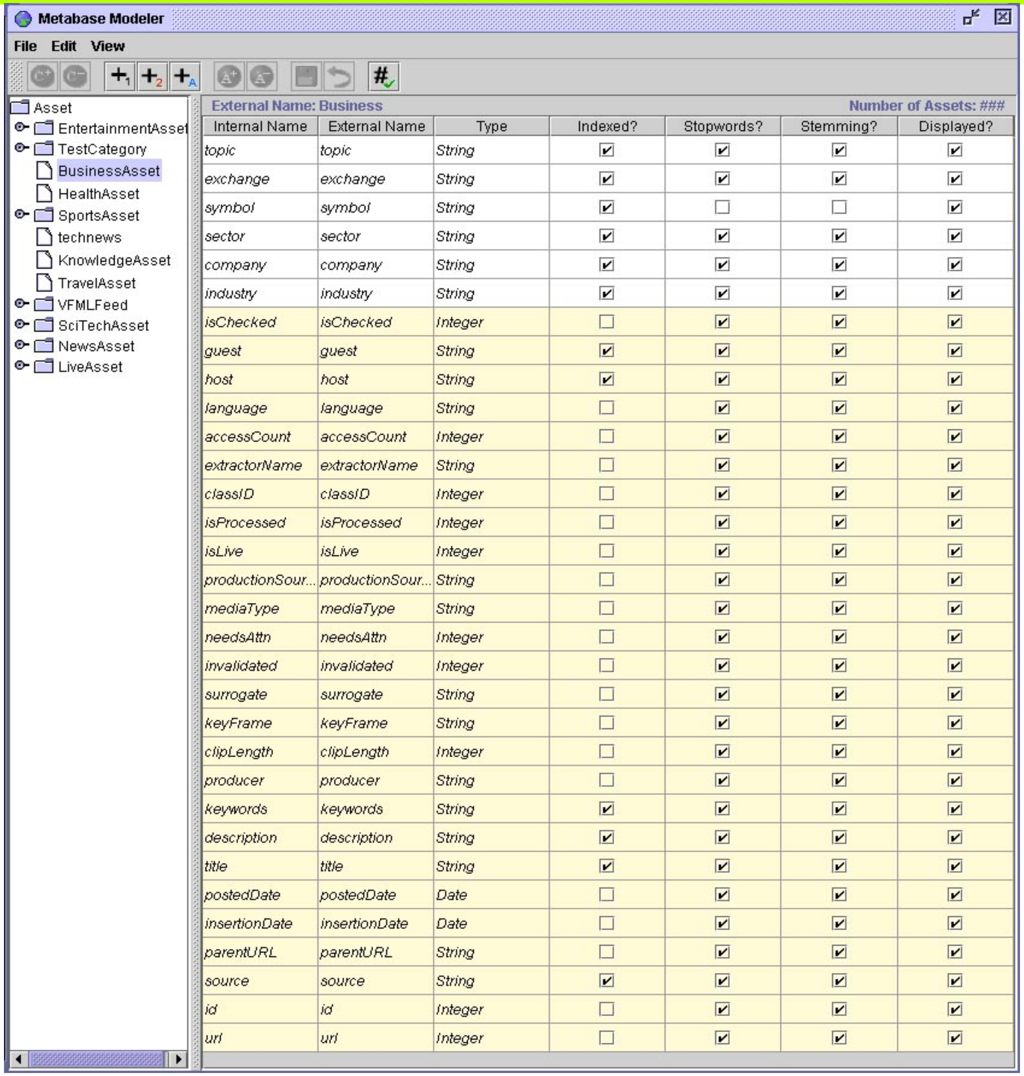
3. Automatic aggregation of Knowledge



4. Querying the Ontology

Step 1: Ontology Model Creation

Create an Ontology Model using Semagix Freedom Toolkit GUIs



External Name: Business

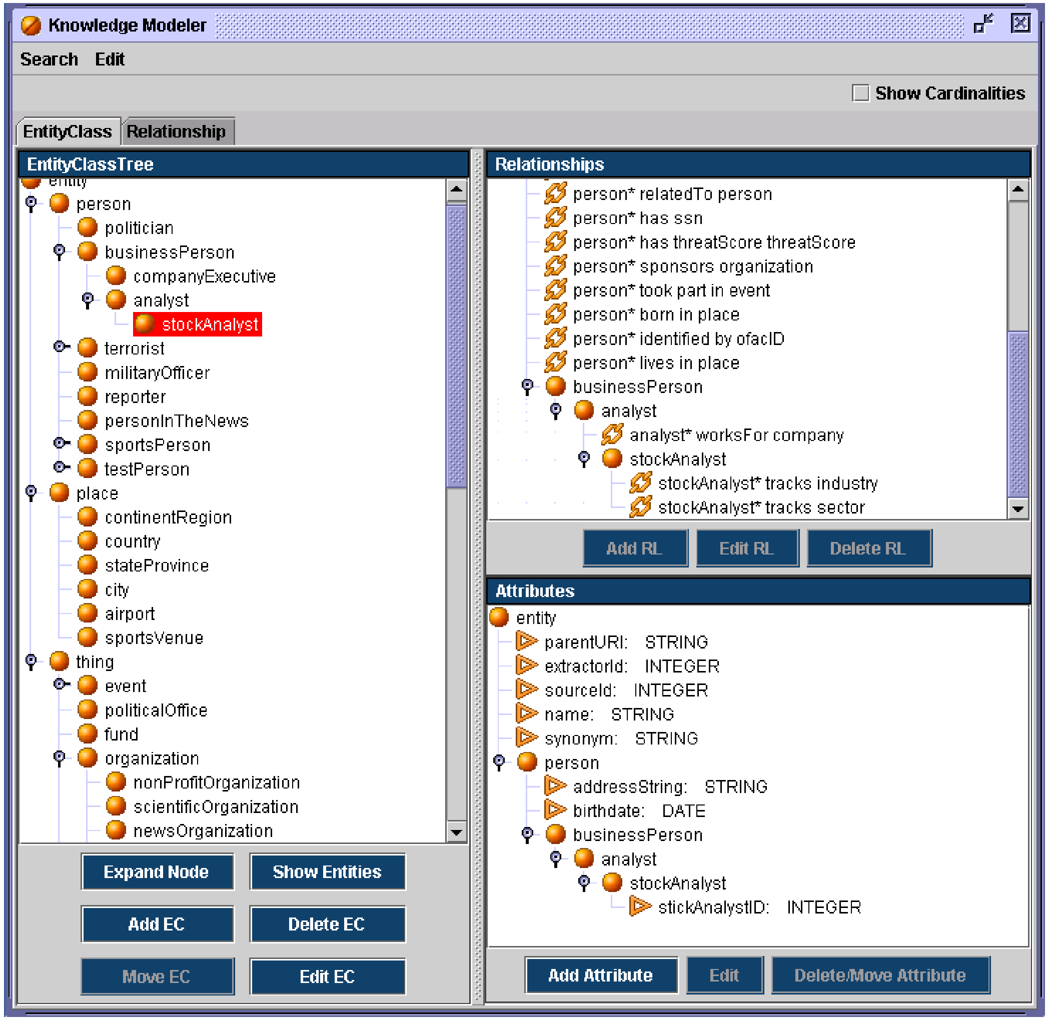
Number of Assets: ###

Internal Name	External Name	Type	Indexed?	Stopwords?	Stemming?	Displayed?
topic	topic	String	☑	☑	☑	☑
exchange	exchange	String	☑	☑	☑	☑
symbol	symbol	String	☑	☐	☐	☑
sector	sector	String	☑	☑	☑	☑
company	company	String	☑	☑	☑	☑
industry	industry	String	☑	☑	☑	☑
isChecked	isChecked	Integer	☐	☑	☑	☑
guest	guest	String	☑	☑	☑	☑
host	host	String	☑	☑	☑	☑
language	language	String	☐	☑	☑	☑
accessCount	accessCount	Integer	☐	☑	☑	☑
extractorName	extractorName	String	☐	☑	☑	☑
classID	classID	Integer	☐	☑	☑	☑
isProcessed	isProcessed	Integer	☐	☑	☑	☑
isLive	isLive	Integer	☐	☑	☑	☑
productionSour...	productionSour...	String	☐	☑	☑	☑
mediaType	mediaType	String	☐	☑	☑	☑
needsAttn	needsAttn	Integer	☐	☑	☑	☑
invalidated	invalidated	Integer	☐	☑	☑	☑
surrogate	surrogate	String	☐	☑	☑	☑
keyFrame	keyFrame	String	☐	☑	☑	☑
clipLength	clipLength	Integer	☐	☑	☑	☑
producer	producer	String	☐	☑	☑	☑
keywords	keywords	String	☑	☑	☑	☑
description	description	String	☑	☑	☑	☑
title	title	String	☑	☑	☑	☑
postedDate	postedDate	Date	☐	☑	☑	☑
insertionDate	insertionDate	Date	☐	☑	☑	☑
parentURL	parentURL	String	☐	☑	☑	☑
source	source	String	☑	☑	☑	☑
id	id	Integer	☐	☑	☑	☑
url	url	Integer	☐	☑	☑	☑

- 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

Step 1: Ontology Model Creation

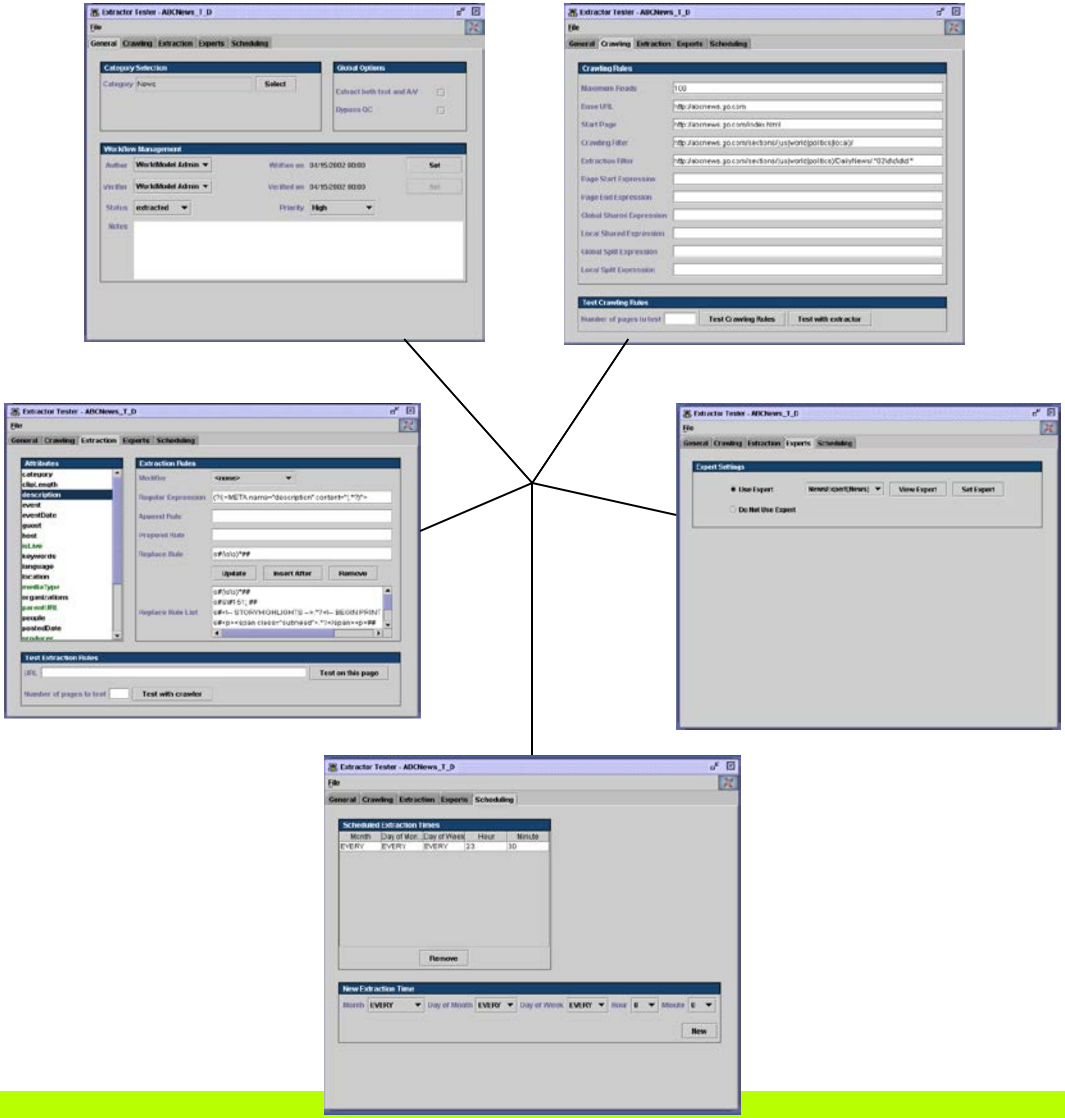
Create an Ontology Model using Semagix Freedom Toolkit GUIs (Cont.)



- 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>**

Step 2: Knowledge Agent Creation

Create and configure Knowledge Agents to populate the Ontology



Extractor Tester - ABCNews_T_D

Category Selection: Category: News

Global Options: Extract both text and A/V ☐ Bypass QC ☐

Workflow Management: Author: WorldModel Admin; Written on: 04/15/2002 00:00; Verifier: WorldModel Admin; Verified on: 04/15/2002 00:00; Status: extracted; Priority: High; Notes:

Crawling Rules:
Maximum Reads: 100
Base URL: http://abcnews.go.com
Start Page: http://abcnews.go.com/index.html
Crawling Filter: http://abcnews.go.com/sections/(us|world|politics|local)/
Extraction Filter: http://abcnews.go.com/sections/(us|world|politics)/DailyNews/.*02\d\d\d\d.*
Page Start Expression:
Page End Expression:
Global Shared Expression:
Local Shared Expression:
Global Split Expression:
Local Split Expression:

Test Crawling Rules: Number of pages to test:

Attributes: category, clipLength, description, event, eventDate, guest, host, isLive, keywords, language, location, mediaType, organizations, parentURB, people, postedDate, producer

Extraction Rules:
Modifier: <none>
Regular Expression: (?(<=META name="description" content=")(.*?)"=
Append Rule:
Prepend Rule:
Replace Rule: s#(\s\s)*##
Replace Rule List:
s#(\s\s)*##
s#&#151; ##
s#<!-- STORYHIGHLIGHTS -->.*?<!-- BEGIN PRINT
s#<p><span class="subhead">.*?</span><p>##

Test Extraction Rules: URL: Number of pages to test:

Expert Settings: ● Use Expert NewsExpert(News) ○ Do Not Use Expert

Scheduled Extraction Times

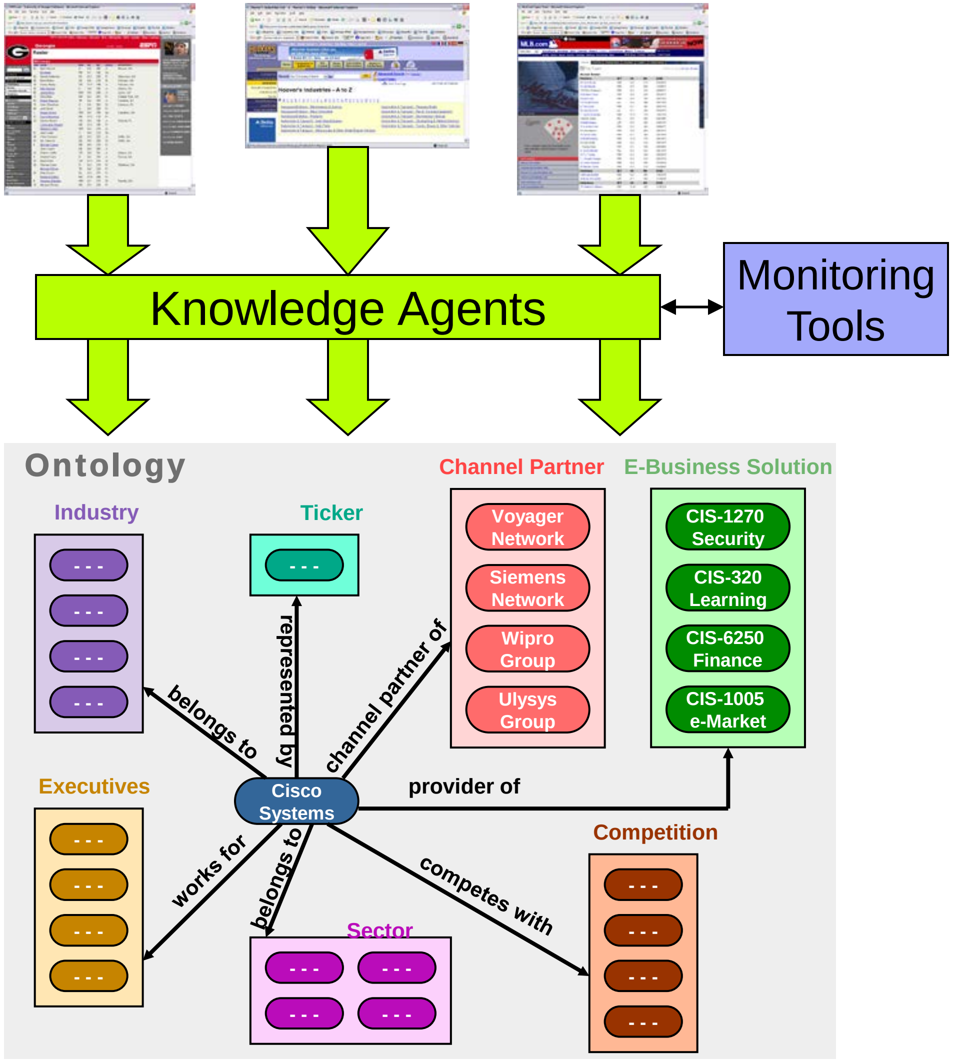
Month	Day of Mon	Day of Week	Hour	Minute
EVERY	EVERY	EVERY	23	30

New Extraction Time: Month: EVERY; Day of Month: EVERY; Day of Week: EVERY; Hour: 0; Minute: 0

- 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 <u>automatically</u>) 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



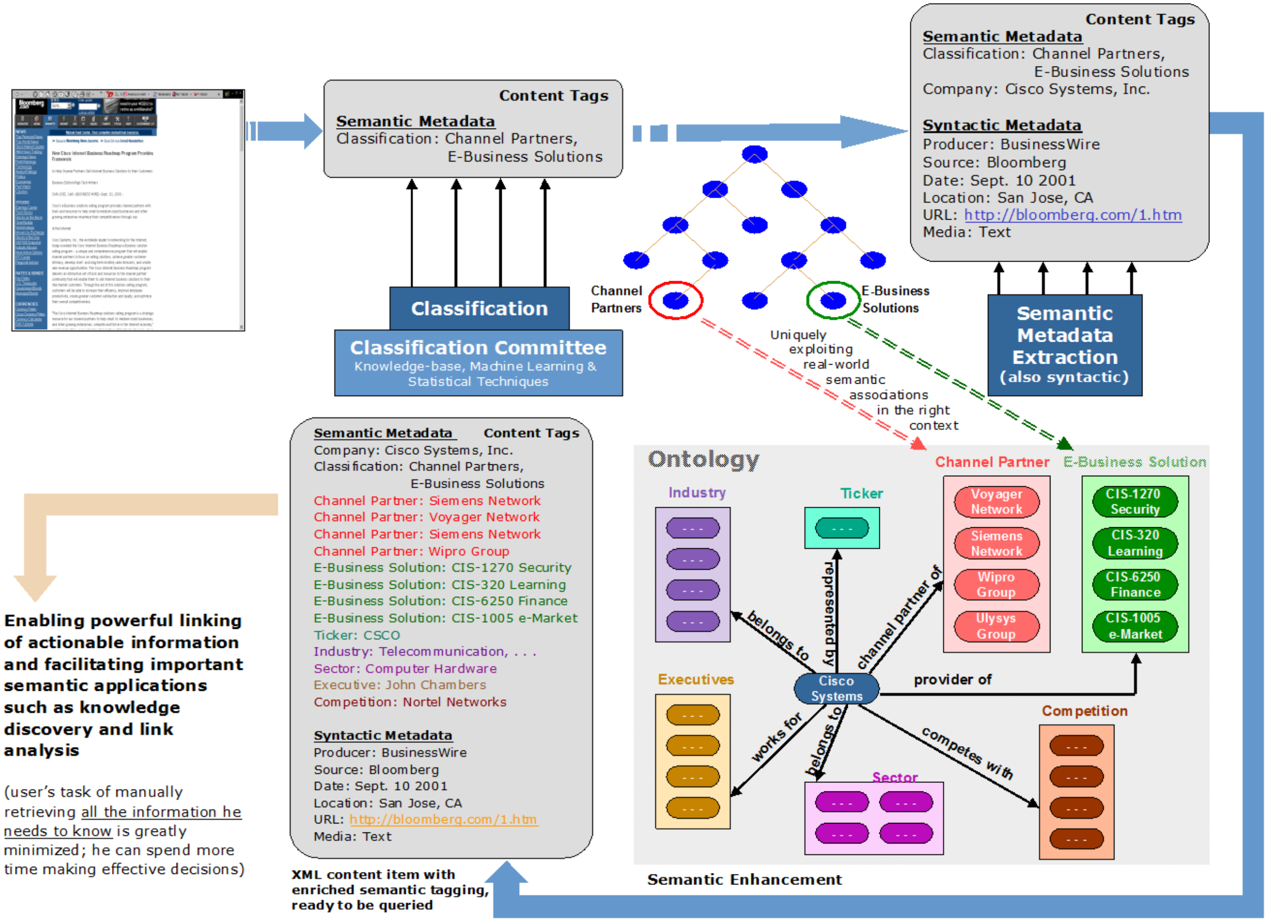
- Automatic aggregation of knowledge at pre-defined intervals fo time
- Supplemented by easy-to-use monitoring tools
- Knowledge Agents extract and organize relevant knowledge into the Ontology, based on the Ontology Model
  - Tools for disambiguation and cleaning
- The Ontology is constantly growing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



Content Tags
Semantic Metadata
Classification: Channel Partners,
E-Business Solutions

Content Tags
Semantic Metadata
Classification: Channel Partners,
E-Business Solutions
Company: Cisco Systems, Inc.

Syntactic Metadata
Producer: BusinessWire
Source: Bloomberg
Date: Sept. 10 2001
Location: San Jose, CA
URL: http://bloomberg.com/1.htm
Media: Text

Semantic Metadata Content Tags
Company: Cisco Systems, Inc.
Classification: Channel Partners,
E-Business Solutions
Channel Partner: Siemens Network
Channel Partner: Voyager Network
Channel Partner: Siemens Network
Channel Partner: Wipro Group
E-Business Solution: CIS-1270 Security
E-Business Solution: CIS-320 Learning
E-Business Solution: CIS-6250 Finance
E-Business Solution: CIS-1005 e-Market
Ticker: CSCO
Industry: Telecommunication, . . .
Sector: Computer Hardware
Executive: John Chambers
Competition: Nortel Networks

Syntactic Metadata
Producer: BusinessWire
Source: Bloomberg
Date: Sept. 10 2001
Location: San Jose, CA
URL: http://bloomberg.com/1.htm
Media: Text

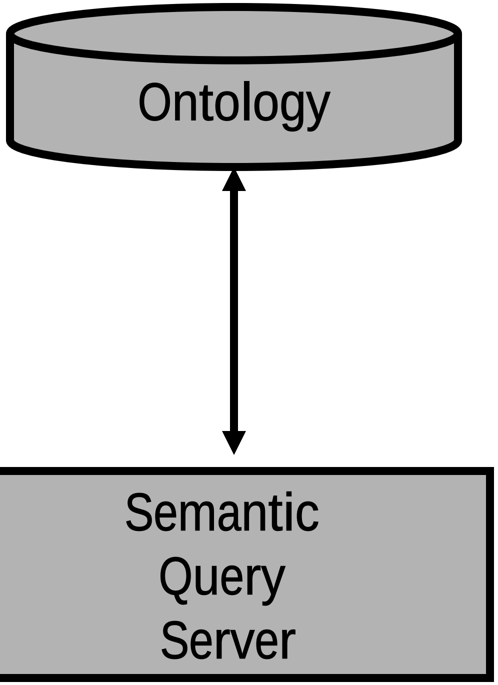
XML content item with enriched semantic tagging, ready to be queried

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

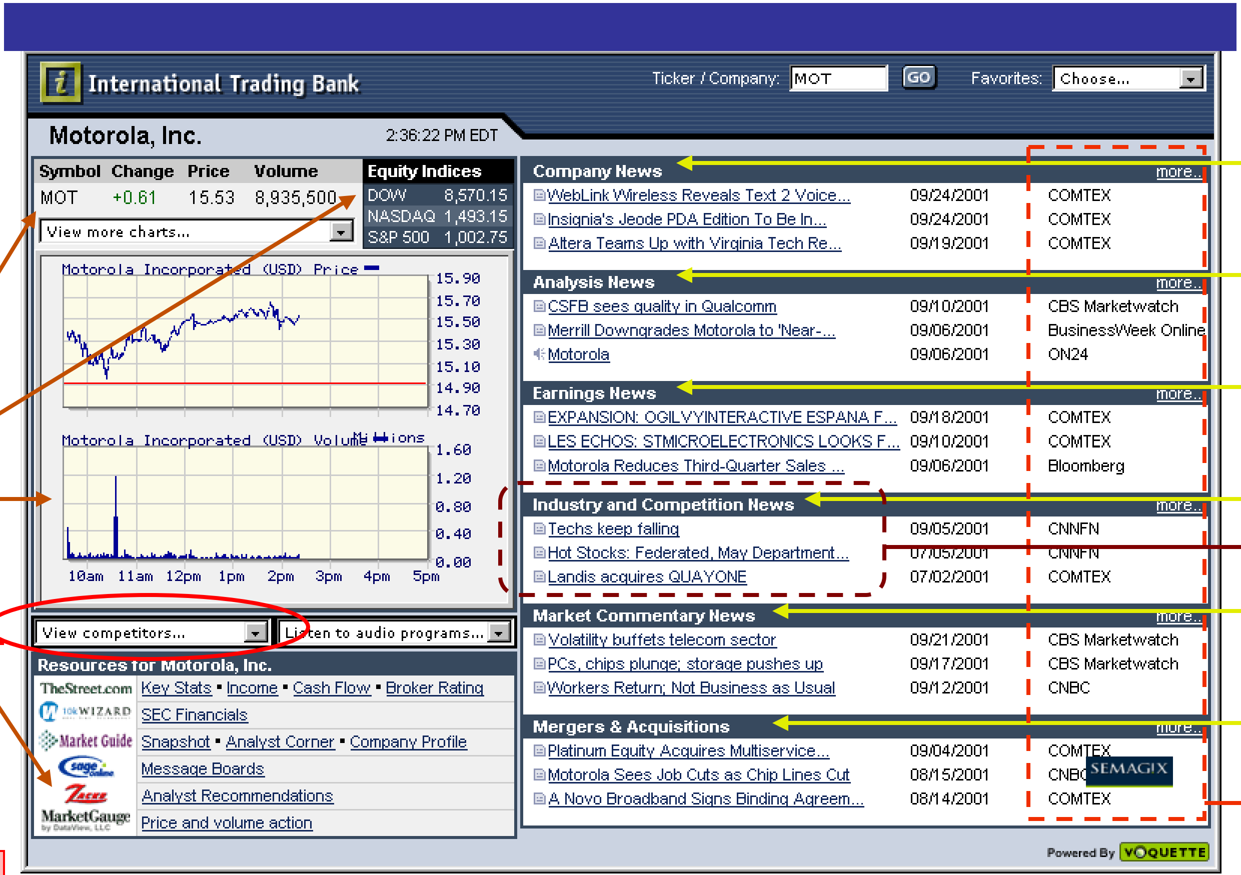
Ontology-based Semagix solutions

- **Equity Analysis Workbench**
  - Heterogeneous internal and 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

Semantic Application Example – Analyst Workbench

Automatic 3rd party content integration

Competitive research inferred automatically



International Trading Bank

Ticker / Company: MOT GO Favorites: Choose...

Motorola, Inc. 2:36:22 PM EDT

Symbol	Change	Price	Volume
MOT	+0.61	15.53	8,935,500

Equity Indices	
DOW	8,570.15
NASDAQ	1,493.15
S&P 500	1,002.75

Resources for Motorola, Inc.

TheStreet.com	Key Stats ▪ Income ▪ Cash Flow ▪ Broker Rating
10kWIZARD	SEC Financials
Market Guide	Snapshot ▪ Analyst Corner ▪ Company Profile
sage online	Message Boards
Zacks	Analyst Recommendations
MarketGauge by DataView, LLC	Price and volume action

Company News		more...
WebLink Wireless Reveals Text 2 Voice...	09/24/2001	COMTEX
Insignia's Jeode PDA Edition To Be In...	09/24/2001	COMTEX
Altera Teams Up with Virginia Tech Re...	09/19/2001	COMTEX
Analysis News		more...
CSFB sees quality in Qualcomm	09/10/2001	CBS Marketwatch
Merrill Downgrades Motorola to 'Near-...	09/06/2001	BusinessWeek Online
Motorola	09/06/2001	ON24
Earnings News		more...
EXPANSION: OGILVYINTERACTIVE ESPANA F...	09/18/2001	COMTEX
LES ECHOS: STMICROELECTRONICS LOOKS F...	09/10/2001	COMTEX
Motorola Reduces Third-Quarter Sales ...	09/06/2001	Bloomberg
Industry and Competition News		more...
Techs keep falling	09/05/2001	CNNFN
Hot Stocks: Federated, May Department...	07/05/2001	CNNFN
Landis acquires QUAYONE	07/02/2001	COMTEX
Market Commentary News		more...
Volatility buffets telecom sector	09/21/2001	CBS Marketwatch
PCs, chips plunge; storage pushes up	09/17/2001	CBS Marketwatch
Workers Return; Not Business as Usual	09/12/2001	CNBC
Mergers & Acquisitions		more...
Platinum Equity Acquires Multiservice...	09/04/2001	COMTEX
Motorola Sees Job Cuts as Chip Lines Cut	08/15/2001	CNBC
A Novo Broadband Signs Binding Agreem...	08/14/2001	COMTEX

Powered By VOQUETTE

SEMAGIX

Focused relevant content organized by topic (*semantic categorization*)

Related relevant content not explicitly asked for (semantic associations)

Automatic Content Aggregation from multiple content providers and feeds

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

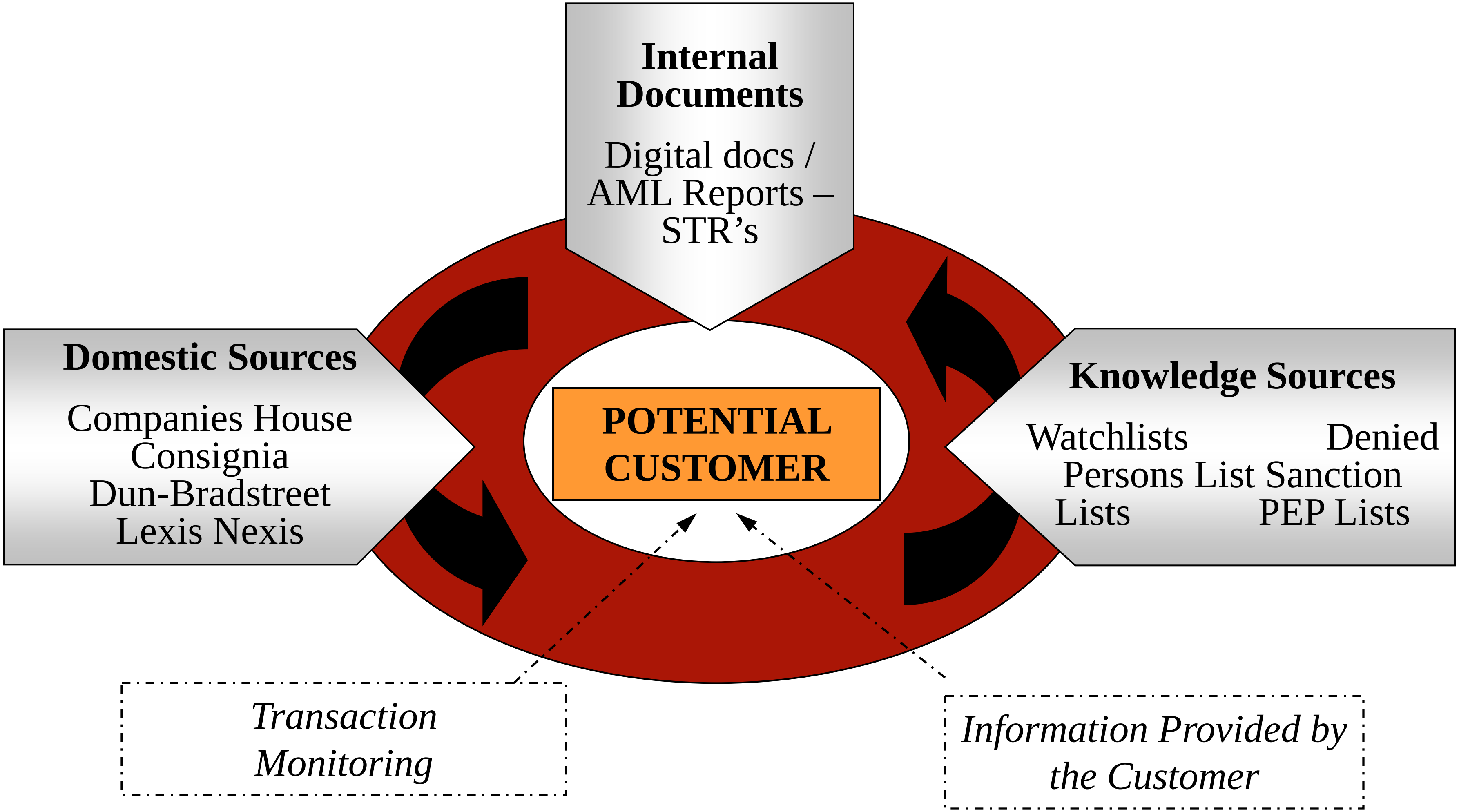
- Difficulty to link different pieces of information – **reduced effectiveness**
- Checks are sequential and resource intensive - **Increase process time and cost**
- Duplication of content – **increased subscription cost**

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’



Internal Documents

Digital docs / AML Reports – STR’s

Domestic Sources

Companies House
Consignia
Dun-Bradstreet
Lexis Nexis

POTENTIAL CUSTOMER

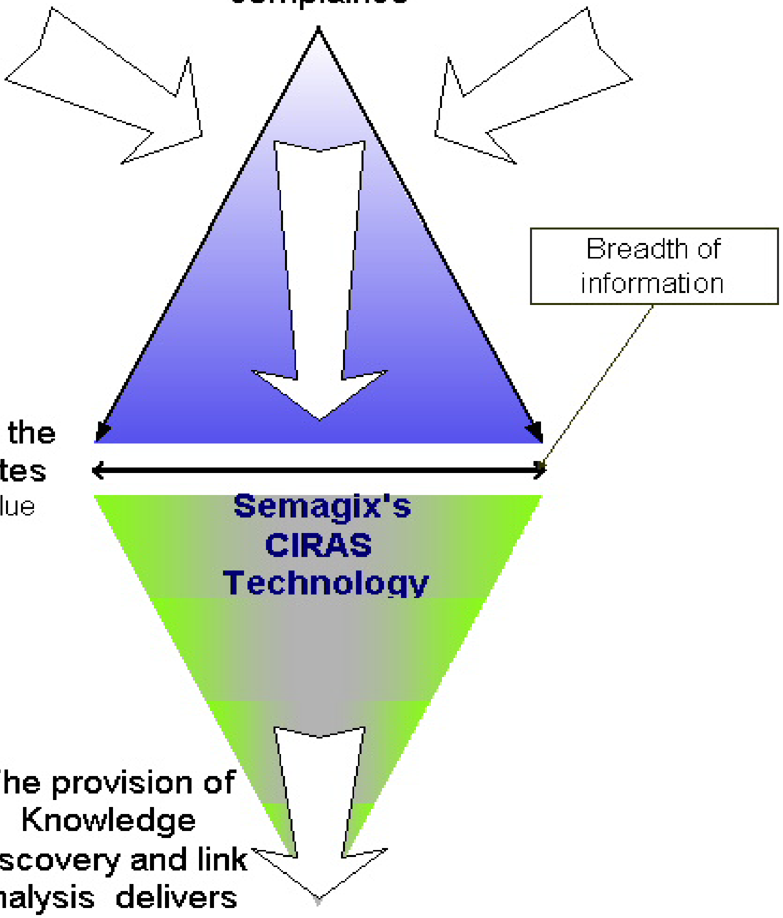
Knowledge Sources

Watchlists Denied Persons List Sanction Lists PEP Lists

Transaction Monitoring

Information Provided by the Customer

What vs. Why



Increased volume of information to meet current complaince

Breadth of information

This helps show the **WHAT** but creates more noise than value

Semagix's CIRAS Technology

The provision of Knowledge Discovery and link analysis delivers the **WHY** aspect

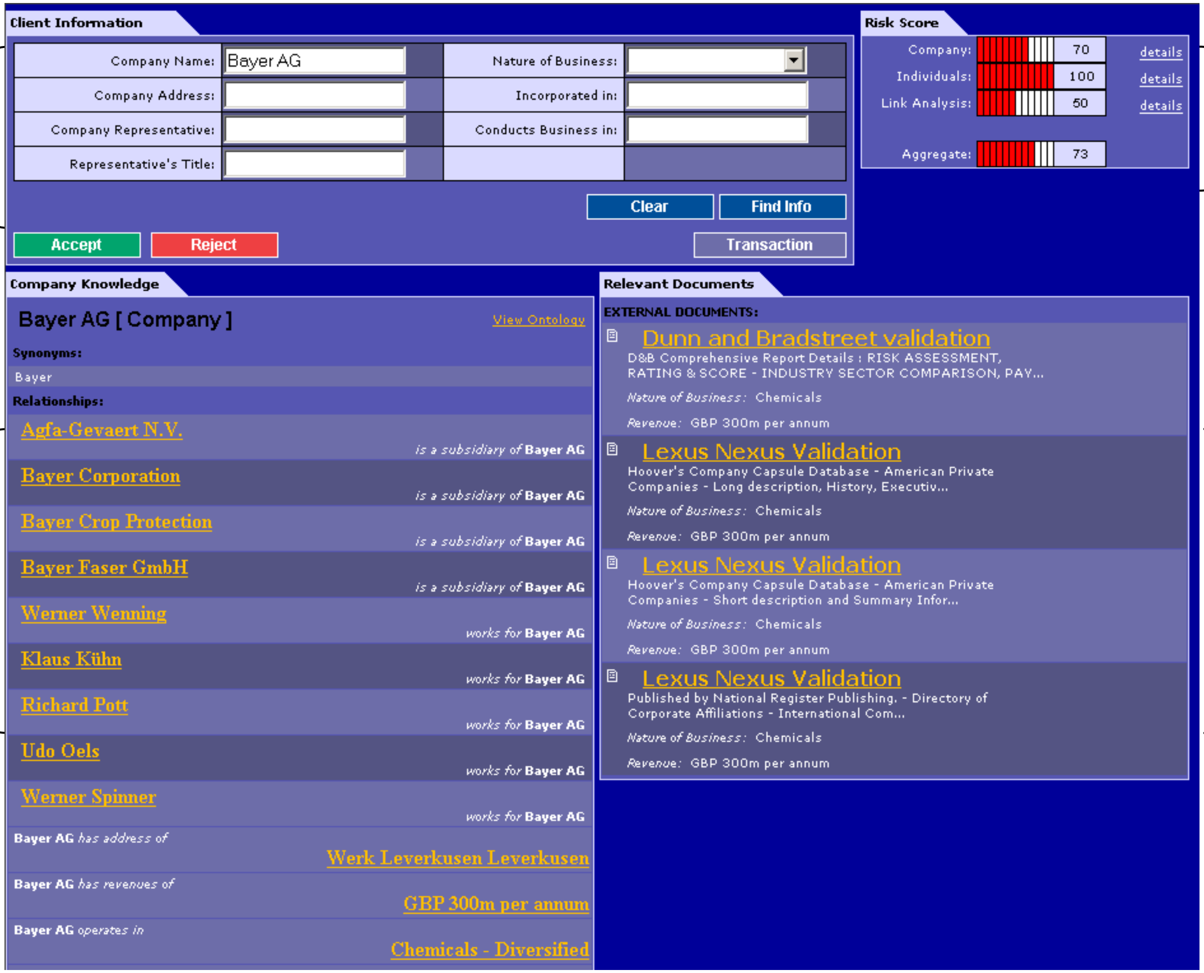
What are the benefits

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

Customer Application Information:

Integration of structured information gathered during the account opening process



Company Name:	Bayer AG	Nature of Business:	
Company Address:		Incorporated in:	
Company Representative:		Conducts Business in:	
Representative's Title:			

Company:	70	details
Individuals:	100	details
Link Analysis:	50	details
Aggregate:	73	

Bayer AG [Company] View Ontology

Synonyms:

Bayer

Relationships:

Agfa-Gevaert N.V.	*is a subsidiary of* **Bayer AG**
Bayer Corporation	*is a subsidiary of* **Bayer AG**
Bayer Crop Protection	*is a subsidiary of* **Bayer AG**
Bayer Faser GmbH	*is a subsidiary of* **Bayer AG**
Werner Wenning	*works for* **Bayer AG**
Klaus Kühn	*works for* **Bayer AG**
Richard Pott	*works for* **Bayer AG**
Udo Oels	*works for* **Bayer AG**
Werner Spinner	*works for* **Bayer AG**
Bayer AG *has address of*	Werk Leverkusen Leverkusen
Bayer AG *has revenues of*	GBP 300m per annum
Bayer AG *operates in*	Chemicals - Diversified

EXTERNAL DOCUMENTS:

Dunn and Bradstreet validation
D&B Comprehensive Report Details : RISK ASSESSMENT, RATING & SCORE - INDUSTRY SECTOR COMPARISON, PAY...
Nature of Business: Chemicals
Revenue: GBP 300m per annum

Lexus Nexus Validation
Hoover's Company Capsule Database - American Private Companies - Long description, History, Executiv...
Nature of Business: Chemicals
Revenue: GBP 300m per annum

Lexus Nexus Validation
Hoover's Company Capsule Database - American Private Companies - Short description and Summary Infor...
Nature of Business: Chemicals
Revenue: GBP 300m per annum

Lexus Nexus Validation
Published by National Register Publishing. - Directory of Corporate Affiliations - International Com...
Nature of Business: Chemicals
Revenue: GBP 300m per annum

Risk Weighting

Relevant Knowledge

Relevant Content

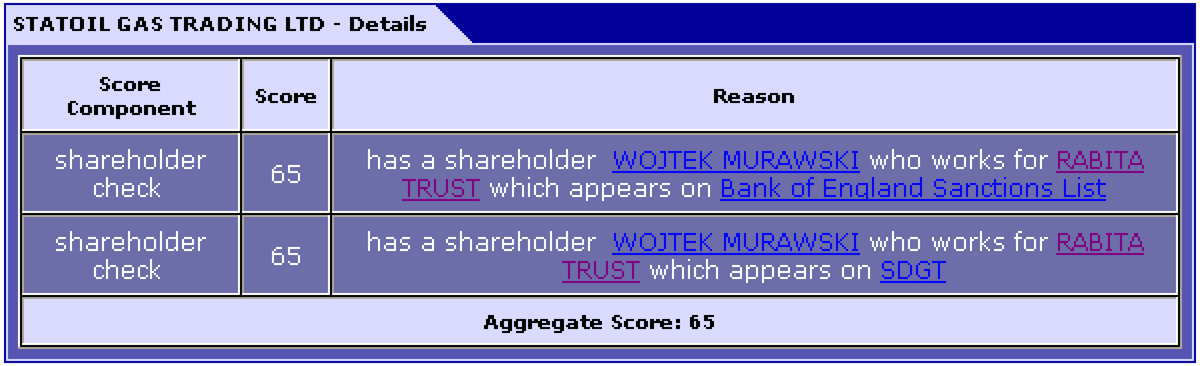
Anti-Money Laundering Ontology

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

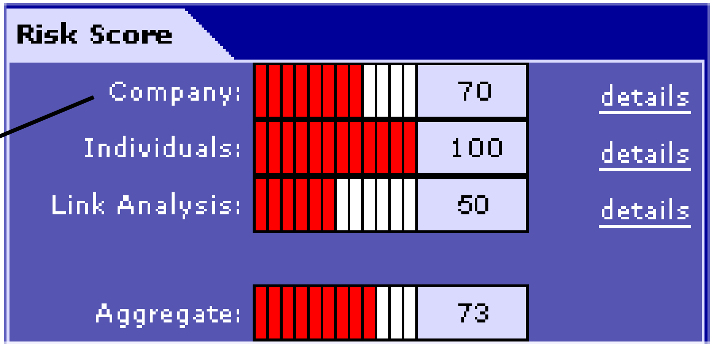
Score Component	Score	Reason
shareholder check	65	has a shareholder WOJTEK MURAWSKI who works for RABITA TRUST which appears on Bank of England Sanctions List
shareholder check	65	has a shareholder WOJTEK MURAWSKI who works for RABITA TRUST which appears on SDGT
Aggregate Score: 65		

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

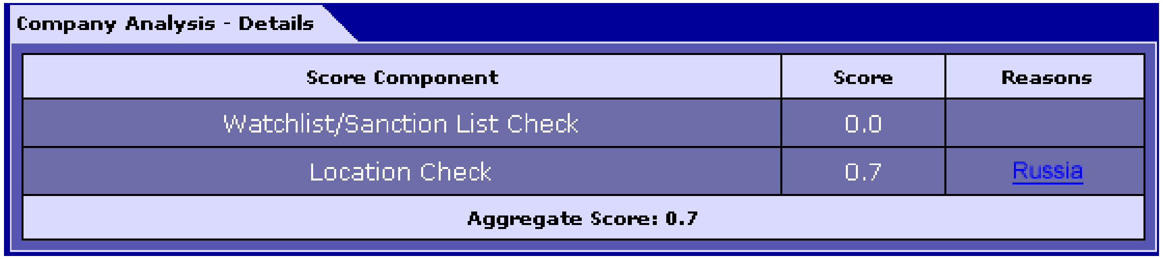
CIRAS’s Components



Risk Score

Company:	▨	70	details
Individuals:	▨	100	details
Link Analysis:	▨	50	details
Aggregate:	▨	73	

1. Company Analysis

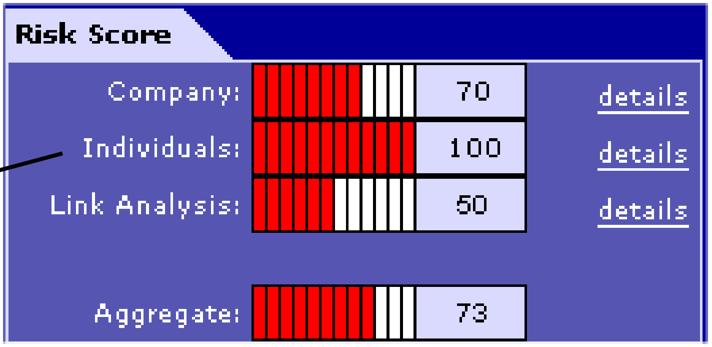


Company Analysis - Details

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

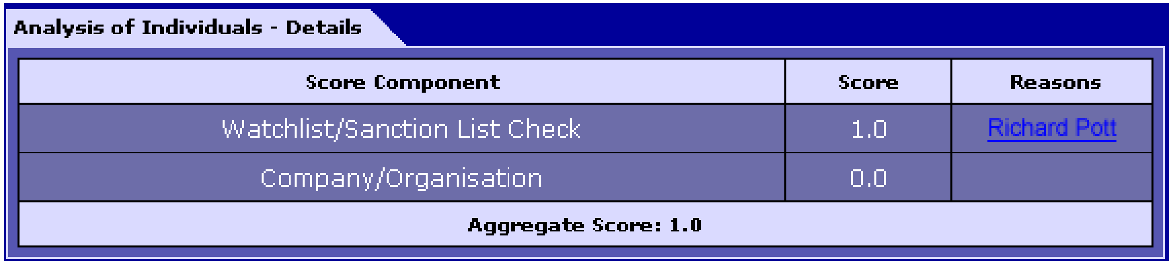
CIRAS’s Components



Risk Score

Company:	70	details
Individuals:	100	details
Link Analysis:	50	details
Aggregate:	73	

2. ID Verification

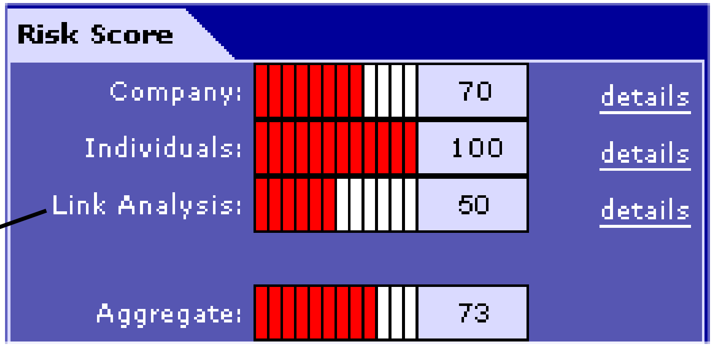


Analysis of Individuals - Details

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

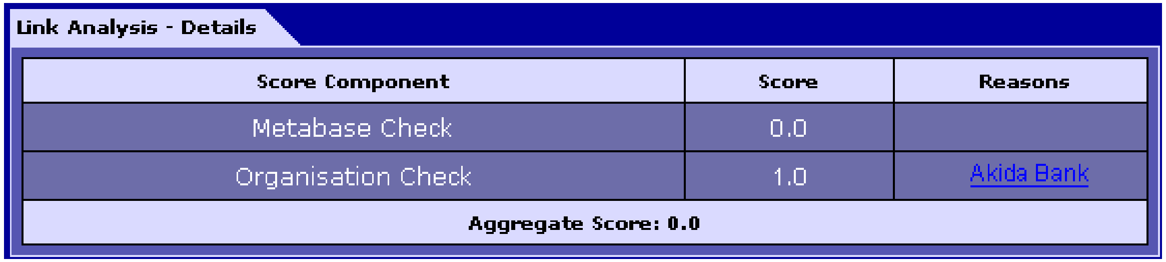
CIRAS's Components



Risk Score

Company:	70	details
Individuals:	100	details
Link Analysis:	50	details
Aggregate:	73	

3. Link Analysis Check



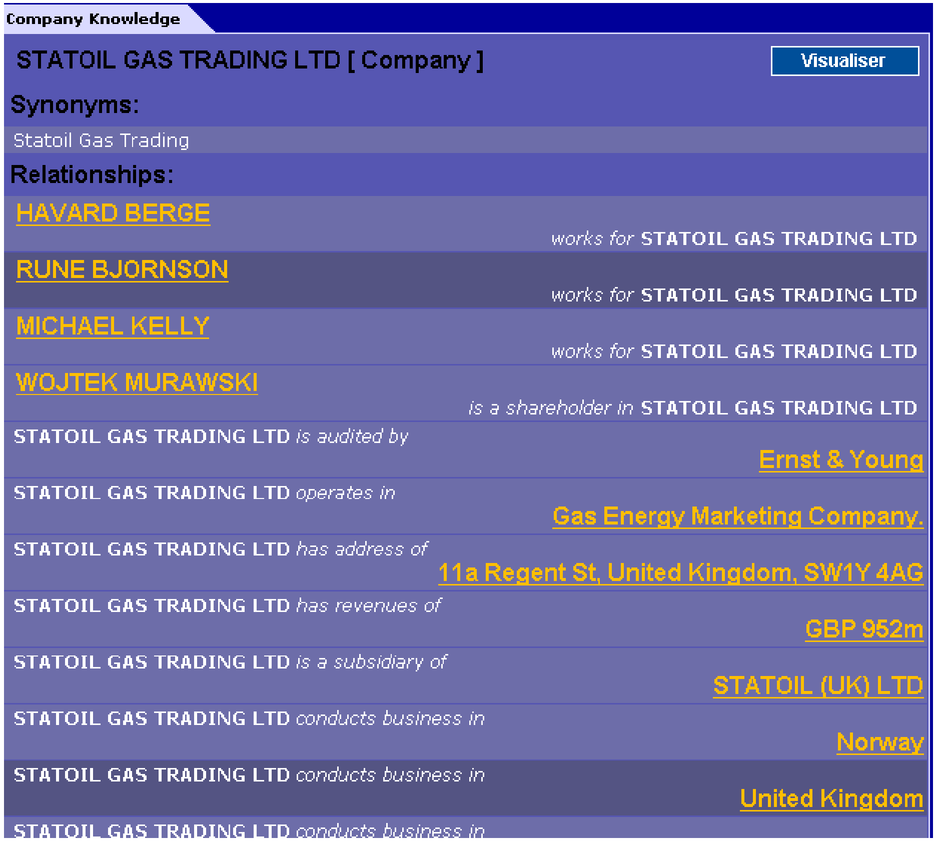
Link Analysis - Details

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



STATOIL GAS TRADING LTD [Company]

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 (UK) LTD

STATOIL GAS TRADING LTD *conducts business in*
Norway

STATOIL GAS TRADING LTD *conducts business in*
United Kingdom

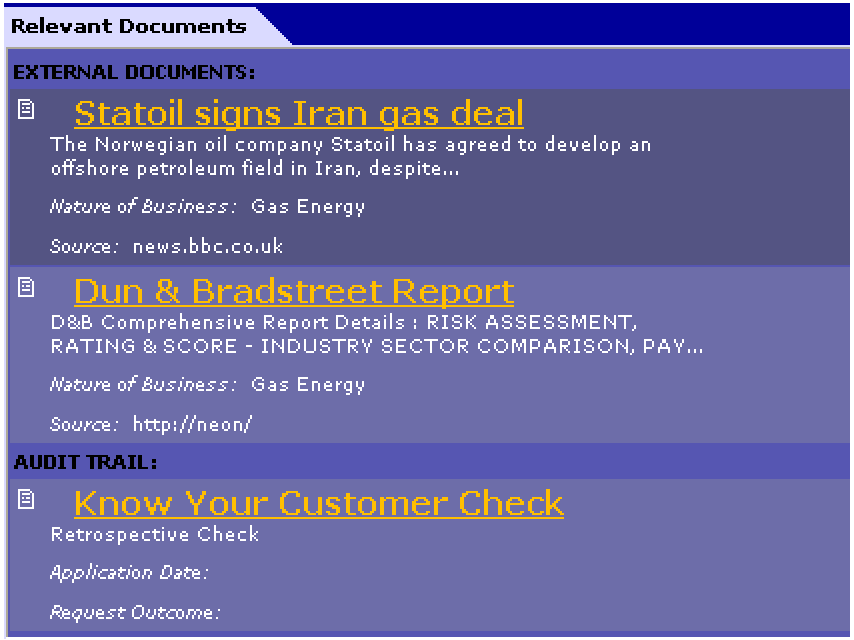
STATOIL GAS TRADING LTD *conducts business in*

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



Relevant Documents

EXTERNAL DOCUMENTS:

Statoil signs Iran gas deal

The Norwegian oil company Statoil has agreed to develop an offshore petroleum field in Iran, despite...

Nature of Business: Gas Energy

Source: news.bbc.co.uk

Dun & Bradstreet Report

D&B Comprehensive Report Details : RISK ASSESSMENT, RATING & SCORE - INDUSTRY SECTOR COMPARISON, PAY...

Nature of Business: Gas Energy

Source: http://neon/

AUDIT TRAIL:

Know Your Customer Check

Retrospective Check

Application Date:

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