

9-23-2003

Semantic Web in Action: Ontology-driven Information Search, Integration and Analysis

Amit P. Sheth

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

Follow this and additional works at: <https://corescholar.libraries.wright.edu/knoesis>



Part of the [Bioinformatics Commons](#), [Communication Technology and New Media Commons](#), [Databases and Information Systems Commons](#), [OS and Networks Commons](#), and the [Science and Technology Studies Commons](#)

Repository Citation

Sheth, A. P. (2003). Semantic Web in Action: Ontology-driven Information Search, Integration and Analysis.
.
<https://corescholar.libraries.wright.edu/knoesis/32>

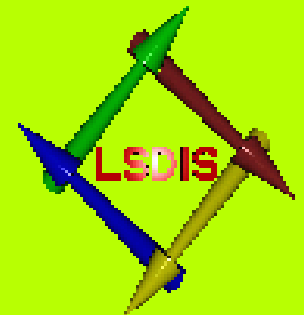
This Presentation is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.

Semantic Web in Action
Ontology-driven information search, integration and analysis
Net Object Days and MATES, Erfurt, September 23, 2003

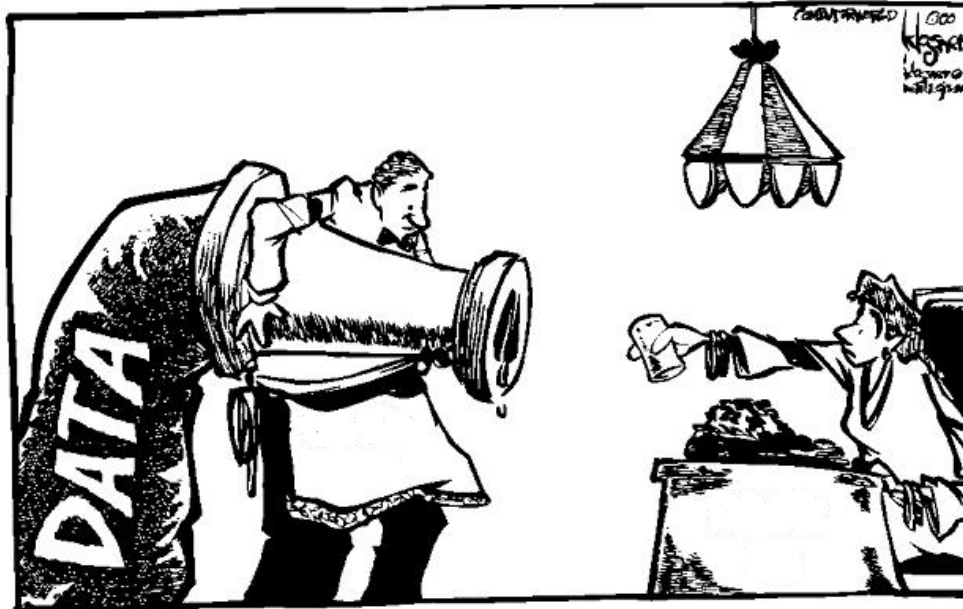
Amit Sheth

Semagix, Inc. and LSDIS Lab, University of Georgia

Talk Abstract



Paradigm shift over time: Syntax -> Semantics



Increasing sophistication in applying semantics

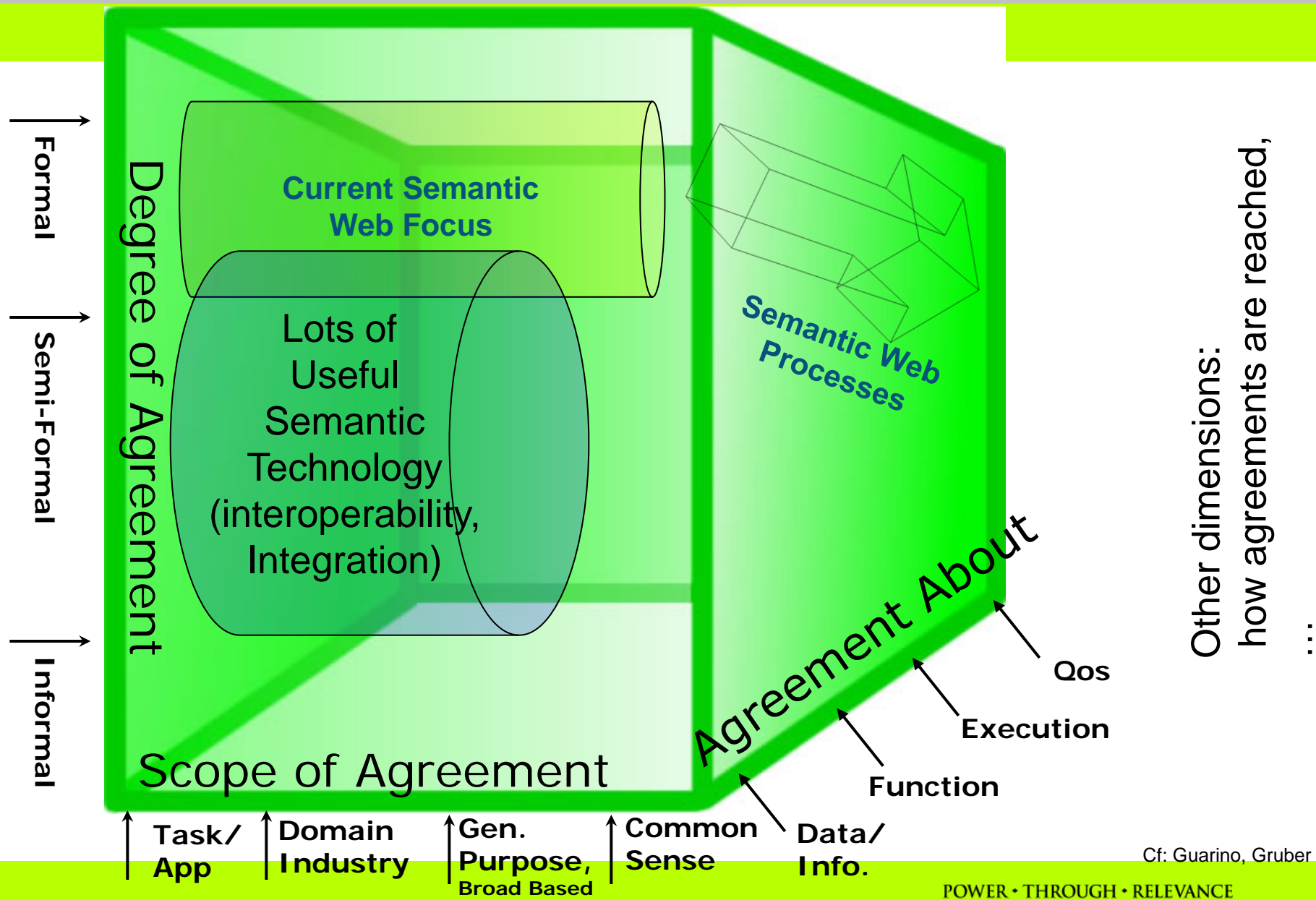
- Relevant Information (Semantic Search & Browsing)
- Semantic Information Interoperability and Integration
- Semantic Correlation/Association, Analysis, Early Warning

Ontology at the heart of the Semantic Web

Ontology provides underpinning for semantic techniques in information systems.

- ◆ A model/representation of the real world (relevant set of interconnected concepts, entities, attributes, relationships, domain vocabulary and factual knowledge).
- ◆ Basis of capturing agreement, and of applying knowledge
- ◆ Enabler for improved information systems functionalities and the Semantic Web

Ontology = Schema (Description) + Knowledge Base (Description Base)
i.e, both T-nodes and A-nodes



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
- ◆ Utilizing semantic metadata and ontology
 - ❖ Semantic search/querying/browsing
 - ❖ Information and application integration - normalization
 - ❖ Analysis/Mining/Discovery - relationships

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

Practical Experiences on Ontology Management today

- ◆ What types of ontologies are needed and developed for semantic applications today?
 - ❖ Is there a typical ontology?
- ◆ How are such ontologies built?
- ◆ Who builds them? How long it takes? How are ontologies maintained?
 - ❖ People (expertise), time, money
- ◆ How large ontologies become (scalability)?
- ◆ How are ontologies used and what are computational issues?

Types of Ontologies (or things close to ontology)

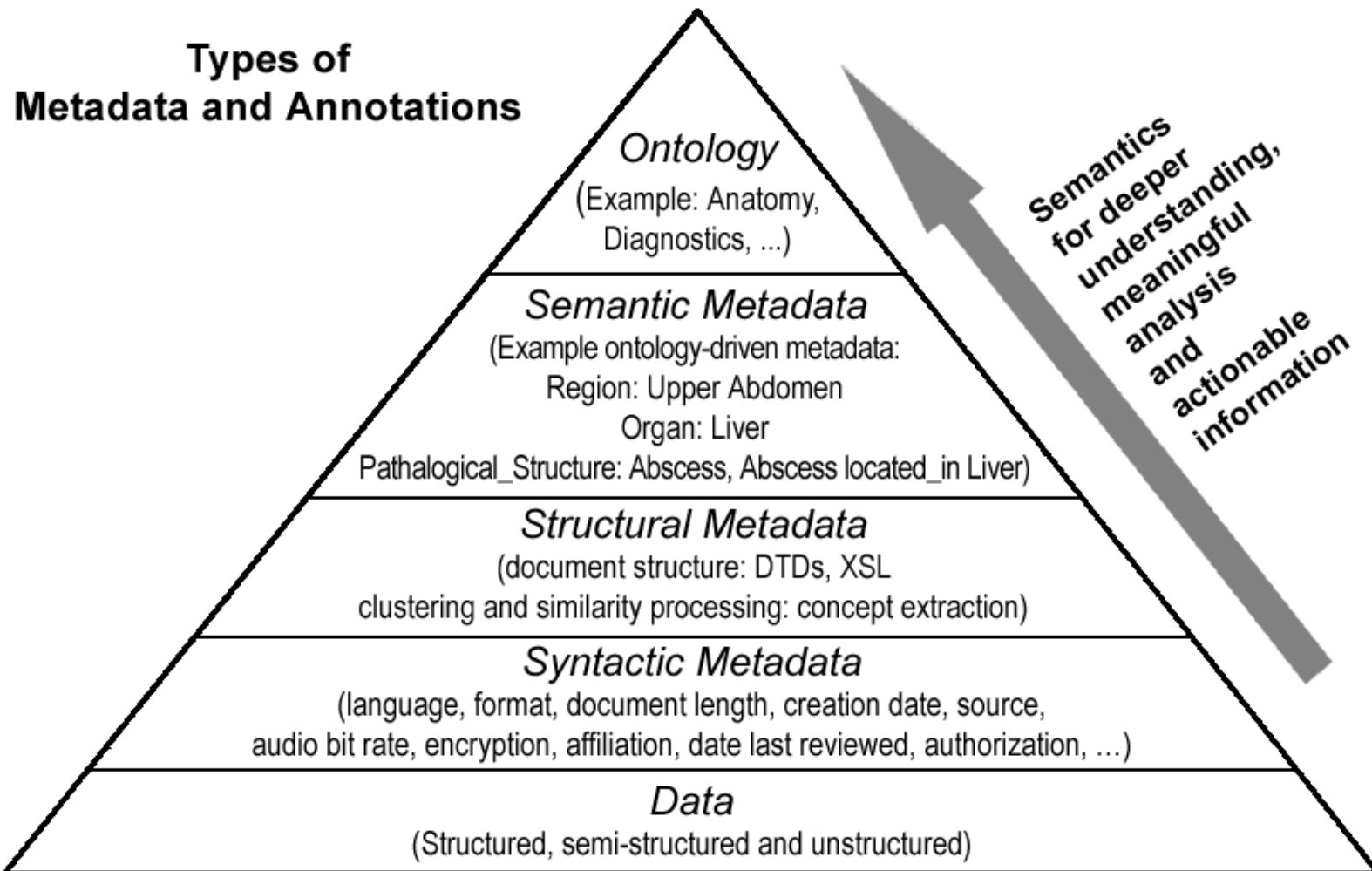
- ◆ Upper ontologies: modeling of time, space, process, etc
- ◆ Broad-based or general purpose ontology/nomenclatures: Cyc, CIRCA ontology (Applied Semantics), *WordNet*
- ◆ Domain-specific or Industry specific ontologies
 - ❖ News: politics, sports, business, entertainment
 - ❖ Financial Market
 - ❖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

Building ontology

- ◆ Three broad approaches:
 - ❖ social process/manual: many years, committees
 - ◆ Based on metadata standard
 - ❖ automatic taxonomy generation (statistical clustering/NLP):
limitation/problems on quality, dependence on corpus, naming
 - ❖ Descriptive component (schema) designed by domain experts; Descriptive base (assertional component, extension) by automated processes

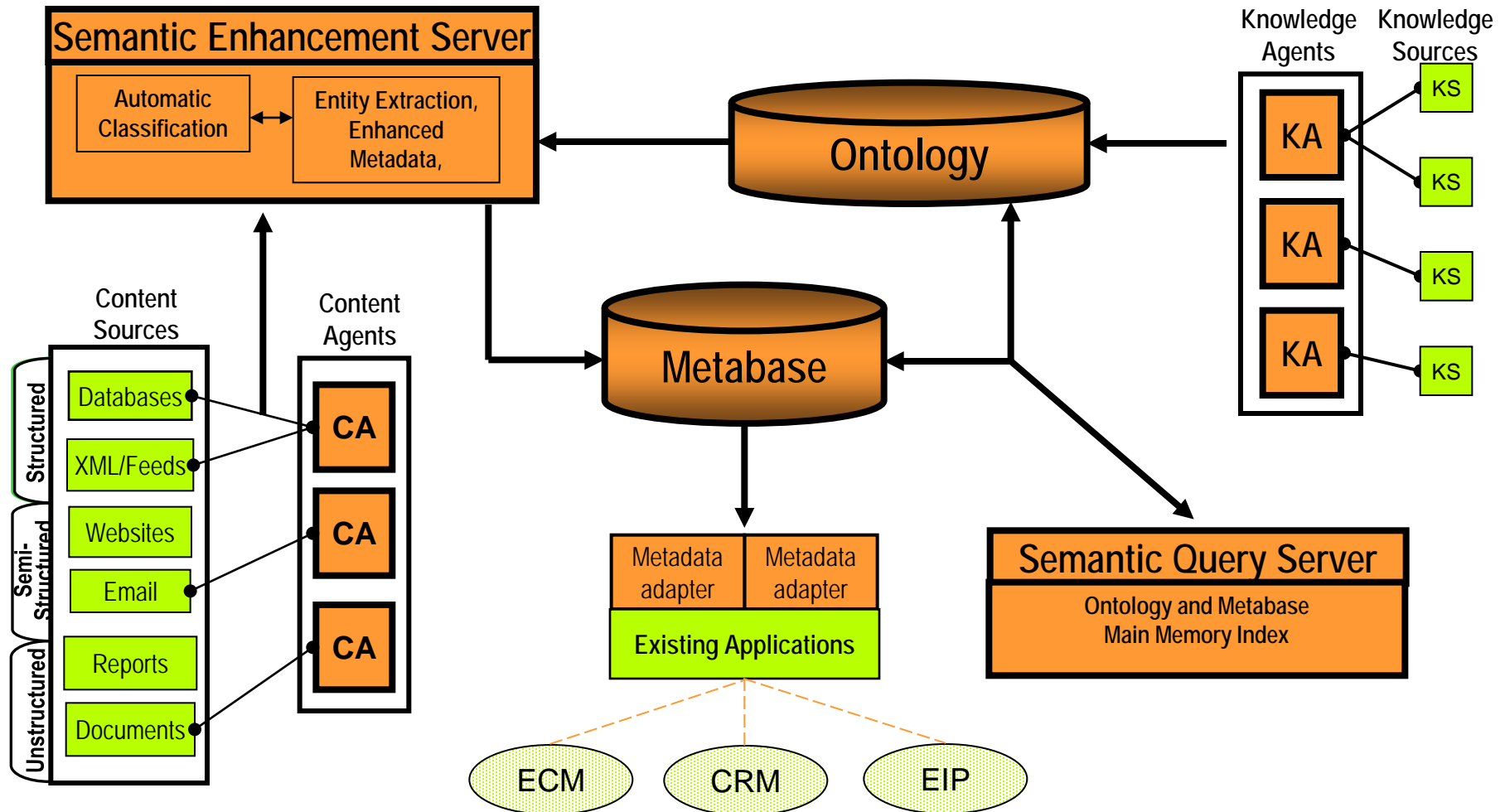
Option 2 is being investigated in an ontology learning system at UGA; Option 3 is currently supported by Semangix Freedom

Metadata and Ontology: Primary Semantic Web enablers



Semagix Freedom Architecture

(a platform for building ontology-driven information system)

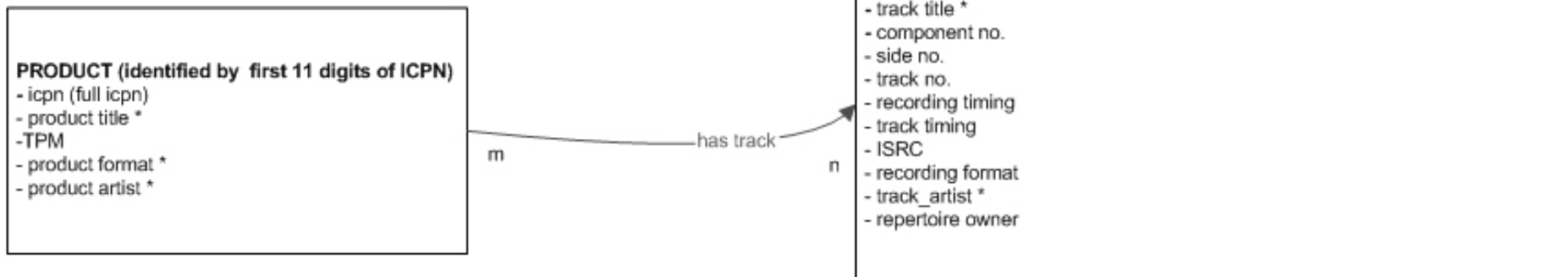


Practical Ontology Development Observation by Semagix

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

Example 1: Ontology with simple schema

- ◆ **Ontology for a customer in Entertainment Industry primarily for repertoire management**
- ◆ **Ontology Schema (Descriptive Component)**
 - ❖ Only few high-level entity classes, primarily **Product** and **Track**
 - ❖ A few attributes for each entity class
 - ❖ Only a few relationship types, e.g.: “*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 (Assertional Component)

- ❖ Description base of 10 to 20 million objects (entity, relationship, attribute instances in ontology)
- ❖ Extracted by Knowledge Agents from 6 disparate databases

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

Technical Challenges Faced

◆ **'Dirty' data**

- ❖ Inconsistent field values
- ❖ Unfilled field values
- ❖ Field values appearing to mean the same, but are different

◆ **Non-normalized Data**

- ❖ Different names to mean the same object (schematic heterogeneity)

◆ **Upper case vs. Lower case text analysis**

◆ **Scoring (for identity resolution) and pre-processing (for normalization) parameters changed frequently by customer, necessitating constant update of algorithms**

Ambiguity
Resolution

◆ **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's main memory system

Effort Involved

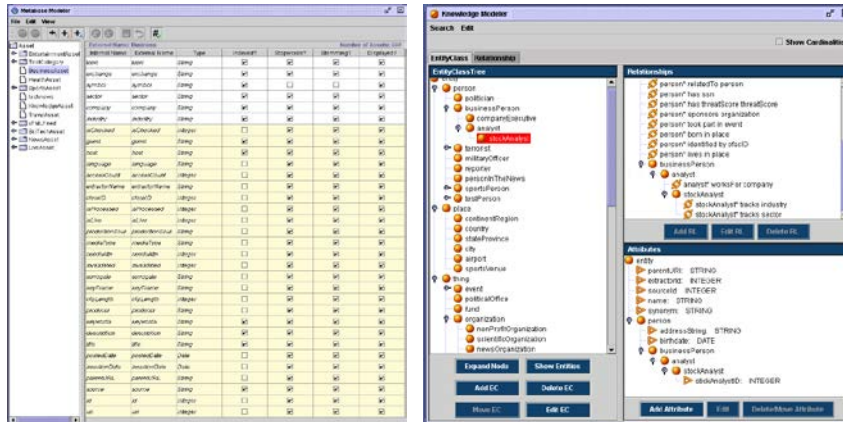
◆ **Ontology Schema Build-Out** (descriptive component)

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

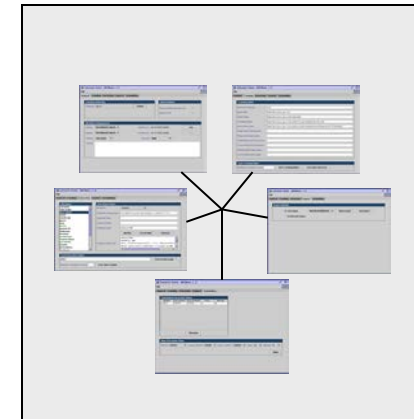
- ❖ Due to iterative decision making process with the multi-national customer, overall finalization of ontology took 3-4 weeks to complete; not complex otherwise
- ❖ **Ontology Population** (assertional component/description base)
- ❖ 6 Knowledge Agents, one for each database; writing agents took about a day
- ❖ Automated extraction using Knowledge Agents took a few days for all the Agents, with a few days of validation

Ontology Creation and Maintenance Process

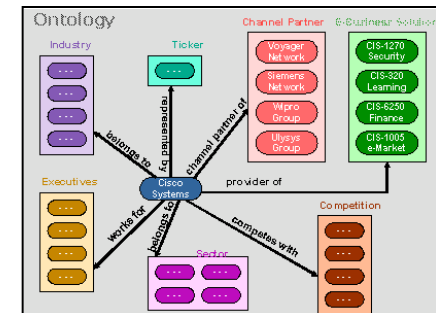
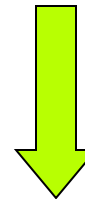
Ontology Creation and Maintenance Steps



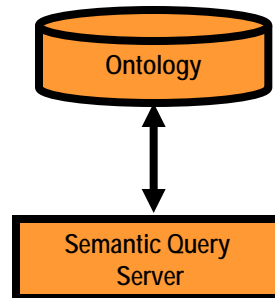
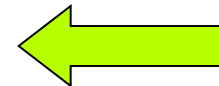
1. Ontology Model Creation (Description)



2. Knowledge Agent Creation



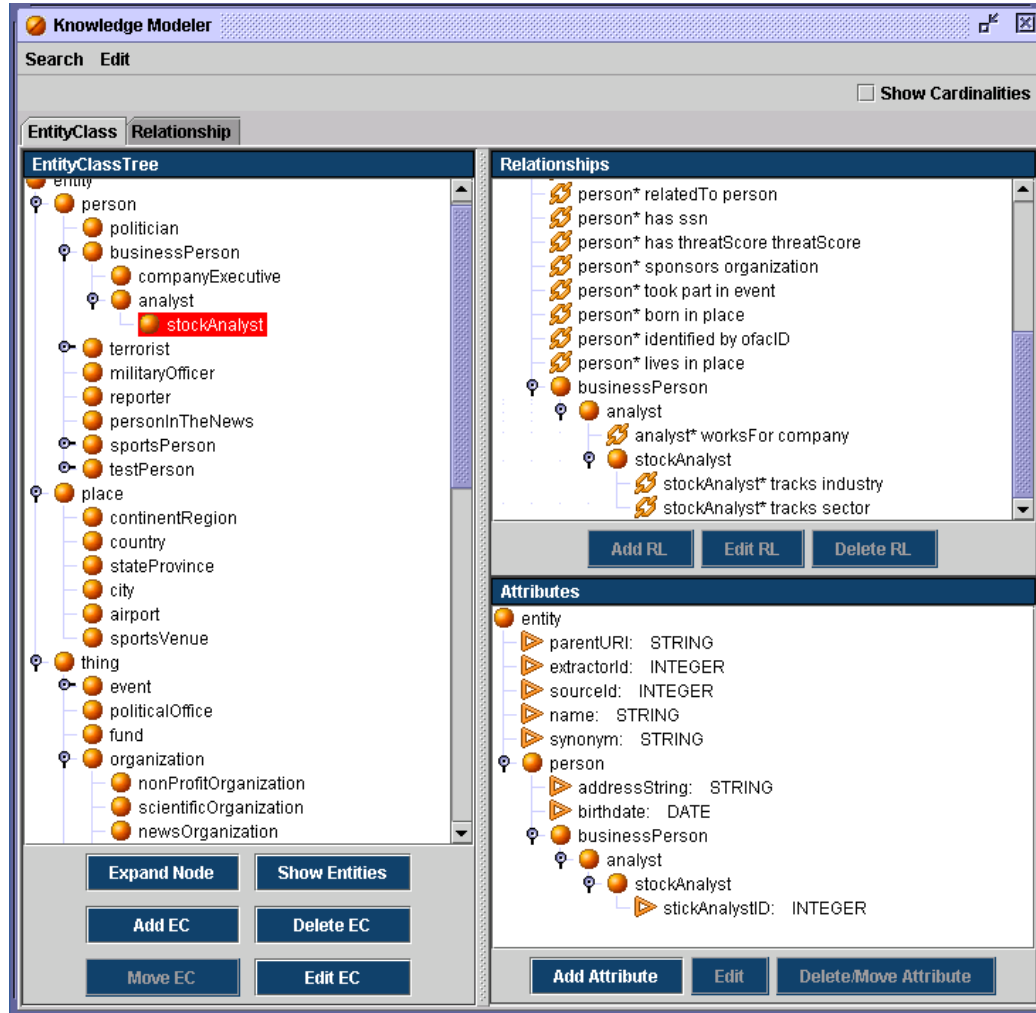
3. Automatic aggregation of Knowledge



4. Querying the Ontology

Step 1: Ontology Model Creation

Create an Ontology Model using Semagix Freedom Toolkit GUIs



- This corresponds to the schema of the description 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 1: Ontology Model Creation

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

Metabase Modeler

File Edit View

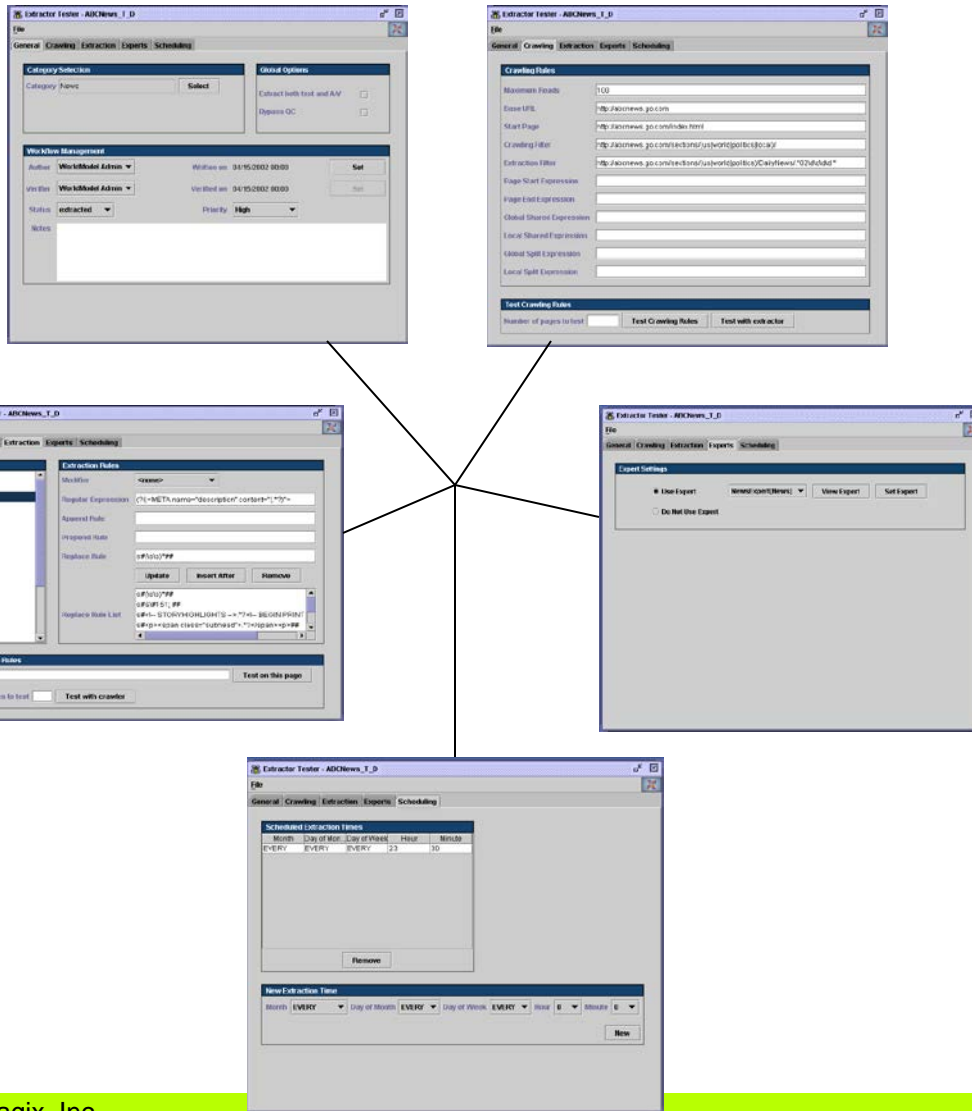
Number of Assets: ###

Internal Name	External Name	Type	Indexed?	Stopwords?	Stemming?	Displayed?
topic	topic	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
exchange	exchange	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
symbol	symbol	String	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
sector	sector	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
company	company	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
industry	industry	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
isChecked	isChecked	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
guest	guest	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
host	host	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
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 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"/>
isLive	isLive	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
productionSour...	productionSour...	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"/>
needsAttn	needsAttn	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
invalidated	invalidated	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
surrogate	surrogate	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"/>
clipLength	clipLength	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"/>

- 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 2: Knowledge Agent Creation (Automation Component)

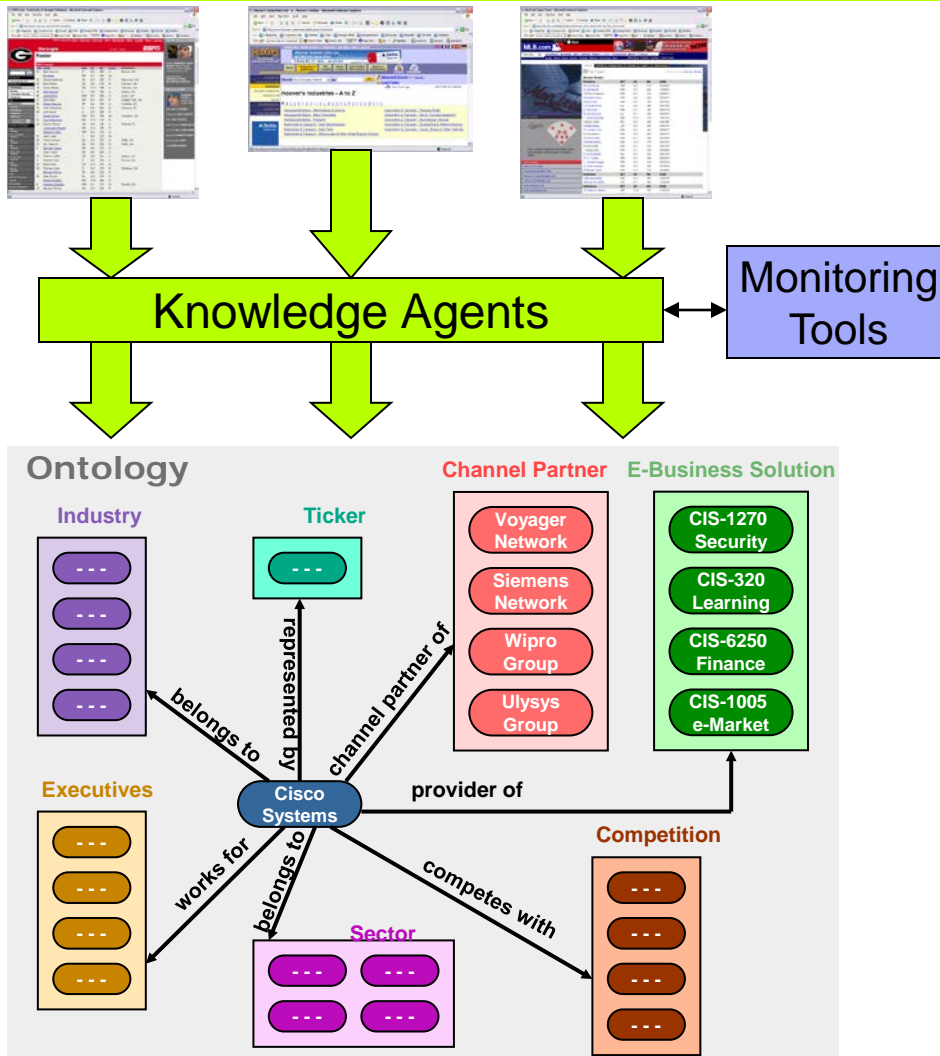
Create and configure Knowledge Agents to populate the Ontology



- 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

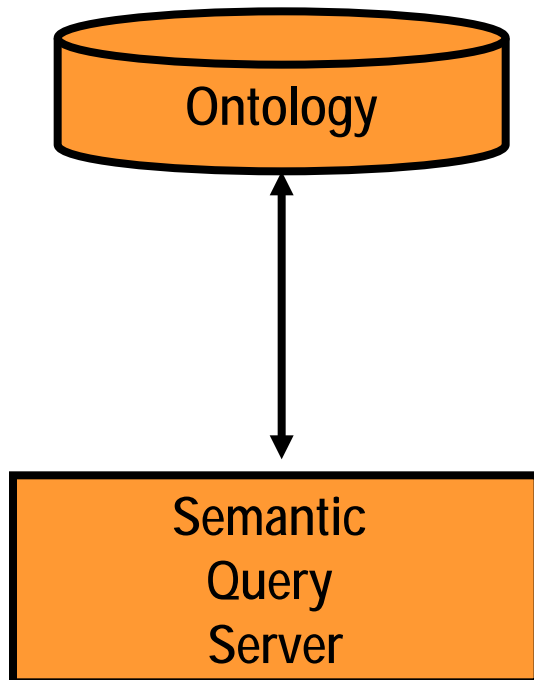
Automatic aggregation of knowledge from knowledge sources



- Automatic aggregation of knowledge at pre-defined intervals of 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

Step 4: Querying the Ontology

Semantic Query Server can now query the Ontology

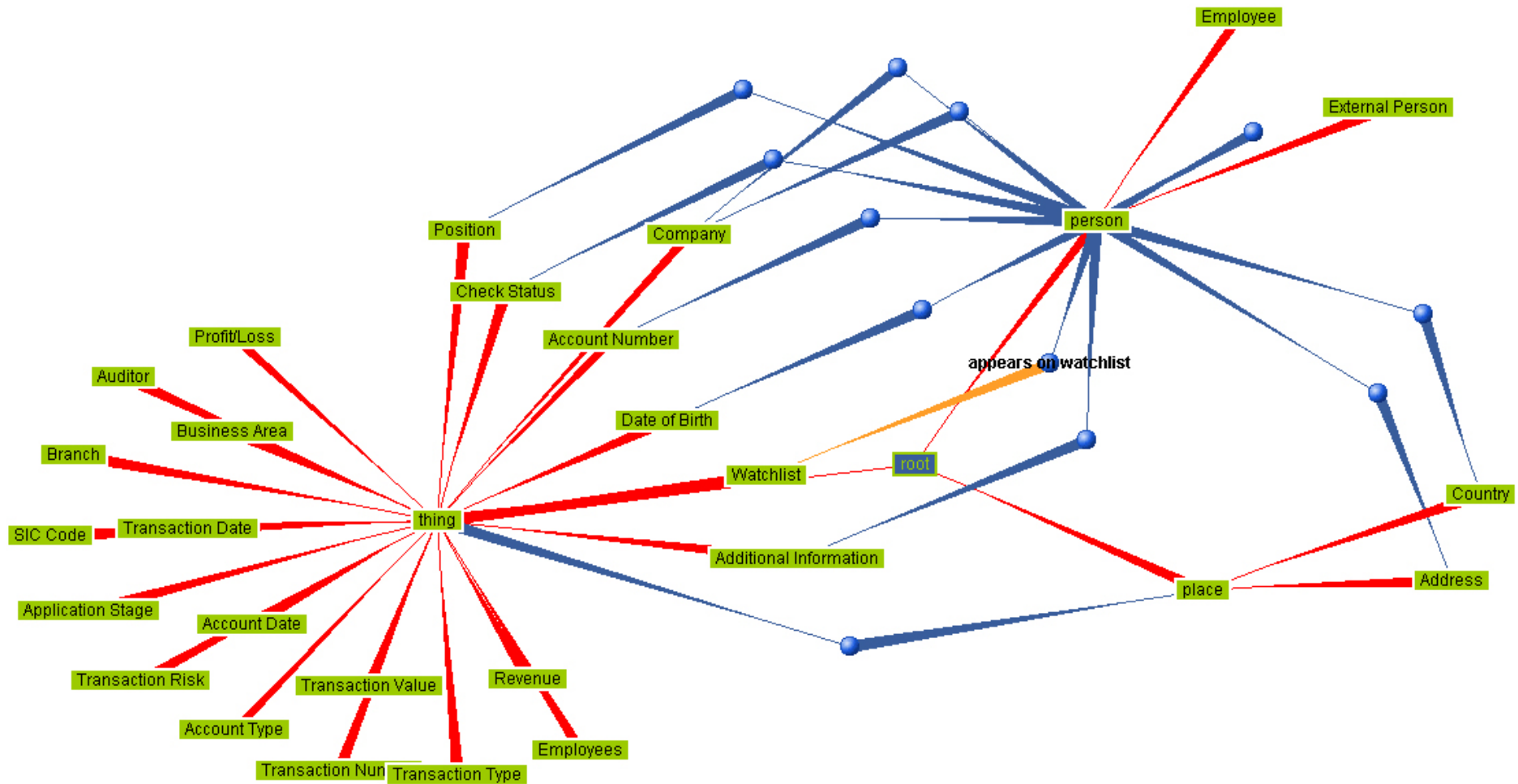


- Incremental indexing
- Distributed indexing
- Knowledge APIs provide a Java, JSP or an HTTP-based interface for querying the Ontology and Metadata

Example2: Ontology with complex schema

- ◆ **Ontology for Anti-money Laundering (AML) application in Financial Industry**
- ◆ **Ontology Schema (Descriptive Component)**
 - ❖ About 50 entity classes
 - ❖ About 100 attribute types
 - ❖ About 60 relationship types between entity classes

AML Ontology Schema (Descriptive Component)



AML (Anti-Money Laundering) Ontology

Ontology Schema (Assertional Component)

- ◆ About 1.5M entities, attributes and relationships
- ◆ 4 primary (licensed or public) sources for knowledge extraction
 - ❖ Dun and Bradstreet
 - ❖ Corporate 192
 - ❖ Companies House
 - ❖ Hoovers

Effort Involved

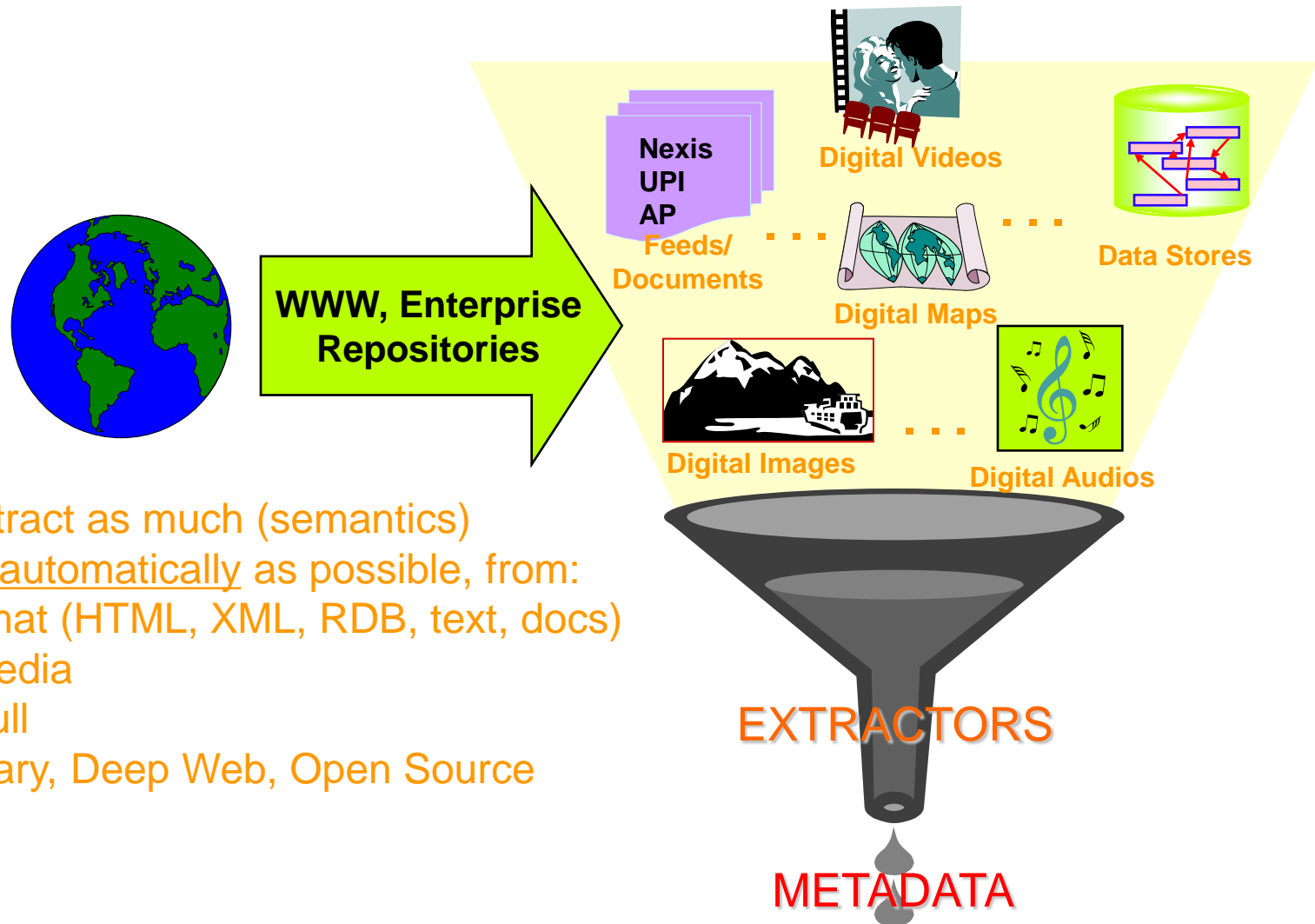
- ◆ Ontology schema design: **less than a week** (with periodic extensions)
- ◆ Automated Ontology population using Knowledge Agents: **a few days**

Technical Challenges Faced

- ◆ **Complex ambiguity resolution at entity extraction time**
- ◆ **Modelling the ontology to capture adequate details of the domain for intended application**
 - ❖ **Ensuring that the risk algorithm (link score analysis) can be implemented with the needed parameters**
- ◆ **Knowledge extraction from sources that needed extended cookie/HTTPS handling**
- ◆ **Adding entities on the fly (dynamic ontology)**

Metadata Extraction from Heterogeneous, Distributed Content:

Metadata extraction from heterogeneous content/data



Create/extract as much (semantics) metadata automatically as possible, from:

- Any format (HTML, XML, RDB, text, docs)
- Many media
- Push, pull
- Proprietary, Deep Web, Open Source

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 [Bobby Bonilla](#).

Ashby agreed to a three-year contract.

Galarraga's contract expired at the end of the season. He had 100 RBIs.

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

Free agents not offered arbitration by their former teams 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:

Classify URL

Select a story
from Virage:

Classification Results

http://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

Locations

Bonilla, Bobby	Sportsperson	Central (1266)
Joyner, Wally	Sportsperson	Atlanta (406)
Kamieniecki, Scott	Sportsperson	
Mulholland, Terry	Sportsperson	
Ashby, Andy	Sportsperson	
Galarraga, Andres	Sportsperson	

Semantic Metadata

Ontology-directed Metadata Extraction (Semi-structured data)

SEMAG!X

Web Page



Extraction
Agent

Enhanced Metadata Asset

A screenshot of a Microsoft Internet Explorer browser window displaying a form titled 'Video Abstract'. The form contains various fields for metadata extraction, including Asset Id, Category, ExtractorName, Keywords, Title, Surrogate, Media Type, URL, Description, Clip Length, Parent URL, Location, and People. The form is populated with data from the BBC News page, such as '31918736' for Asset Id, 'NewsEvent' for Category, and 'BBCWorld' for ExtractorName. The form also includes buttons for 'Previous', 'Next', 'Update', 'Update/Next', 'Delete/Next', and 'Run Experts'.

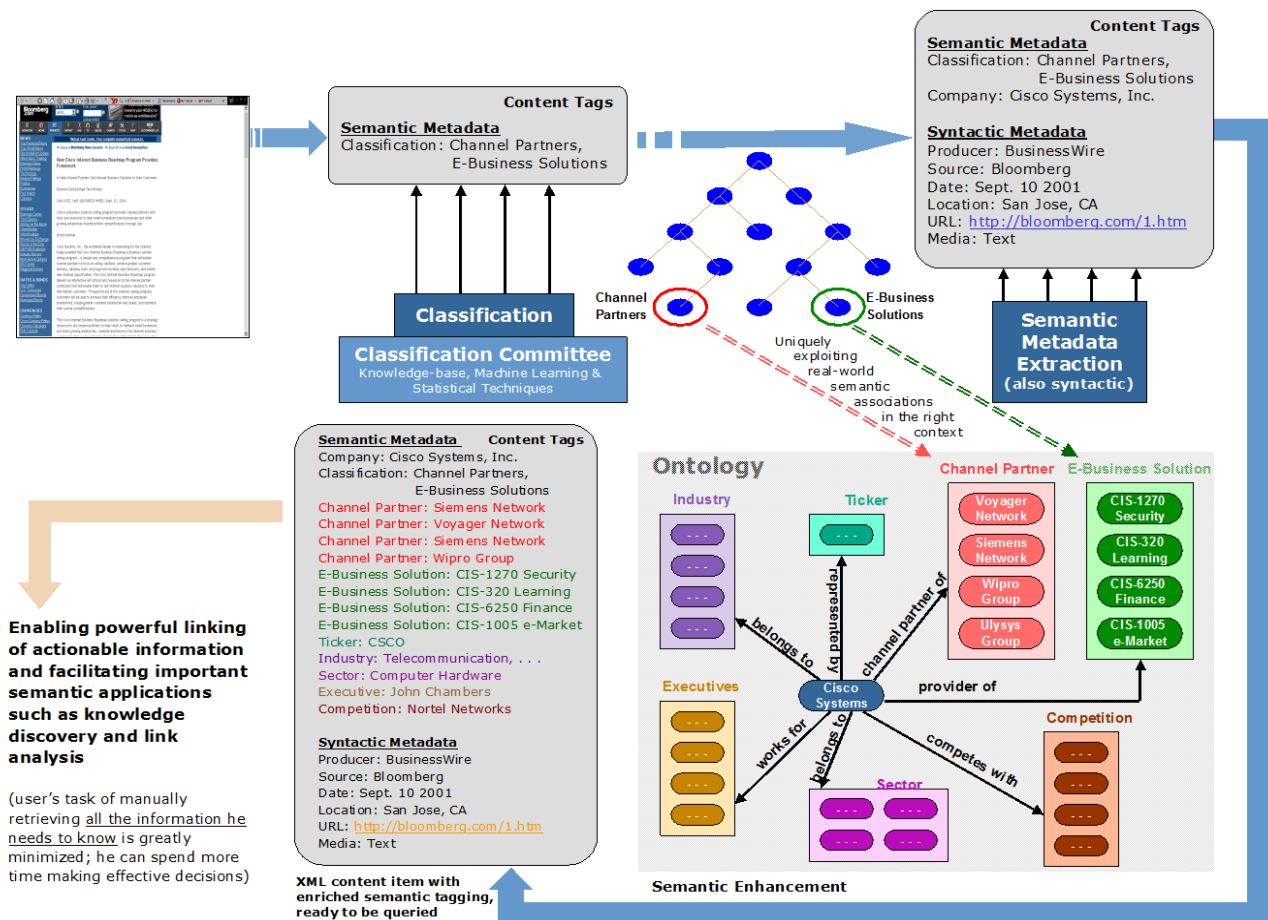
Semantic Enhancement Server

Semantic Enhancement

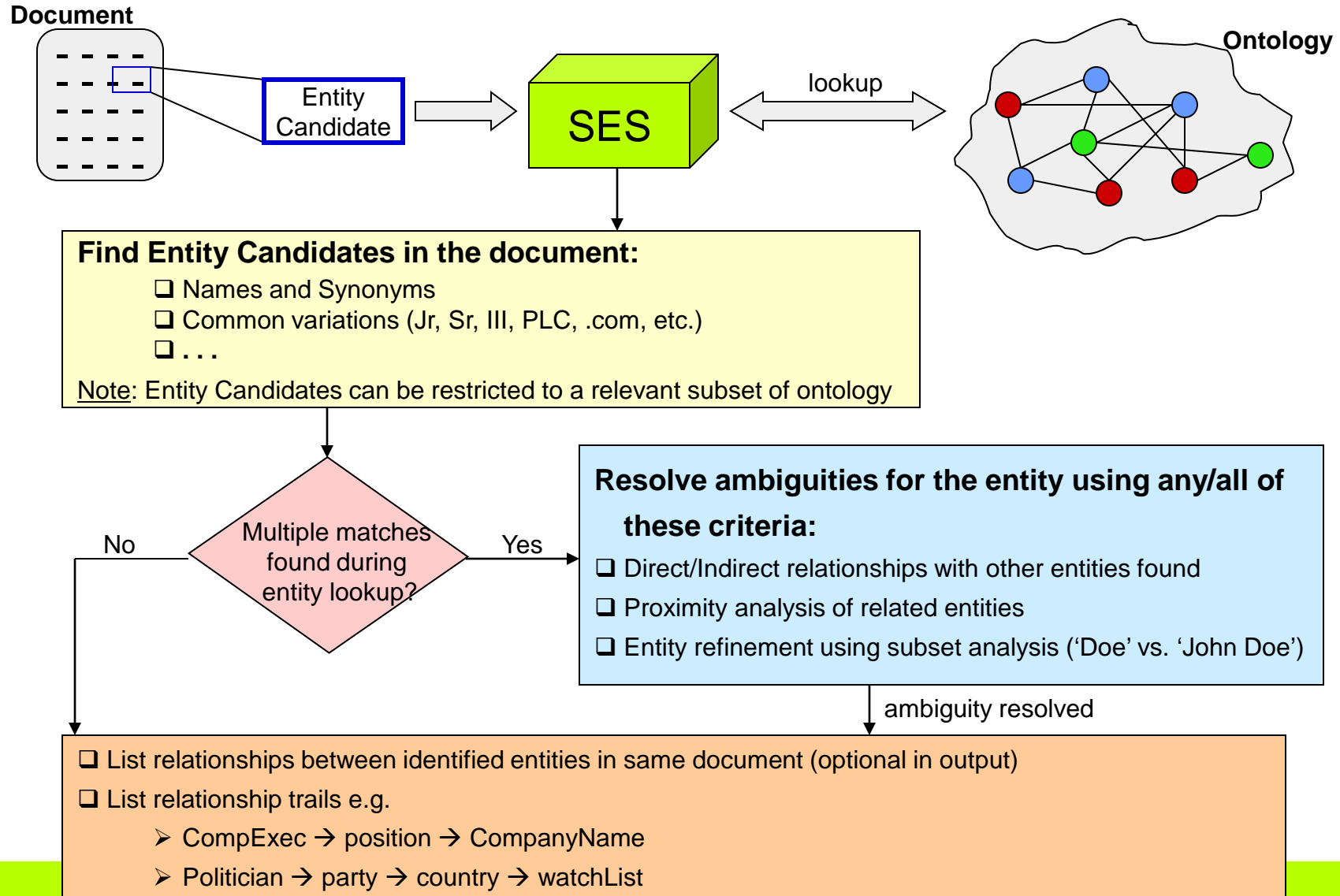
Server: Semantic Enhancement Server classifies content into the appropriate topic/category (if not already pre-classified), and subsequently performs entity extraction and content enhancement with semantic metadata from the Semagix Freedom Ontology

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



Ambiguity Resolution during Metadata Extraction from content text



Overcoming the key issue of resolving ambiguities in facts & evidence

- ◆ Aggregation and normalization of any type of fact and evidence into the domain ontology
 - ❖ Resolution of issues over terminology
 - ◆ i.e. “Benefit number” *is an alias* of “SSN”
 - ❖ Resolution of issues over identity
 - ◆ i.e. is *executive* “Larry Levy” an existing entity or a new entity?
 - ❖ Enabling decisions to be made on the trustworthiness of existing facts
 - ◆ Which source did the data originate from?
 - ◆ How much supporting evidence was there?
 - ❖ Validating and enforcing constraints, e.g. cardinality
 - ◆ President of the United States (has cardinality) = Single
 - ◆ Terrorist (has cardinality) = Multiple

Overcoming the key issue of resolving ambiguities in facts & evidence (Contd...)

- ◆ Managing temporal aspects of the domain
 - ❖ Expiration of entity instances
 - ❖ E.g., “Hillary Clinton” *is no longer the First Lady* of the United States *but was until* “May 3rd 2001”
- ◆ Providing auditing capabilities
 - ❖ Stamping evidence with *date, time* and *source*
 - ❖ E.g., *Terrorist*: “Seamus Monaghan”; *date extracted*: “2003-01-30; *time extracted*: 16:45:27; *source*: FBI Watch list
- ◆ Ontological relationships makes for more expressive model and provide better semantic description (compared to taxonomies)
 - ❖ Information can be presented in natural language format
 - ❖ E.g., “Bob Scott” *is a founder member of business entity* “AIX LLP” *that has traded in* “Iran” *that is on* “FATF watch-list”

Example Scenario 1

Sample content text

Have you ever been to Athens?
How about Japan?

Ontology Matches:

- A: Athens[, Greece, Europe]
- B: Athens[, Georgia, United States of America, North America]
- C: Athens[, Ohio, United States of America, North America]
- D: Athens[, Tennessee, United States of America, North America]
- E: Japan[, Asia]

Scores:

A, B, C, D and E all scored equally – hence no ambiguity resolution possible

Example Scenario 2

Sample content text

Have you ever been to Athens?
Or anywhere else in Georgia?
How about Japan?

Ontology Matches:

- A: Athens[, Greece, Europe]
- B: Athens[, Georgia, United States of America, North America]
- C: Athens[, Ohio, United States of America, North America]
- D: Athens[, Tennessee, United States of America, North America]
- E: Georgia[, Asia]
- F: Georgia[, United States of America, North America]
- G: Georgia On My Mind, Inc.
- H: Japan[, Asia]

Scores:

B and F scored highest because of exact text match and relationship

Result:

Entity Ambiguity Resolved

Automatic Semantic Annotation of Text: Entity and Relationship Extraction

SEMAG!X

Blue-chip bonanza continues

Dow above 9,000 as [HP](#), [Home Depot](#) lead advance; [Microsoft](#) upgrade helps techs.

date time
August 22, 2002: 11:44 AM EDT

phrase phrase
By: Alexandra Twin, CNN/Money Staff Writer

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

time financial index
Around 11:40 a.m. ET, the [Dow Jones industrial average](#) gained 65.06 to 9,022.09, continuing a more
date stock exchange
than 1,300-point resurgence since July 23. The [Nasdaq](#) composite gained 9.12 to 1,418.37.

financial index
[The Standard & Poor's 500 index](#) rose 9.61 to 958.97.

company stockSym \$ \$
[Hewlett-Packard](#) ([HPQ](#): up \$0.33 to \$15.03, Research, Estimates) said a report shows its share of

the printer market grew in the second quarter, although another report showed that its share of the

continent region continent
computer server market declined in [Europe](#), the [Middle East](#) and [Africa](#).

company stockSym \$ \$
[Home Depot](#) ([HD](#): up \$1.07 to \$33.75, Research, Estimates) was up for the third straight day after
topping fiscal second-quarter earnings estimates on Tuesday.

tech category company
Tech stocks managed a turnaround. [Software](#) continued to rise after [Salomon Smith Barney](#) upgraded

company stockSym \$ \$
No. 1 software maker [Microsoft](#) ([MSFT](#): up \$0.55 to \$52.83, Research, Estimates) to "outperform"

\$ \$ company
from "neutral" and raised its price target to \$59 from \$56. Business software makers [Oracle](#)
stockSym \$ \$ company stockSym \$ \$
([ORCL](#): up \$0.18 to \$10.94, Research, Estimates), [PeopleSoft](#) ([PSFT](#): up \$1.17 to \$20.67,

company stockSym \$ \$
Research, Estimates) and [BEA Systems](#) ([BEAS](#): up \$0.28 to \$7.12, Research, Estimates)

all rose in tandem.

KB, statistical
and linguistic
techniques

competes
with

Automatic Semantic Annotation

COMTEX Tagging

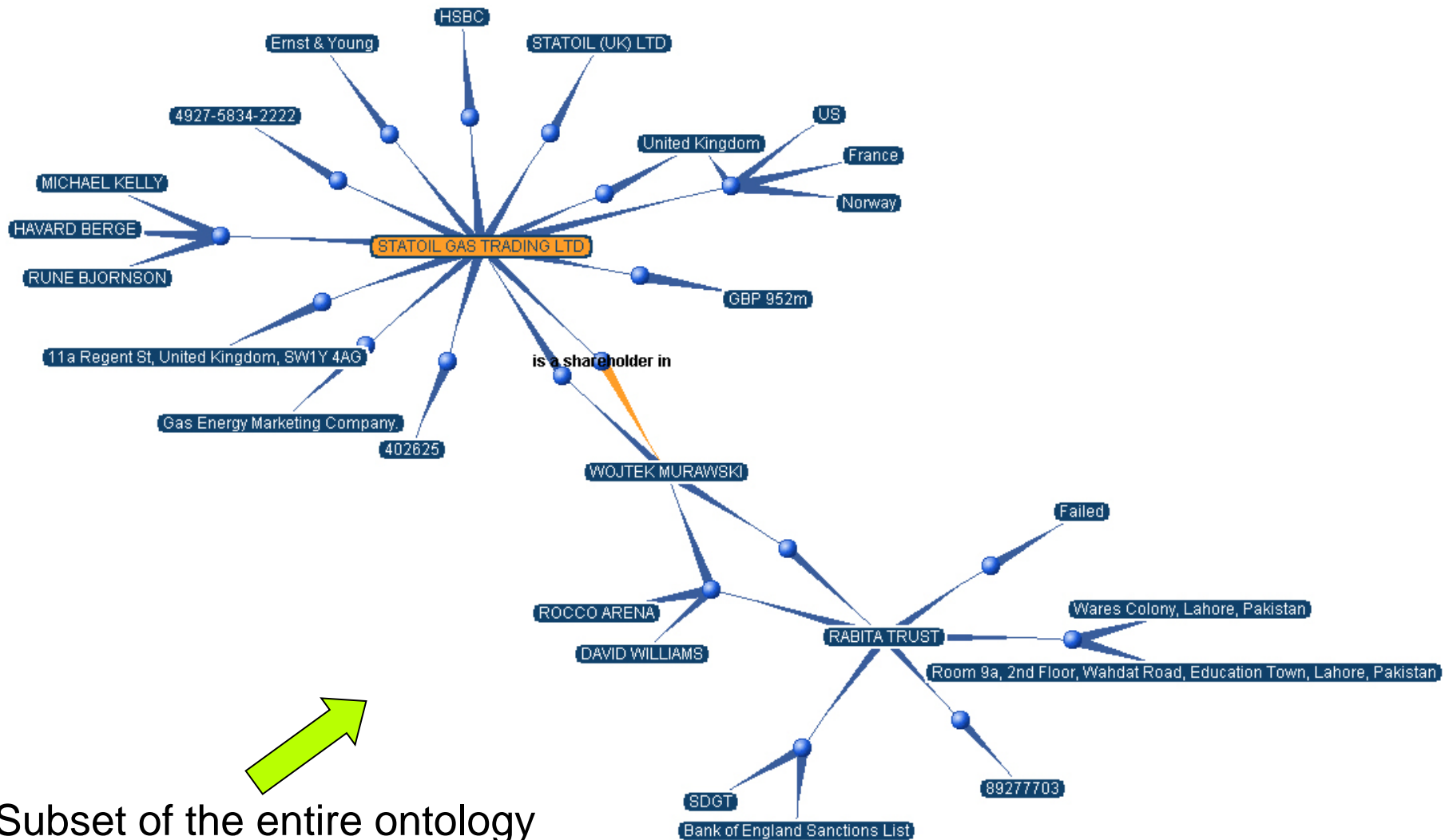
Value-added Semagix Semantic Tagging

Content 'Enhancement' Rich Semantic Metatagging

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

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

AML Ontology Schema (Assertional Component)



Subset of the entire ontology

Performance Issues

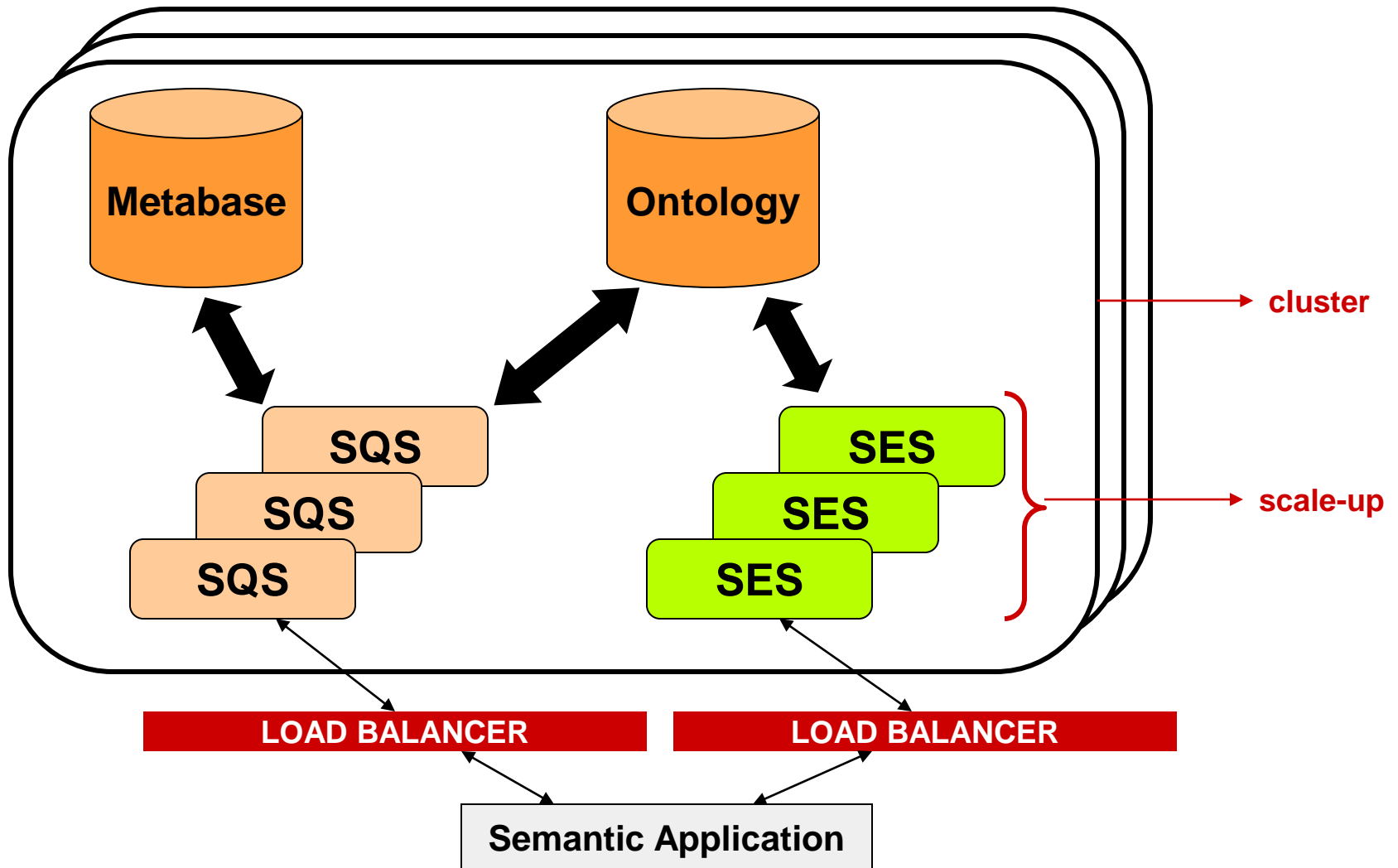
Ontology Storage and Access

- ◆ Ontology typically stores millions of entities, attributes and relationships for any given application
- ◆ Natural implication → how to store it efficiently and most optimally so that accessing ontology does not degrade performance?
- ◆ What are the storage scheme possibilities?
 - ❖ Database storage (RDBMS)
 - ❖ can logic-based /prolog systems handle this size and computation?
 - ❖ . . .
- ◆ Any of the above typical storage schemes poses performance challenges for mass applications

Semantic Query Processing and Analytics

- Solution: In-memory semantic querying (semantic querying in RAM)
 - Complex queries involving 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

Scalable Architecture



Few Application Examples

BLENDED BROWSING & QUERYING INTERFACE

VideoAnywhere

Back Forward Reload Home Search Netscape Print Security Stop

Bookmarks Location: <http://adept.cs.uga.edu/VideoAnywhere/search.htm>

Instant Message WebMail Contact People Yellow Pages Download Channels

VideoAnywhere

About VA Demos Team Members Resources

ATTRIBUTE & KEYWORD QUERYING

SEMANTIC BROWSING

All Assets

- Business
- Movie Assets
 - Film Festivals
 - Interviews
 - Movies
 - Reviews
- News
- Travel

Title: Alien
 Contents:
 Classification: **Movie Assets-Movies**
 Film Name:
 Cast:
 Director:

Submit Clear

Targeted e-shopping/e-commerce

amazon.com

SEARCH VIDEO BROWSE GENRES TOP SELLERS MUSIC NEW & FUTURE RELEASES GIFTS DVD e-CARDS & FAMILY AUCTIONS YOUR ACCOUNT AWARDS & RECOMMENDATIONS

Search Video: Go!

☐ VHS only
☐ DVD only
☒ All Formats

Search Video: [Title, Actor, Director](#) [Video Search Tips](#)

Your Search Results

We found no matches for keywords: "Alien+Resurrection+". Below are results for include "alien resurrection". If you prefer, you may try another search.

Top matches for this search:

- [Alien Resurrection](#) (1997) -- Sigourney Weaver, Winona Ryder, VHS
- [Alien Resurrection](#) (1997) -- Sigourney Weaver, Winona Ryder, DVD
- [Alien Resurrection](#) (1997) -- Sigourney Weaver, Winona Ryder, VHS Widescreen

Full Results: 5 items are shown below.

[Alien Resurrection](#) ~ VHS
 Sigourney Weaver, Winona Ryder / Subtitled in Spanish
 Our Price: \$12.99 ~ You Save: \$1.99 (13%)
[Read more about this title...](#)

assets access

uniform view of worldwide distributed assets of similar type

Details

Title: Alien-Resurrection

Director:

Actors: Sigourney Weaver Winona Ryder

The experiments begin. An unholy combination of human and alien genetics, made possible by an uneasy alliance between a renegade band of smugglers and a zealous cadre of scientists and officials. One subject is familiar -- a woman horrifically linked to the alien species that now elicits so much scrutiny. Ripley is back and all is not what it seems... Alien Resurrection stars Sigourney Weaver as Ripley, the role she originated in Ridley Scott's 1979 classic, Alien, and reprised in James Cameron's Aliens (for which she received an Academy Award nomination)

Contents

Search Results

Title	Details	View	Buy It
Alien-Resurrection		VDO	
Alien Resurrection		MOV	
Alien Resurrection		MOV	
Alien Resurrection		ASX	
Alien Resurrection		ASX	

low_alienrest.asx - Windows M...

File View Play Favorites Go Help

Showcase

Clip: Alien Resurrection

Author: Hollywood Online

Copyright:

Playing 02:22 / 02:31

POWER • THROUGH • RELEVANCE

Semantic Enhancement used in Semantic Search

SEMAG!X

home | your feedback



Precise
MEDIA SEARCH

SEARCH FOOTBALL

Enter search term in appropriate field:

Player
Jamal Anderson
Team
Location
Other Keywords

One field is usually enough.
Entering more yields more precise,
but fewer, results.

Select media:
☒ Audio ☒ Video
☐ Recent ☐ Week
☐ Month ☒ All

Find It

Choose another category to search in:
Football

home



Precise
MEDIA SEARCH

Search results in Football for jamal , anderson
6 results found

PRECISE SEARCH

Query time:0.001 seconds

Exact match for your search...

1. [Week 3 top 10: Anderson TD run](#)
Jamal Anderson runs 26 yards for a touchdown against the Panthers.
Source : NFL.com
League : NFL
Posted Date : 9/20/2000
Teams : Atlanta Falcons

Close match for your search...

1. [Week 13 top 10: Anderson run](#)
Mike Anderson rumbles 80 yards down the sideline
Source : NFL.com
League : NFL
Posted Date : 12/14
Teams : Kentucky Wildcats

NFL.com: NFL Films 00: Top 10 plays of Week 3 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh

Address http://www.nfl.com/multimedia/00top10p3.html

Links CollegeClub Syndicate Partner Log In Demos

NEWS STATS ROSTERS DEPTH

TEAMS PLAYERS SCHEDULES FANTASY NFL FILMS NFL INSIDER TV & RADIO NEWSLETTER TICKETS AUCTIONS SEARCH

NFL FILMS

Top 10 plays of Week 3

The Week 3 top 10 turned, with covered. Mike D... touchdown for the... Jordan scored after blocking a punt and the... Batch hit... Germane Crowell for 50-yard scoring strike

Click below for the rest of the exciting highlights


Week 3 top 10: Blake to Poole TD
Jeff Blake hits Keith Poole on a 49-yard passing strike against the Seahawks.
RealVideo: 56k | 300k | G2

Week 3 top 10: Trevor Pryce fumble return
Trevor Pryce picks up a fumble and runs 28 yards for a touchdown.
RealVideo: 56k | 300k | G2

Week 3 top 10: Anderson TD run
Jamal Anderson runs 26 yards for a touchdown against the Panthers.
RealVideo: 56k | 300k | G2

Week 3 top 10: Coleman TD

home



Precise
MEDIA SEARCH

Football

RESULTS SEARCH

Week 3 top 10: Anderson TD run

Click to play
REAL

http://www.nfl.com

Jamal Anderson runs 26 yards for a touchdown against the Panthers.
Produced by: NFL.com
Posted Date: 9/20/2000
League: NFL
Teams: Atlanta Falcons
Players: Jamal Anderson

View the original source HTML page. Verify that the source page contains no mention of Team name and League name. They are value-additions to the metadata to facilitate easier search.

Click on first result for Jamal Anderson

View metadata. Note that Team name and League name are also included in the metadata

POWER • THROUGH • RELEVANCE

Help Content Knowledge

Search Results for: drug Procrit in category Medical

FDA notifications. Watch out for counterfeit Procrit, 2 lot
(no description available)

Darbepoetin alfa administered
The objectives of this study were to assess the efficacy and safety of darbepoetin alfa (Aranesp) administered with solid tumors receiving chemotherapy.

Erythropoietic agents as neuroprotectants
Erythropoietin is the primary physiologic agent that exerts its effect by binding to cell surface receptors. It has been shown that both erythropoietin and its analogs have neuroprotective effects in various models of neuronal injury.

Epoetin alfa: current and future indications and nursing implications.
Cancer-related anemia commonly is associated with fatigue and decreased quality of life (QOL). Treatment to achieve hemoglobin levels in patients receiving chemotherapy can improve QOL.

Pure Red-Cell Aplasia and Response to Erythropoietin
To the Editor: Casadevall et al. (Feb. 1, 2000) reported that patients with pure red-cell aplasia and antierythropoietin antibodies responded to recombinant erythropoietin (Eprex).

Role of oral versus IV iron supplementation in the treatment of iron deficiency
BACKGROUND: Preoperative treatment with oral iron (Janssen-Cilag) or PROCIT, Ortho Biotech Products, L.P., increases the erythropoietic response to transfusion.

Erythropoietin (Procrit; Epogen)
(no description available)

Role of iron in optimizing response to erythropoietin
Approximately 50% of cancer patients have iron deficiency. Since the late 1980s, recombinant human erythropoietin (rHuEPO, epoetin alfa [Epogen,

Zoom

procrit Find Entity

Classes Instances

Content Details

Epoetin alfa: current and future indications and nursing implications.

Cancer-related anemia commonly is associated with fatigue and decreased quality of life (QOL). Treatment to achieve hemoglobin levels in patients receiving chemotherapy can improve QOL.

Authors Buchsel, Patricia C Murphy, Barbara J

Side Effects fatigue

Drug Class recombinant hormone

Drugs Epoetin Alfa Procrit

Companies Ortho Biotech Products, L.P.

Hormones erythropoietin

Symptoms fatigue

producer PubMed

Java Applet Window

therapy related)

medica

(chronic disease)

fat

of breath

male appe

blood cell count

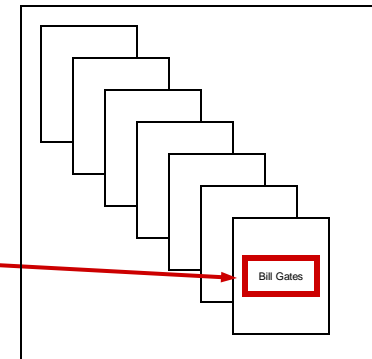
in

Semantic Information Integration spanning three layers of semantic relationships

Single document belonging to a corpus

Blue-chip bonanza continues

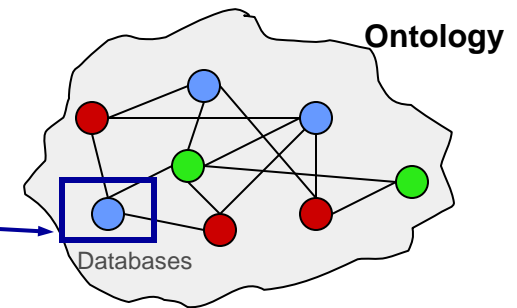
company company company
Dow above 9,000 as HP, Home Depot lead advance; Microsoft upgrade helps techs.
date time
August 22, 2002: 11:44 AM EDT
phrase phrase
By Alexandra Twin, CNN/Money Staff Writer
city company
New York (CNN/Money) - An upgrade of software leader Microsoft and strength in blue chips including Hewlett-Packard and Home Depot were among the factors pushing stocks higher at midday Thursday.
financial index
with the Dow Jones industrial average spending time above the 9,000 level.
time financial index
Around 11:40 a.m. ET, the Dow Jones industrial average gained 65.06 to 9,022.09, continuing a more than 1,300-point resurgence since July 23. The Nasdaq composite gained 9.12 to 1,418.37.
date stock exchange
The Standard & Poor's 500 index rose 9.61 to 958.97.
company stockSym \$ \$
Hewlett-Packard (HPQ: up \$0.33 to \$15.03, Research, Estimates) said a report shows its share of the printer market grew in the second quarter, although another report showed that its share of the computer server market declined in Europe, the Middle East and Africa.
continent region continent
company stockSym \$ \$
Home Depot (HD: up \$1.07 to \$33.75, Research, Estimates) was up for the third straight day after topping fiscal second-quarter earnings estimates on Tuesday.
tech category company
Tech stocks managed a turnaround. Software continued to rise after Salomon Smith Barney upgraded 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.



Corpus of documents

relationships
across documents
in the same corpus

relationships
within text in
the document



relationships
across documents
outside of
the same corpus

Application to semantic analysis/intelligence

- ◆ Documentary content and factual evidence are integrated semantically via semantic metadata

Cocaine scandal sets society hearts fluttering

Investigators are attempting to establish whether the suspect, Palermo businessman Alessandro Martello, was bluffing when he claimed to work as Mr Micciche's assistant and to have the use of an office in the ministry's Rome headquarters.

Mr Martello's arrest warrant, signed last week along with 10 others, alleged that he had not hesitated to deliver a consignment of cocaine inside the ministry itself, confident in the knowledge that his influential connections would protect him from suspicion.

The cocaine scandal has been a gift for the opposition, which promptly tabled a parliamentary question for the **economics minister** **Giulio Tremonti**, asking how many times the alleged pusher had visited his ministry and whether it was true that he had the use of an office there.

The minister has yet to reply, but it has emerged that Mr Martello's frequent visits were the result of his work as a consultant for a company promoting investment in southern Italy. "What he does in his private life has nothing to do with us," his now ex-employer said.

The businessman, who is now in prison, asked the junior minister to intercede on his behalf with a bank where he was having trouble opening an account. Mr Micciche, addressed familiarly as "Gianfrancuccio", said he would see what he could do.

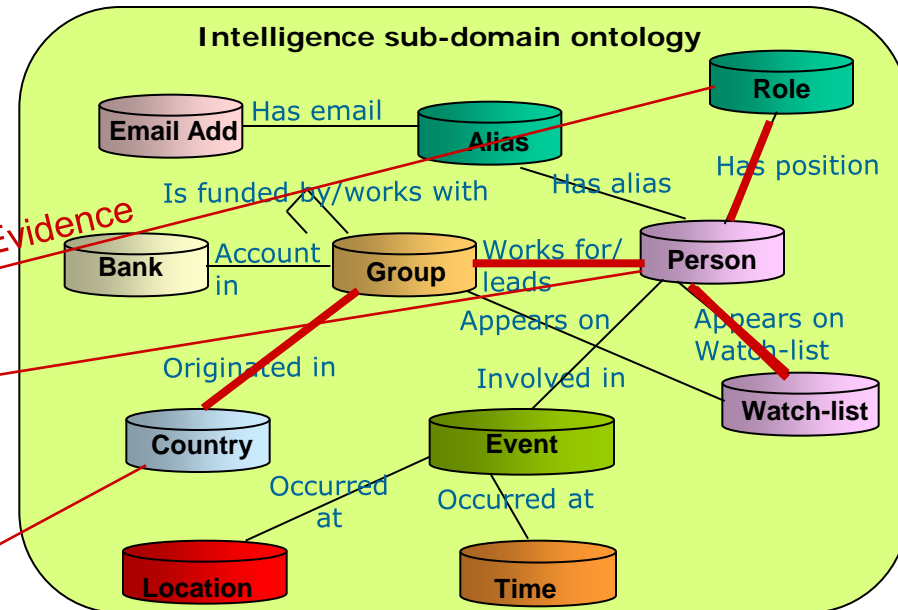
Prime minister Silvio Berlusconi is already engaged in an extenuating personal battle with Milan's anti-corruption magistrates, so the alleged drug entanglements of a junior Sicilian minister are the last thing he needs.

Italian government

[Italian parliament](#)

[Italian president](#)

[Italian government](#)



Classification Metadata: Cocaine seizure investigation

Semantic Metadata extracted from the article:

Person is "Giulio Tremonti"

Position of "Giulio Tremonti" is "Economics Minister"

"Giulio Tremonti" appears on Watchlist "PEP"

Group is Political party "Integrali"

"Integrali" is the "Italian Government"

"Italian Government" is based in "Rome"

Semantic Application Example: Equity Research Dashboard with Blended Semantic Querying and Browsing



Automatic
3rd party
content
integration

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

Competitive
research
inferred
automatically

Semantic Information Integration in *Portals*

Address <http://beta.knowledgeplex.org/topic.html?c=246> Go Links Semagix Email Hotmail ReliableCom The Hub Yahoo! Groups CNN.com DesiRadio

knowledgeplex
The professional resource for affordable housing and community development

your focus
Welcome, **Yash Warke**
Topic: **Workforce Development**
Role: **Public**
Other topics from your profile: **Best Practices & Models, Public Housing, Health Care / Child Care**
Locales: **GA**
Roles: **Public**

topics
All Topics
Affordable Housing Development & Finance
Housing Preservation / Expiring Use
Multifamily Housing
Property Management
Senior / Special Needs Housing
Single Family Housing
Economic Revitalization
Business Recruitment & Retention
Commercial Development Finance
Historic Preservation
Microenterprise
Workforce Development
Fair Housing
Homelessness
Best Practices & Models
Development & Finance Programs
Homeownership / Mortgage Markets
Homebuyer Assistance
Mortgage Lending / Predatory Lending
Land Use & Housing Planning
Environmental Issues
Gentrification & Preserving Affordability
Housing Market Trends & Studies
Landbanking
Planning & Zoning
Smart Growth
Transportation & Parking
Organizational Development
CDC Capacity
Leadership / Skills Training
Measuring Success / Impact
Organizational Effectiveness & Performance
Personal Finance & Asset Creation
Financial Literacy
Financial Strategies & Tools
Underserved Financial Markets

Workforce Development
This page summarizes KnowledgePlex content about **Workforce Development**.
Workforce Development is a general term to describe various efforts to improve the work and work-readiness skills of all citizens. Workforce development efforts are often referred to generically as "Job Training" or "Jobs" programs, and often administered by single-purpose employment organizations or by multi-purpose community-improvement organizations as a part of larger community economic development programs.

documents
HUD's Daily Focus Sep 15 2003
Some Tips on Collecting Data Sep 14 2003
Fundraising: Multiple Mail Appeals Sep 14 2003
Profile of a Healthy Fundraising Program, by Kim Klein Sep 14 2003
Trends in Naturalization Sep 17 2003
Current news on **Workforce Development**.

news
Trends in Naturalization Sep 17 2003 [read]
Families that Receive Government Benefits after Leaving Welfare Are Less Likely to Return Sep 11 2003 [read]
Half of Welfare Recipients without Barriers are Working Sep 11 2003 [read]
Half of Former Welfare Recipients Are Working or Worked Recently Sep 11 2003 [read]

discussion
New Sources Of Development Finance? Sep 11 20:29
Success In Financing Commercial Deals Sep 11 20:30
Immigration And Fair Housing Complaints Sep 11 20:32
Foster Children And Support Housing Sep 11 20:34
Low Interest Rates And Mortgage Lending Abuses Sep 11 20:36
Smart Growth Partnerships Sep 11 20:39
Accountability For Non-profits Sep 11 20:40
Ida Program Growth And Success Sep 11 20:42
Future Of Hope Vi Sep 11 20:44
Arts & Cultural Program Examples Sep 11 20:45

calendar
Sidebars to Cover Story Sep 15 2003 Journal
Public Hotel (Seattle, WA) Sep 15 2003 Journal
Anderson Gardens Sep 15 2003 Journal
Insight Sep 15 2003 Journal
USDA Hopes to Fund Pilot Agriculture Innovation Centers in 2003 Sep 15 2003 Study

ROUGH • RELEVANCE

User profile as a context for semantic integration of diverse yet relevant content

Semantic integration and presentation of various types of personalized content items in one place

Sample content item that is explicitly or implicitly associated semantically to facets in user profile

Anti Money Laundering – Know Your Customer

Transaction and Customer List - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print Mail New

CIRAS 10:39:51 8/1/2003 [View Ontology](#)

Branches

Sort Code	Branch Name	Score
402625	Knutsford	100
408923	Winchester	100
402077	Head office	85
401734	London- Princess St	85
400967	Manchester Regional	0
401754	Lewisham	0
409867	Bedford	0
406582	Jersey	0
401428	University of Sussex, Brighton	0
401432	Lewis	0
402976	Swansea- Sketty	0

Knutsford

Accounts

Account	Account Name	R/B	Customer Since	Aggregate Risk Score
74922753	GIULIO TREMONTI	Retail	1999-05-09	100
18357401	BAYER AG	Business	2001-02-02	0
74523456	TERENCE D CULLEN	Retail	2002-10-13	0

[Show All](#)

Risk Profiles are developed for individuals or companies. If the risk profile changes based on new information the individuals Risk Profile and Branch Aggregate Risk Profile is automatically updated

Retrospective Client Check All Transactions Investigate Client Investigate Transaction New

Internet



View Risk Scores for a specific company or customer

Transaction and Customer List - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media

CIRAS 10:43:56 8/1/2003 [View Ontology](#)

Client Information

Company: STATOIL GAS TRADING LTD
 Individuals:
 Locations:
 Others:

Risk Score

STATOIL GAS TRADING LTD 65
 Aggregate 65 [Details](#)

Company Knowledge

STATOIL GAS TRADING LTD [Company]

Synonyms:
 Statoil Gas Trading

Relationships:
[HAVARD BERGE](#) works for
[RUNE BJORNSEN](#) works for
[MICHAEL KELLY](#) works for
[WOJTEK MURAWSKI](#) is a shareholder in STATOIL GAS TRADING LTD
 STATOIL GAS TRADING LTD is audited by [Ernst & Young](#)

Score Breakdown - Microsoft Internet Explorer

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		

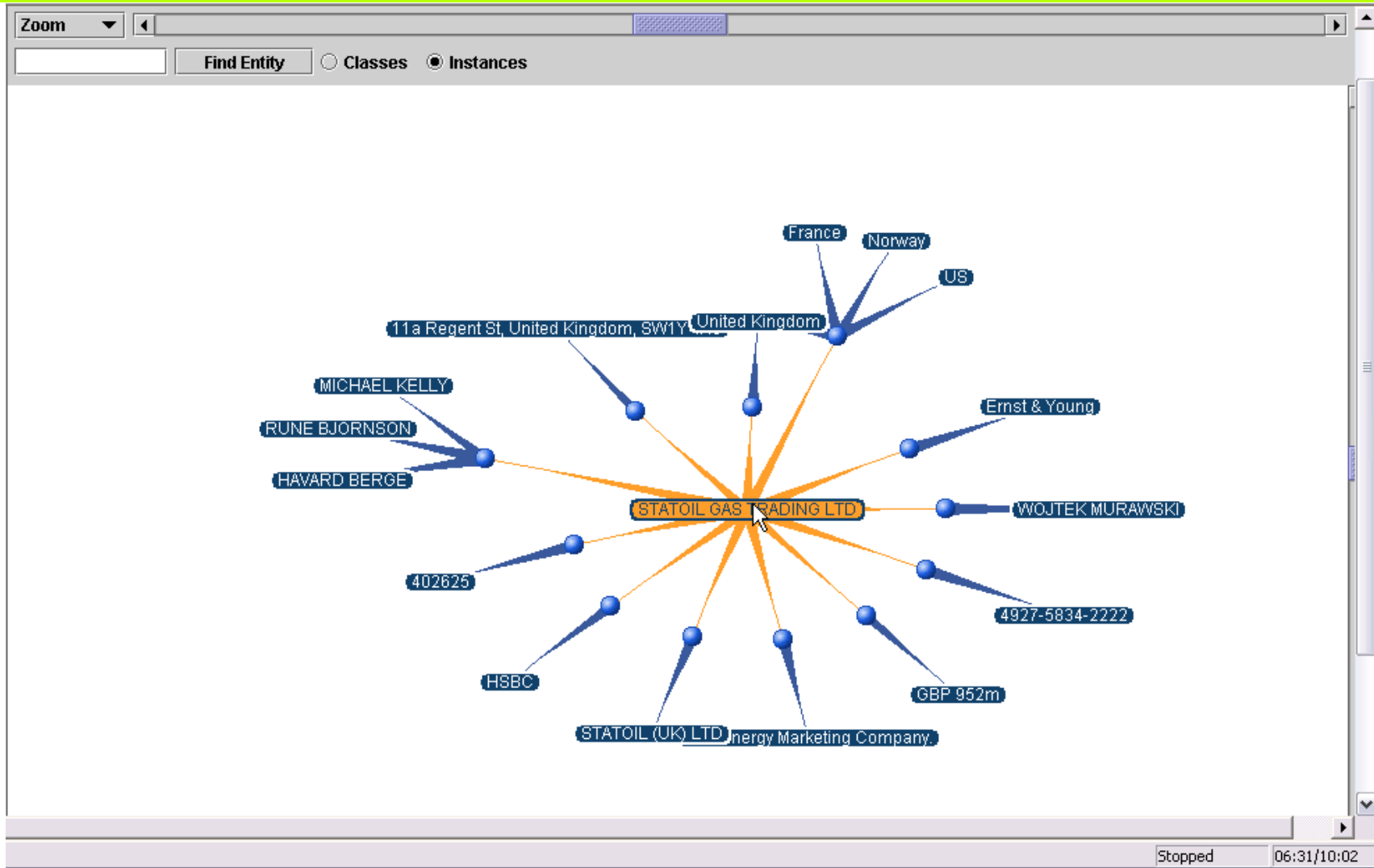
AUDIT TRAIL:
[Know Your Customer Check](#)
 Retrospective Check
 Application Date: 29/01/2003
 Request Outcome: Failed

Buttons: Accept Reject Retrospective Client Check All Transactions Investigate Client Investigate Transaction New

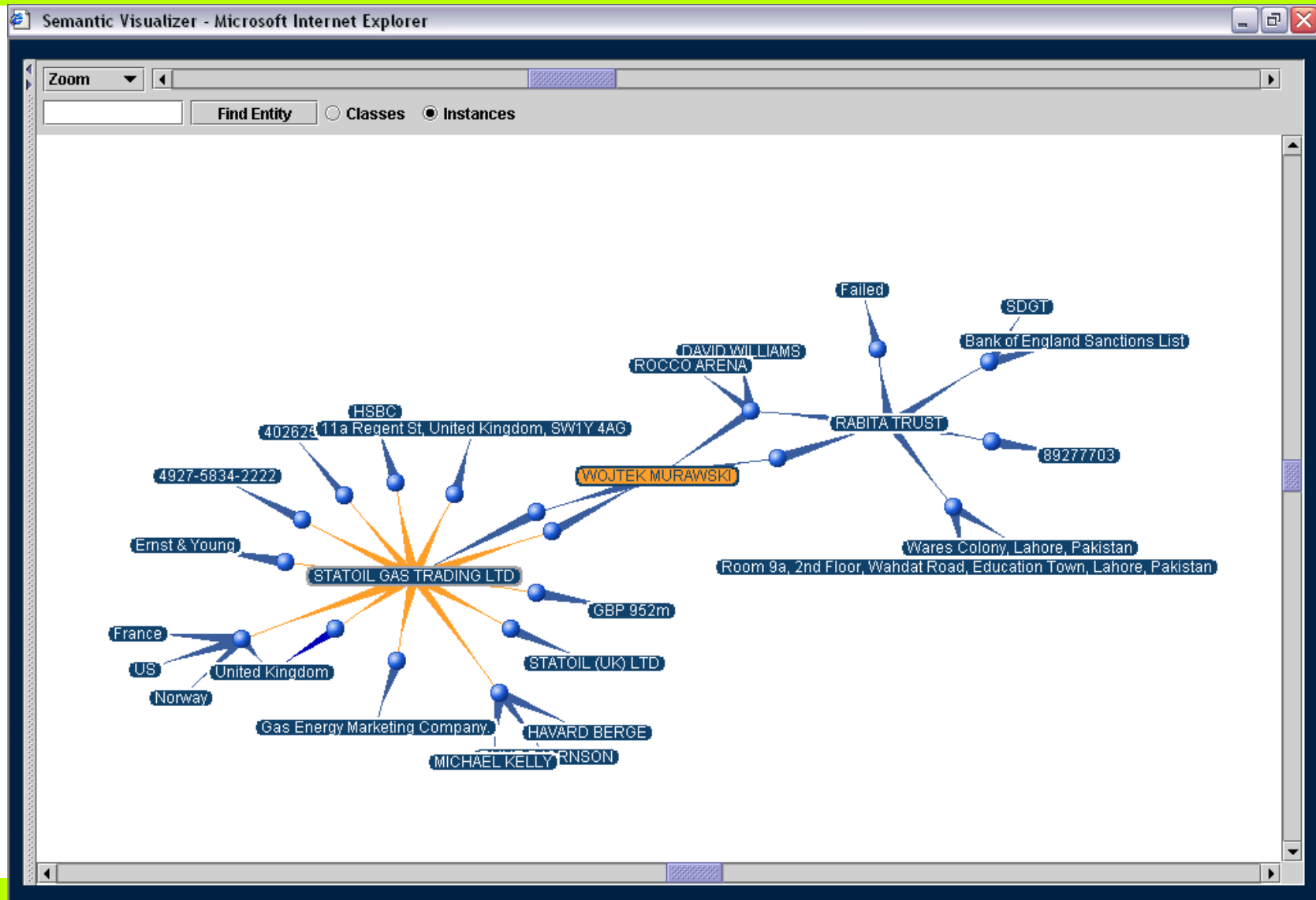
Taskbar: start > RE: WF conten... CD Drive (D:) Camtasia Internet Expl... Adobe Acrobat - ... 100% 10:43 AM Friday 8/1/2003



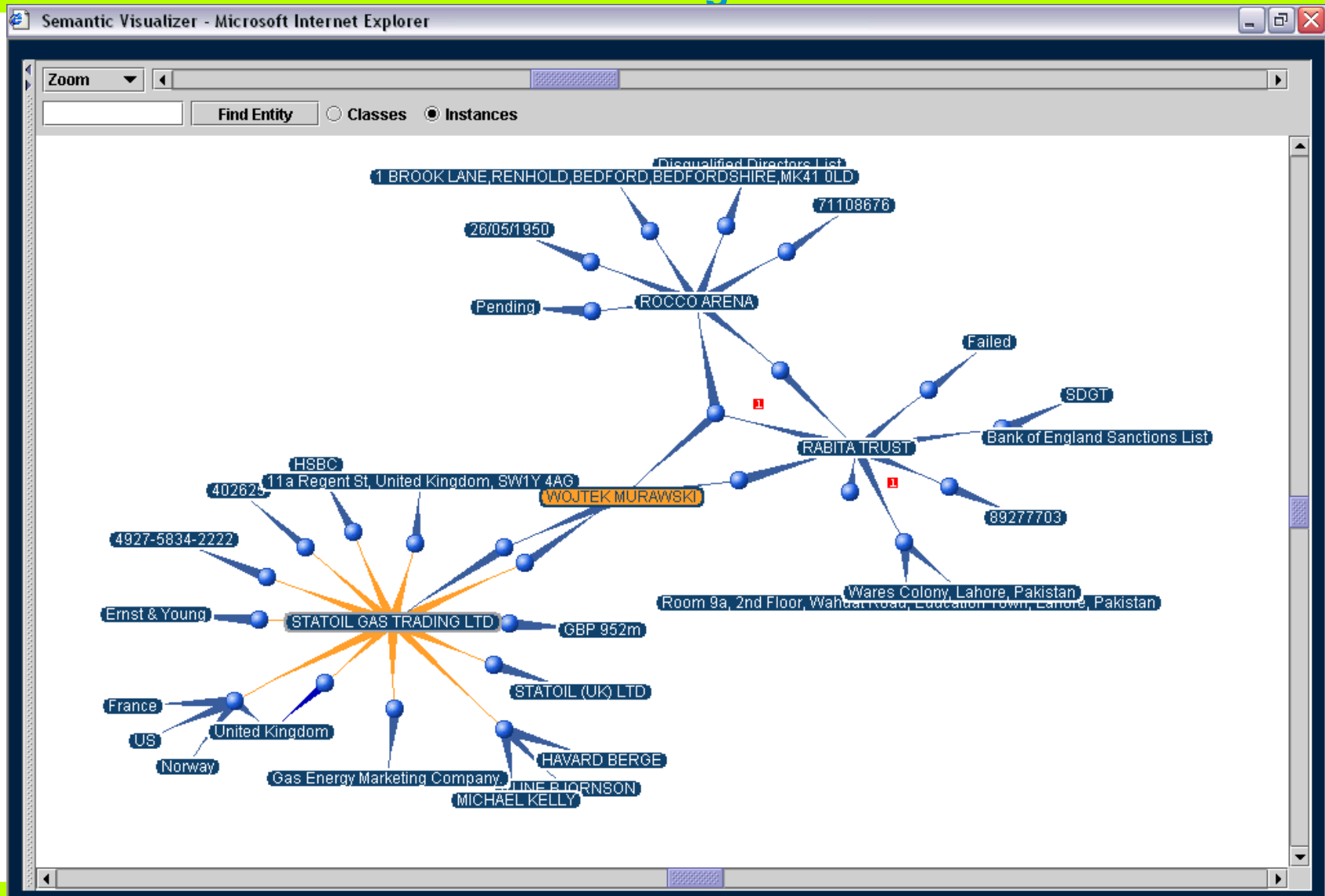
Additional tools allow the user to navigate around the content



Additional tools allow the user to navigate around the content



Additional tools allow the user to navigate around the content



Conclusion

- ◆ Great progress from work in semantic information interoperability/integration of early 90s until now, re-energized by the vision of Semantic Web, related standards and technological advances
- ◆ Technology beyond proof of concept
- ◆ But lots of difficult research and engineering challenges ahead
- ◆ More:
 - (Technology) <http://www.semagix.com/downloads/downloads.shtml>
 - (Research) <http://lsdis.cs.uga.edu/proj/SAI/>
- ◆ Demos available