

From Semantic Search & Integration to Analytics

KMWorld/Intranets 2004

Session: Information Architecture, Track F, October 26, 2004

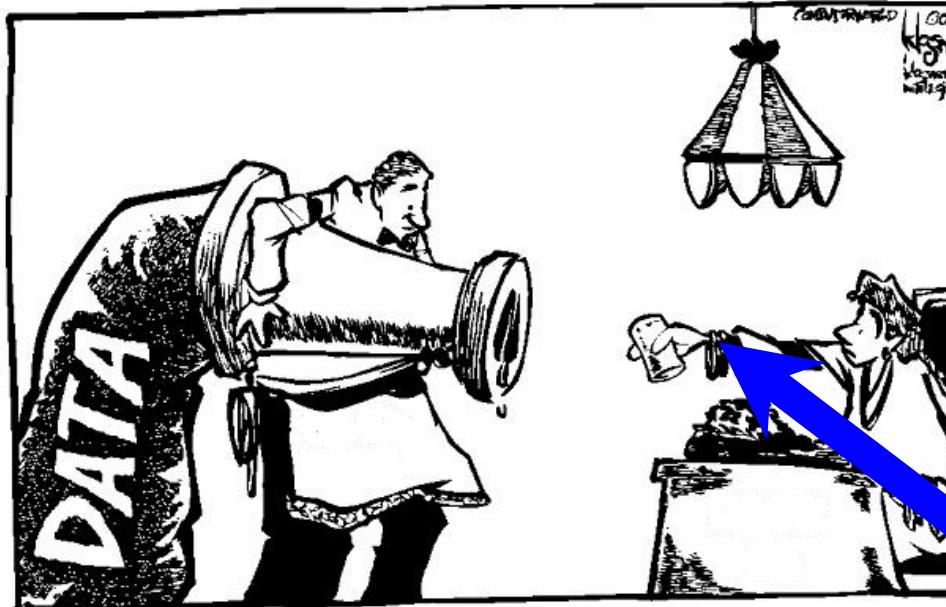
Amit Sheth

Semagix, Inc. and LSDIS Lab, University of Georgia
amit.sheth@semagix.com, amit@cs.uga.edu

© Semagix (when marked by Semagix logo); UGARF, Inc. and Amit Sheth (when marked by LSDIS logo)



What do you want to do with information?



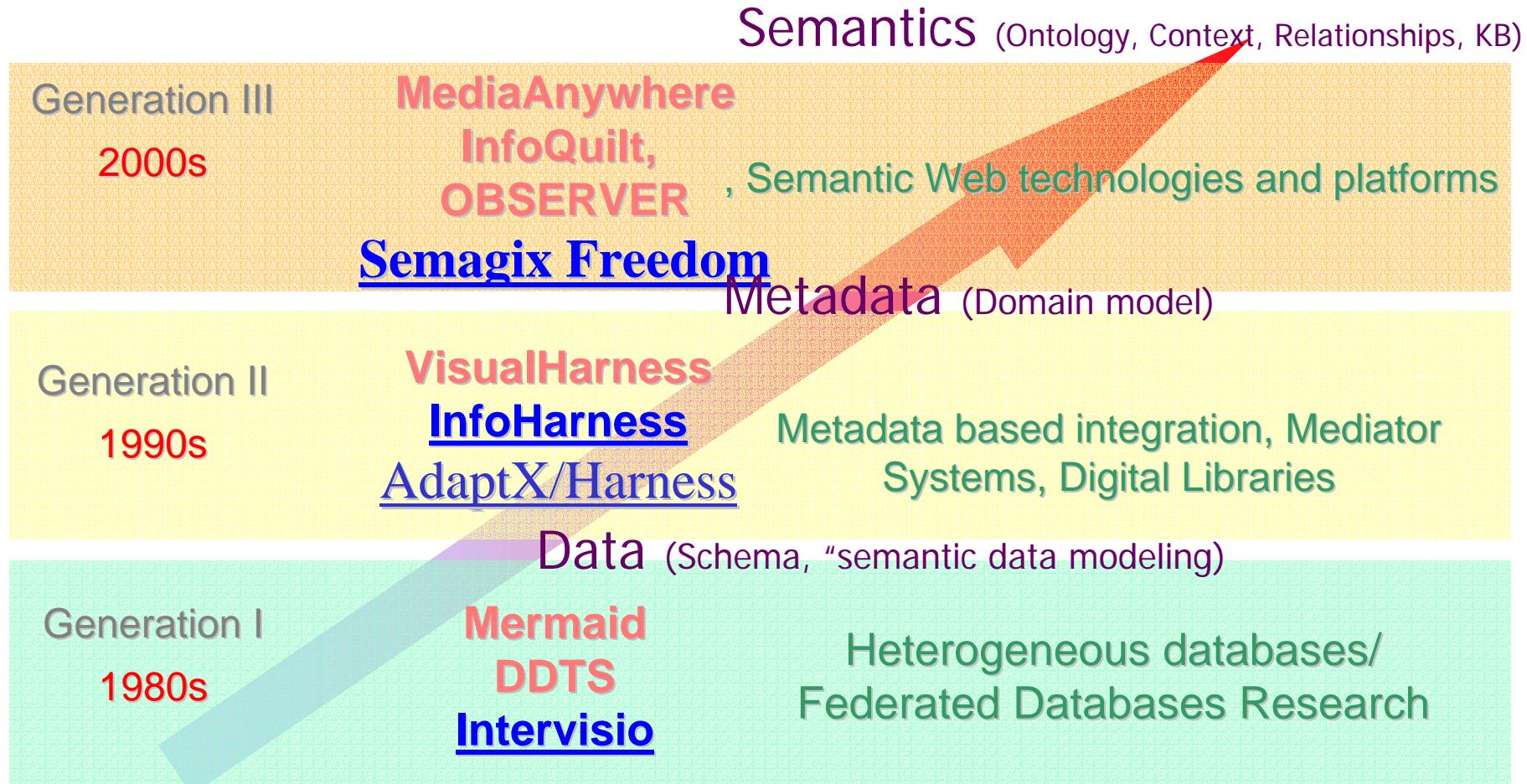
Increasing sophistication in applying semantics

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

Not just relevant data,
but actionable information,
analysis and insight



Three generation of Information Systems: Where we have come from, where we are going



Outline

- Observations
- Semantic Web and its key enablers
 - Ontology: What do real world ontologies look like, how are they created
 - Metadata: what enables automatic metadata extraction, how metadata enables integration of heterogeneous content
- What type of industry and scientific Semantic Applications are built
- Conclusions: Ontologies as a best-in-class approach; analytics as in most demanding applications

Emerging Trends, Changing Focus

- From syntax/structure to semantics
- From techniques that focus on either unstructured data (text) or structured content, to both types and semi-structured data
- From directly analyzing data (warehousing and mining) to ontology based processes of creating high quality metadata and analyzing metadata
- From search and browsing for delivering relevant documents; from locating entities within contents to discovering complex relationships and delivering actionable information with insights;
from semantic search to integration, mining, analytics

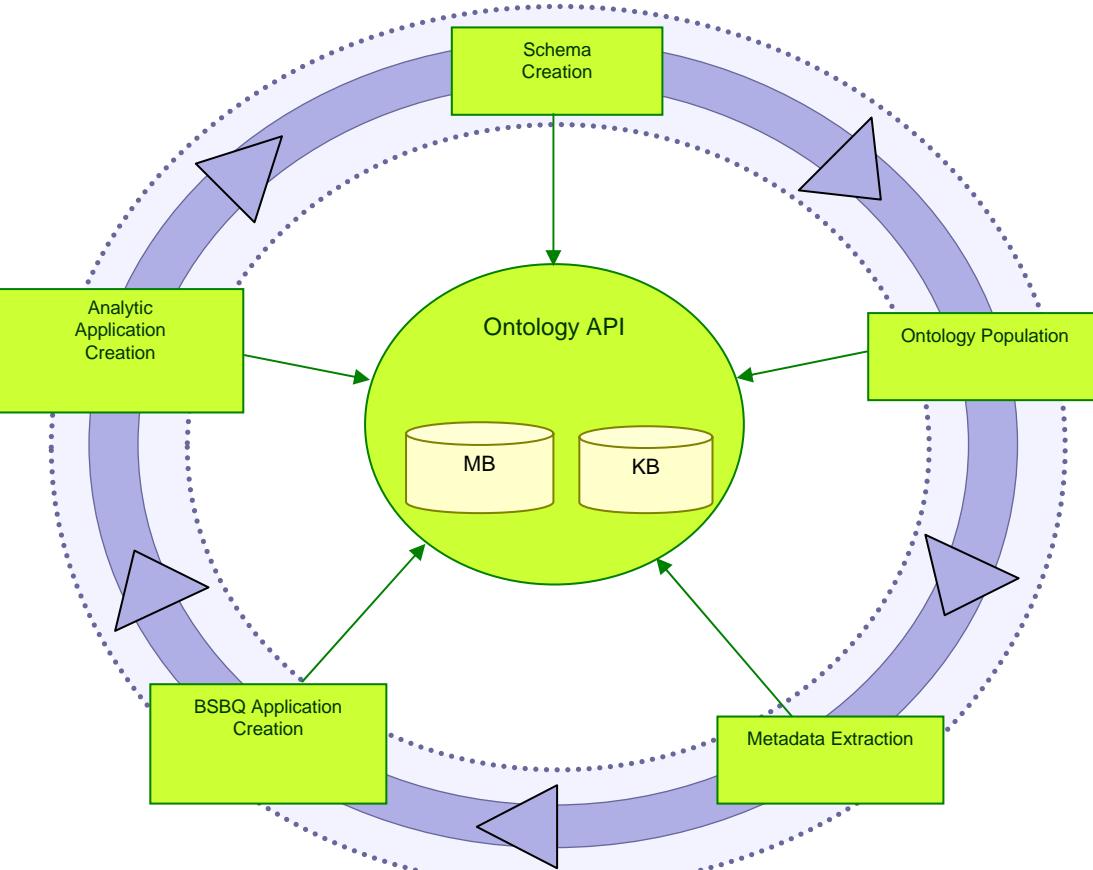
Empirical observations based on real-world efforts

- Comprehensive commercial products support development Ontology-based information systems and Semantic Applications
 - Several commercial companies with offerings that vary from tools to comprehensive enterprise software platforms
 - Deployed, business critical applications can be found
 - Applications validate the importance of ontology the current semantic approaches
 - Significant impact of academic research in early stage companies
- Empirical observations in this talk are based on development of several real-world ontologies and implemented/deployed semantic applications by Semagix and its partners/customers

Semantic Web in a nutshell

- Ontology as the centerpiece
- Metadata that associate meaning to content
- Computing (complex querying, inferencing, other reasoning) that support semantic applications

Ontology-driven Information System Lifecycle



Building a scalable and high performance system with support for:

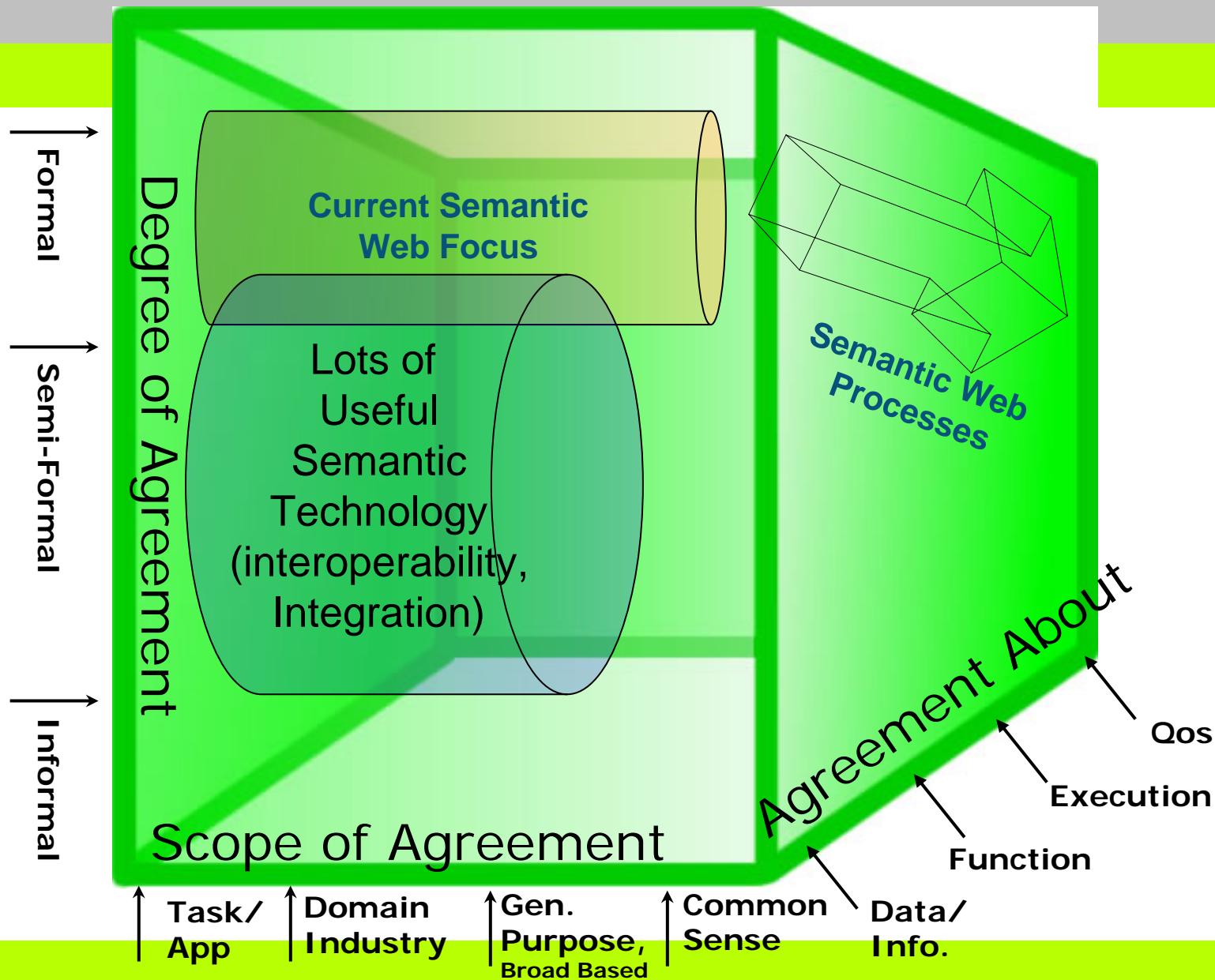
- Ontology creation and maintenance
- Knowledge-based (and other techniques) supporting Automatic Classification
- Ontology-driven Semantic Metadata Extraction/Annotation
- Utilizing semantic metadata and ontology
 - Semantic search/querying/browsing
 - Information and application integration - normalization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

Central Role of Ontology

- Ontology represents agreement, represents common terminology/nomenclature
- Ontology is populated with extensive domain knowledge or known facts/assertions
- Key enabler of semantic metadata extraction from all forms of content:
 - unstructured text (and 150 file formats)
 - semi-structured (HTML, XML) and
 - structured data
- Ontology is in turn the center piece that enables
 - resolution of semantic heterogeneity
 - semantic integration
 - semantically correlating/associating objects and documents



Broad Scope of Semantic (Web) Technology



Other dimensions:
how agreements are reached,
...



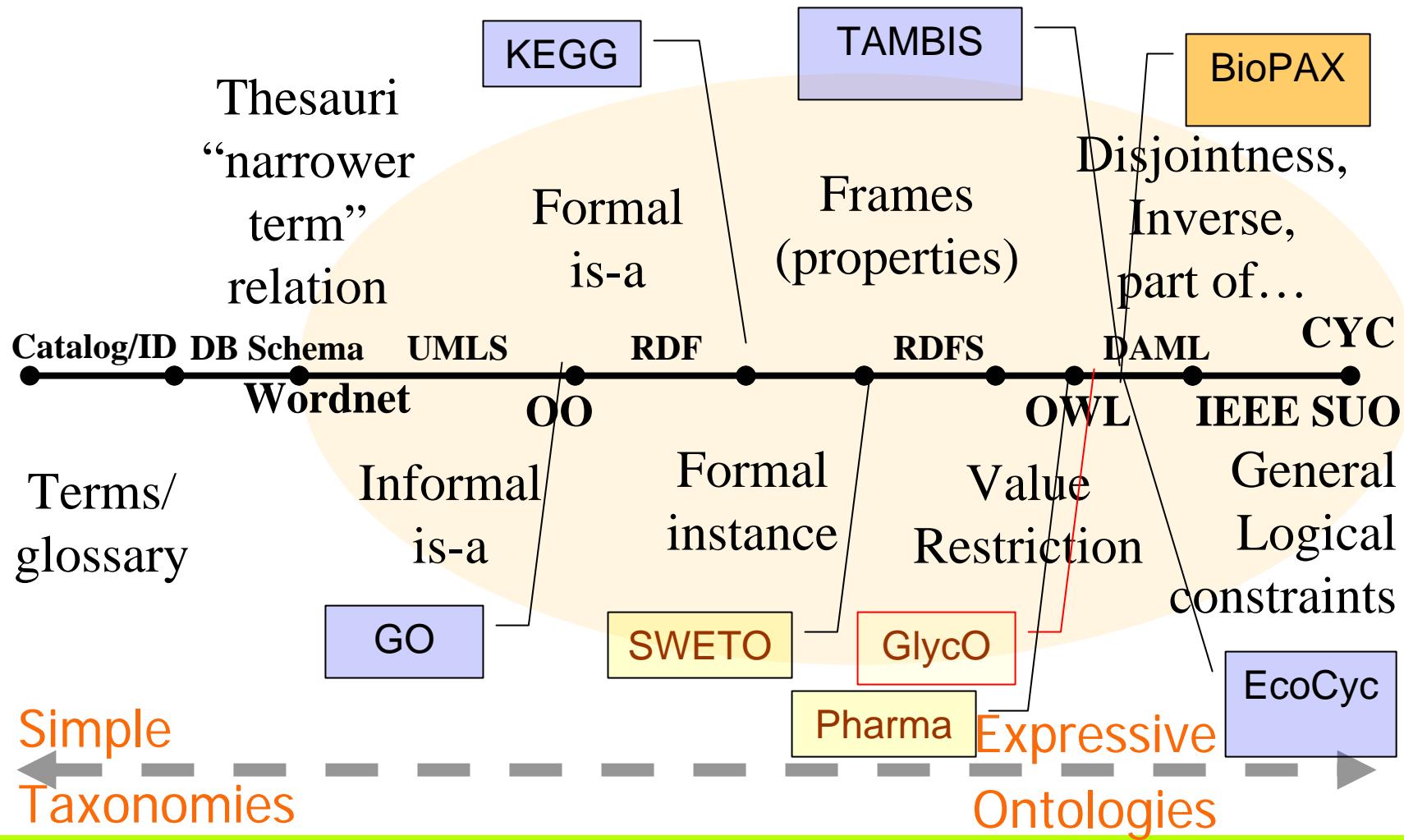
Types of Ontologies (or things close to ontology)

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

Fundamentally different approaches in developing ontologies
at the two end of the above spectrum



Expressiveness Range: Knowledge Representation and Ontologies





Building ontology

- Three broad approaches:
 - social process/manual: many years, committees
 - Can be 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; Description base (assertional component, extension) by automated processes

Option 2 is being investigated in several research projects;

Option 3 is currently supported by Semagix Freedom

Ontology can be very large

Semantic Web Ontology Evaluation Testbed – SWETO v1.4
is

- Populated with over 800,000 entities and over 1,500,000 explicit relationships among them
- Continue to populate the ontology with diverse sources thereby extending it in multiple domains, new larger release due soon
- Two other ontologies of Semagix customers have over 10 million instances, and requests for even larger ontologies exist

GlycO

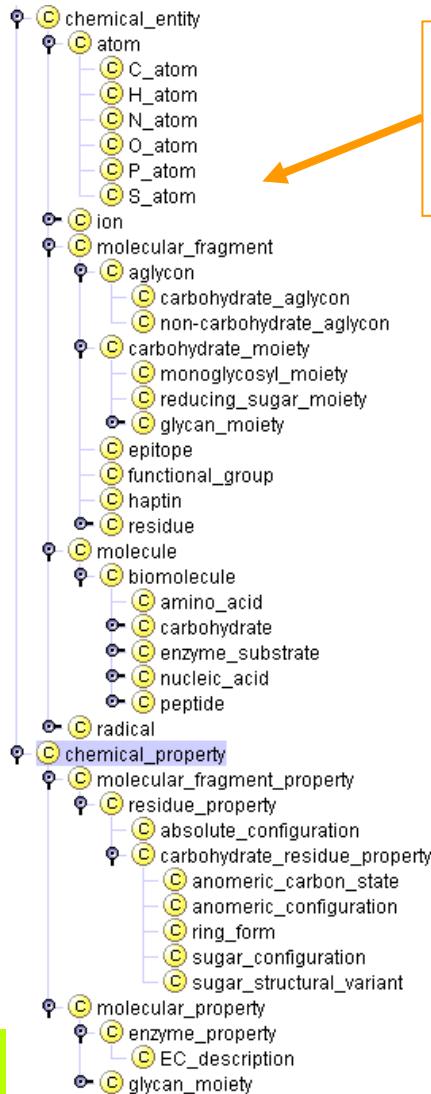
- *is a focused ontology for the description of glycomics*
- *models the biosynthesis, metabolism, and biological relevance of complex glycans*
- models complex carbohydrates as sets of simpler structures that are connected with rich relationships

Ontology schema can be large and complex

GlycO statistics now (and growing)

- 767 classes
- 142 slots
- Instances Extracted with Semagix Freedom:
 - 69,516 genes (From PharmGKB and KEGG)
 - 92,800 proteins (from SwissProt)
 - 18,343 publications (from CarbBank and MedLine)
 - 12,308 chemical compounds (from KEGG)
 - 3,193 enzymes (from KEGG)
 - 5,872 chemical reactions (from KEGG)
 - 2210 N-glycans (from KEGG)

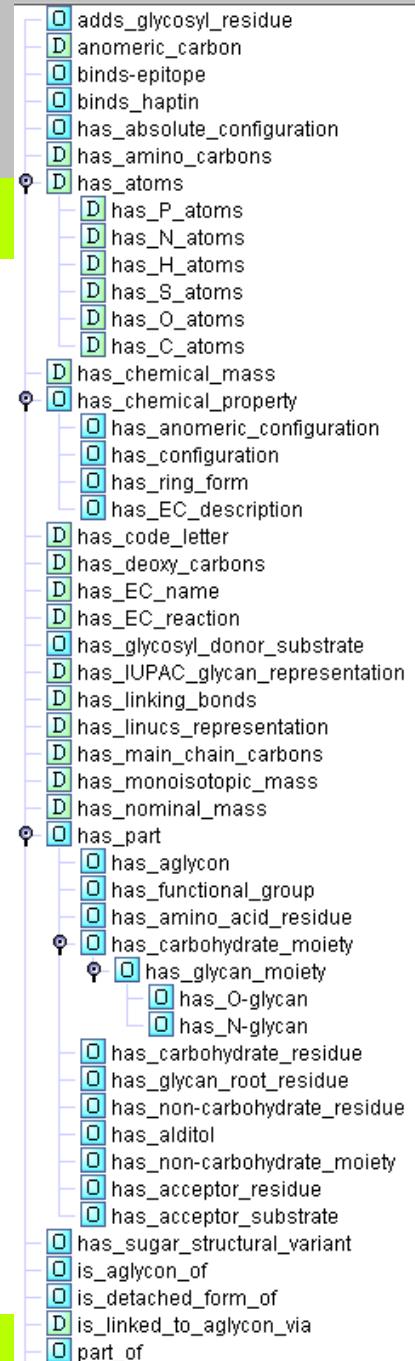
GlycO taxonomy



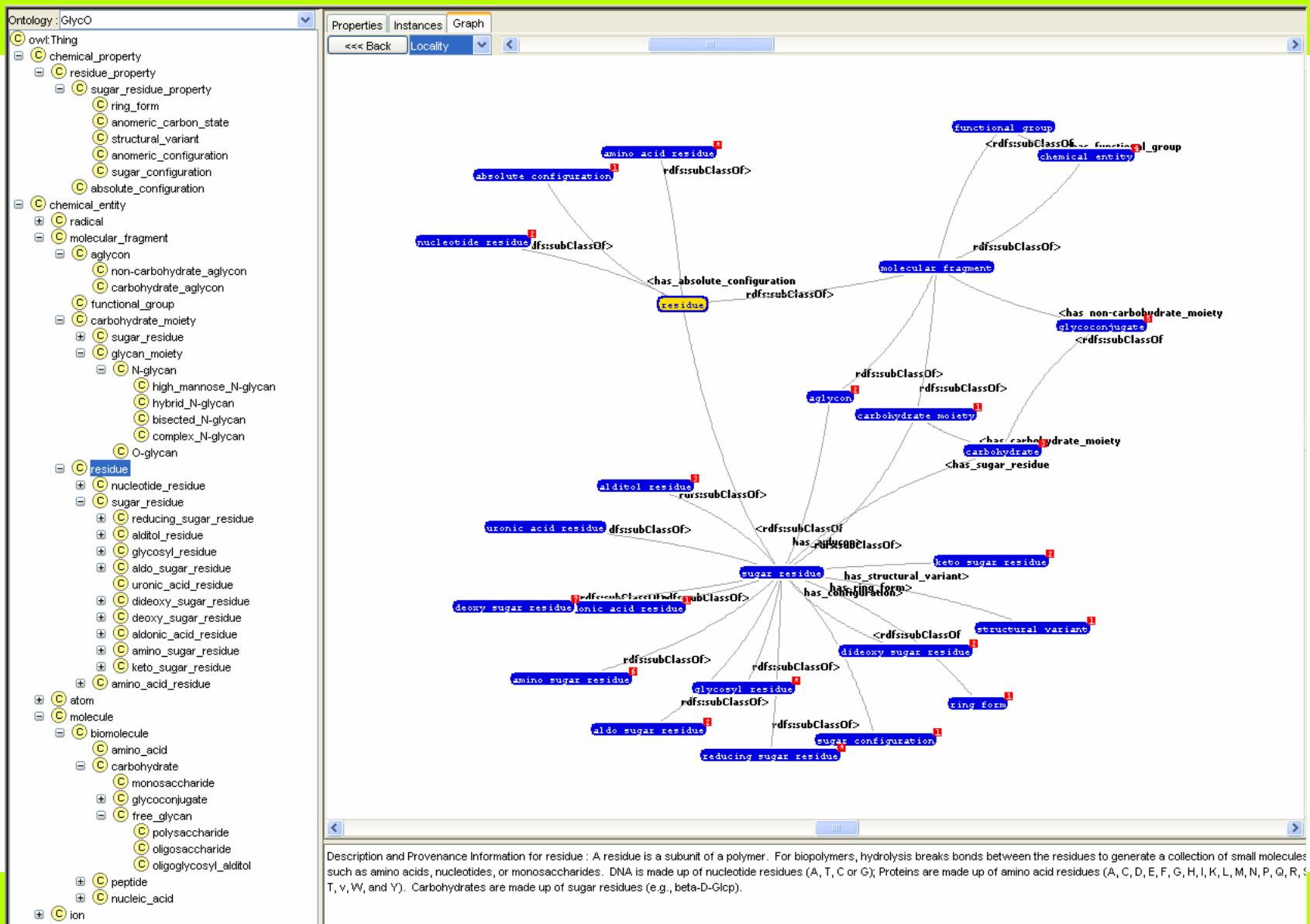
The first levels of the GlycO taxonomy

Most relationships and attributes in GlycO

GlycO exploits the expressiveness of OWL-DL. Cardinality constraints, value constraints, Existential and Universal restrictions on Range and Domain of properties allow the classification of unknown entities as well as the deduction of implicit relationships.



Query and visualization



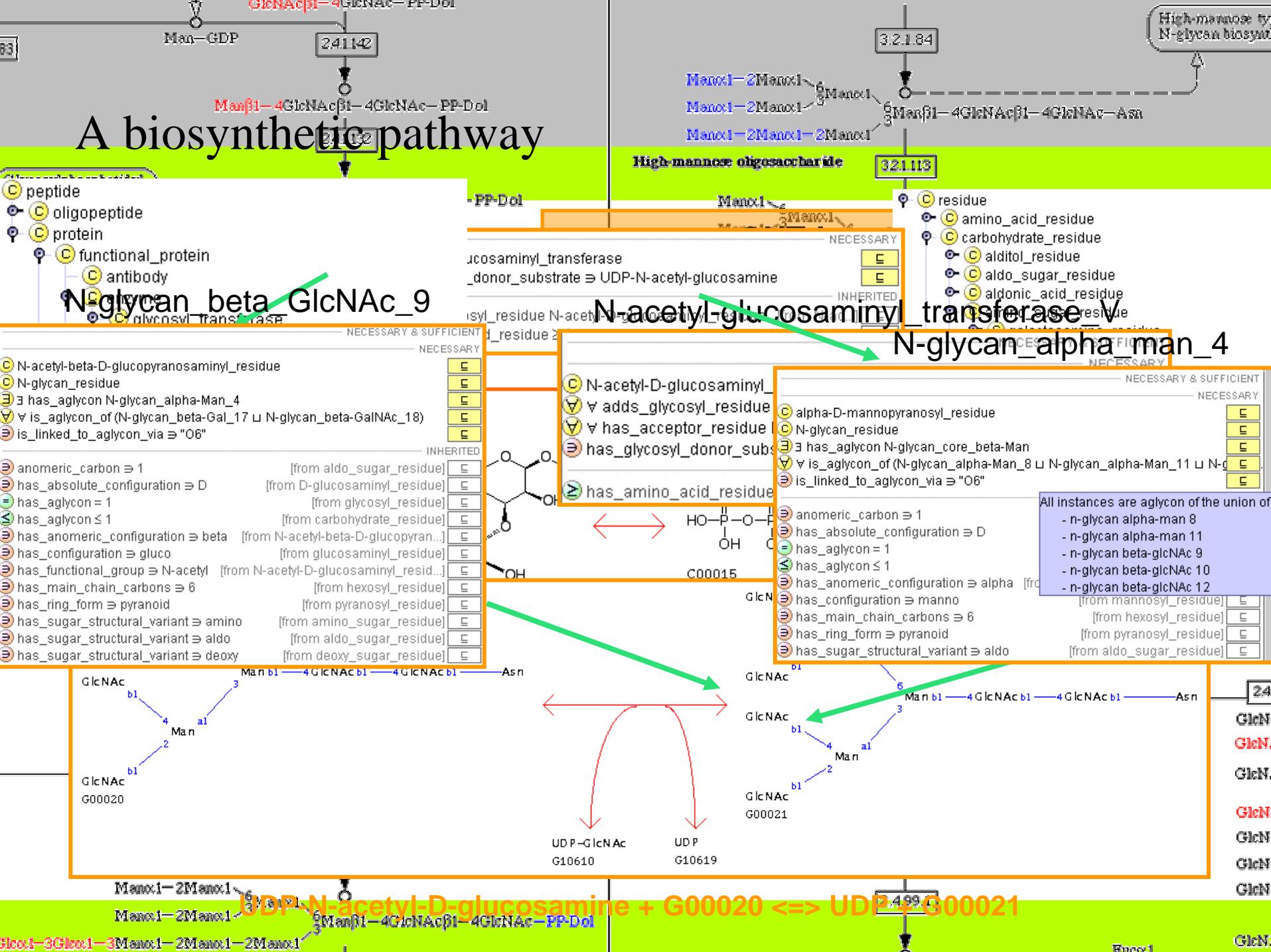
Query and visualization

owl:Thing

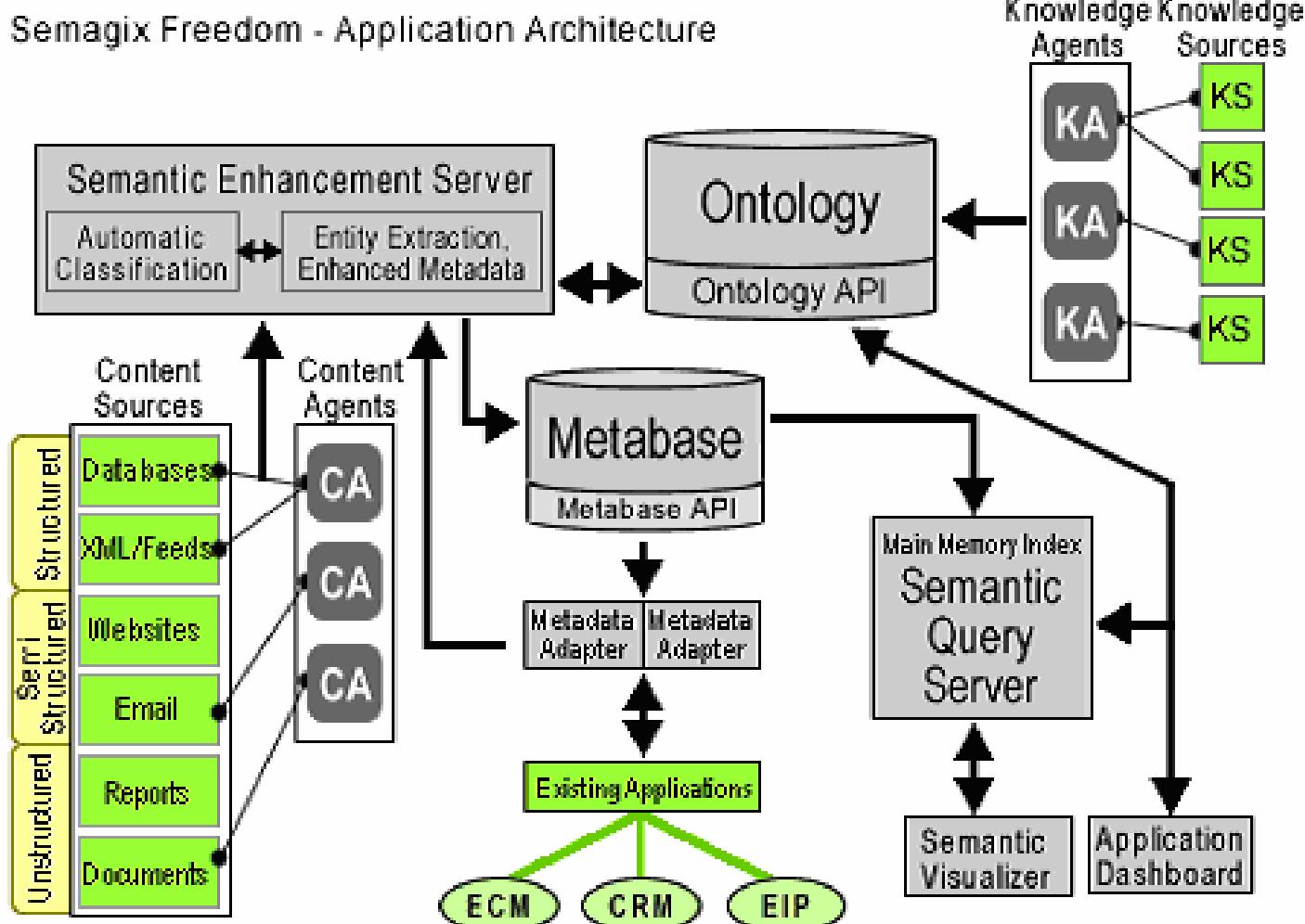
- C chemical_property
 - + C molecular_property
 - + C molecular_fragment_property
- C chemical_entity
 - + C radical
 - C molecular_fragment
 - C epitope
 - + C aglycon
 - C haptin
 - C functional_group
 - C carbohydrate_moiety
 - C reducing_sugar_moiety
 - + C glycan_moiety
 - C monoglycosyl_moiety
 - C residue
 - + C nucleotide_residue
 - + C carbohydrate_residue
 - + C amino_acid_residue
- + C atom
- + C molecule
- + C ion

<<< Back Rotate >>>

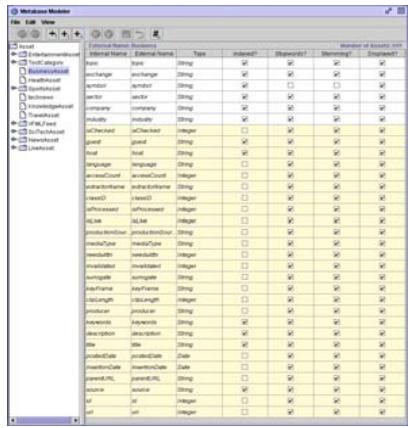
Description and Provenance Information for **glycan_moiety** : a glycan moiety is a glycan that is part of a glycoconjugate, which has a non-carbohydrate moiety



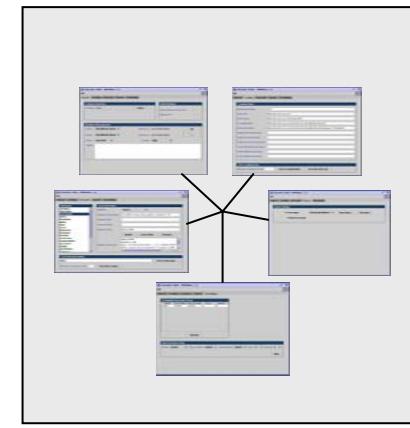
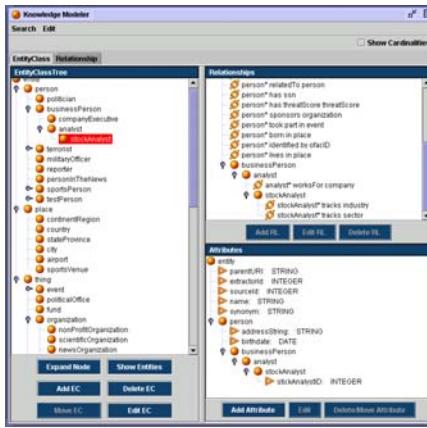
Semagix Freedom Architecture: for building ontology-driven information system



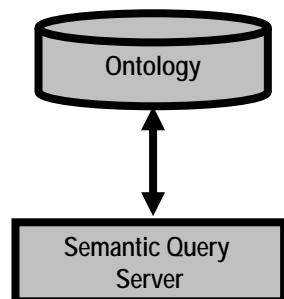
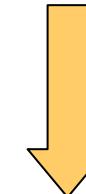
Ontology Creation and Maintenance Steps



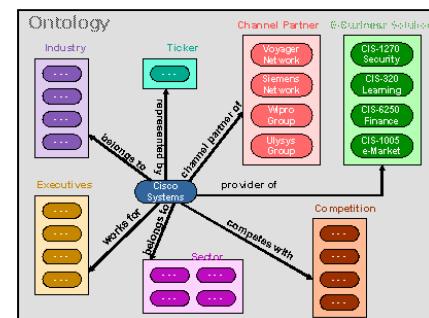
1. Ontology Model Creation (Description)



2. Knowledge Agent Creation

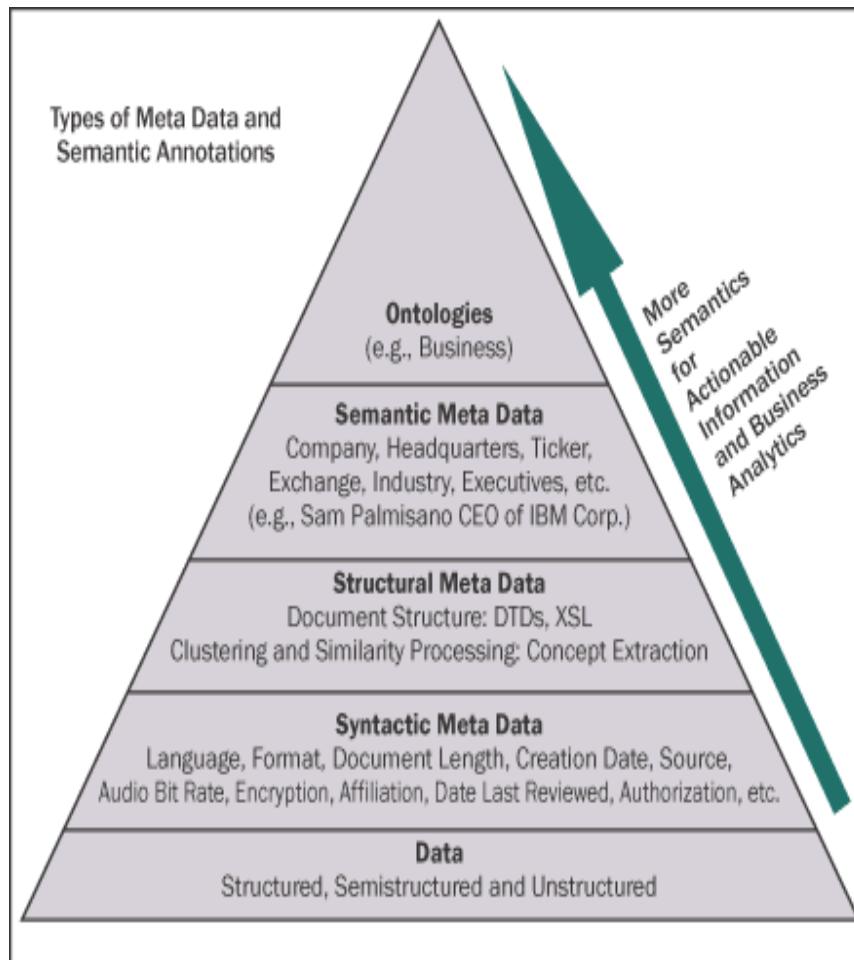


4. Querying the Ontology



3. Automatic aggregation of Knowledge

The Evolution of Meta Data

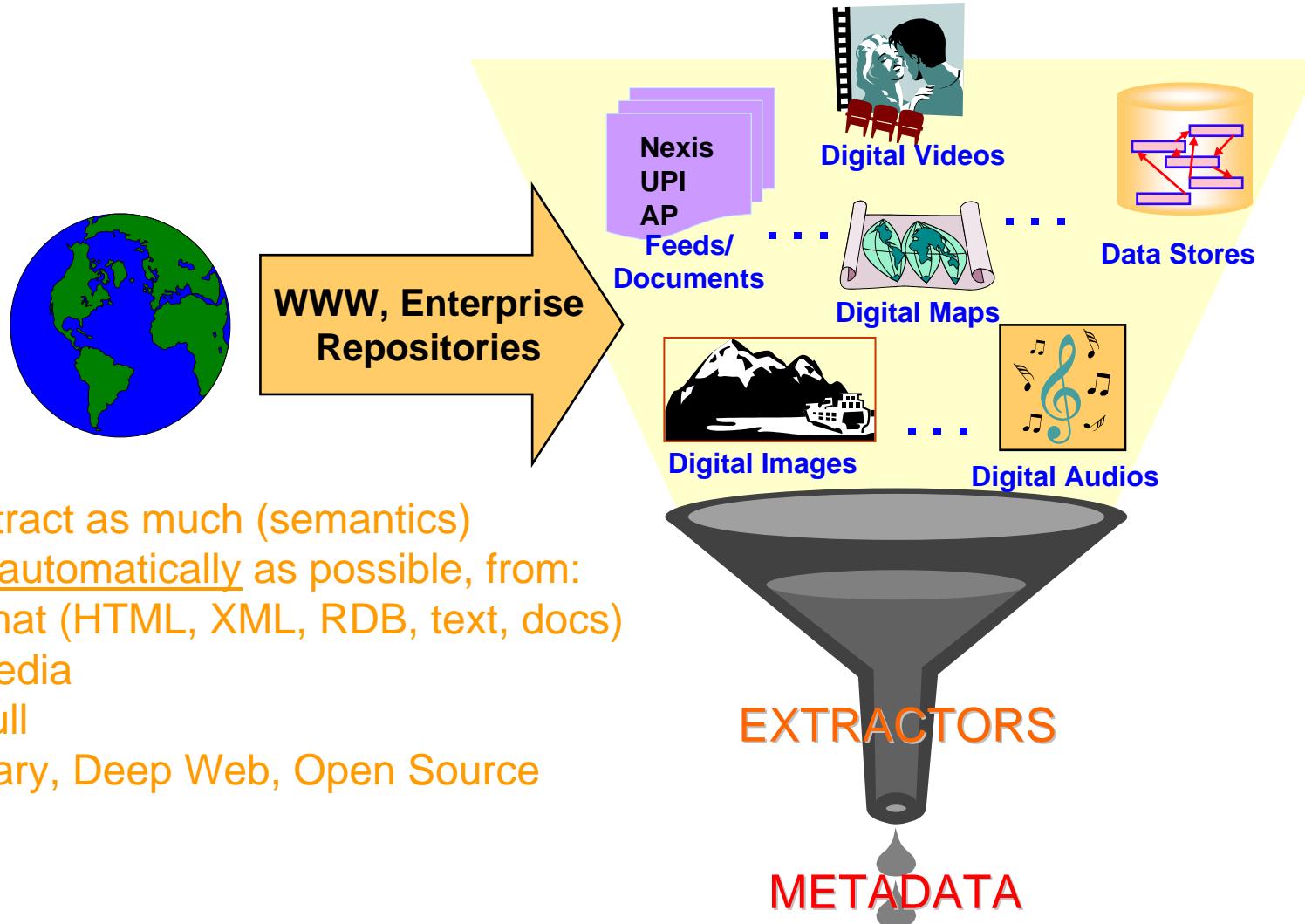


The more sophisticated technologies:

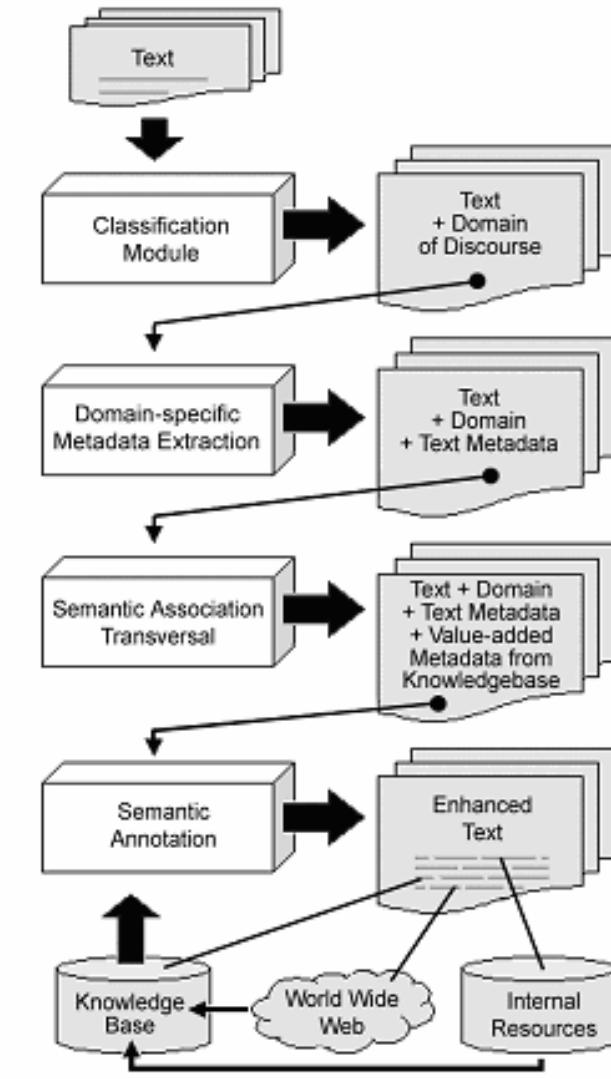
- Provide scalability and flexibility
- Handle all types of data (unstructured, semi-structured, structured)
- Accommodate SmartQuerying – flexible, intelligent querying
- Create SmartData – enhancing raw data with context and relationships
- Enable powerful enterprise decisionmaking



Metadata extraction from heterogeneous content/data



Metadata Extraction and Semantic Enhancement



Semantic Annotation/ Metadata Extraction

+

Enhancement

[Bancroft, Hammond, Sheth]

Blue-chip bonanza continues

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

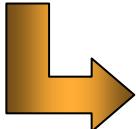
August 22, 2002: 11:44 AM EDT

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
Se
mc
del
els
Go
<p>S
loc
<p>M
Tin
co
sai
<p>G
</body>
</html>

```



```

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

```

The enhanced XML includes additional tags and values:

- Language**: `<Language FormalName="en" />`
- Property** (`urn:newsml:comtexnews.net:20010201:DomesticPublicCompanies:1`):
 - `<Property FormalName="Company" Value="Phillips Petroleum Co." />`
 - `<Property FormalName="StockSymbol" Value="P" />`
 - `<Property FormalName="Competitor">`
 - `<Property FormalName="CompanyName" Value="BP p.l.c." />`
 - `<Property FormalName="Competitor">`
 - `<Property FormalName="CompanyName" Value="Ultramar Diamond Shamrock Corp." />`
 - `<Property FormalName="Competitor">`
 - `<Property FormalName="CompanyName" Value="Royal Dutch/Shell Group" />`
 - `<Property FormalName="Headquarters" Value="Bartlesville, Oklahoma, United States of America" />`
 - `<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="Hornet, 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 FormalName="CompanyExecutive" Value="Turner, Kathryn C." />`
 - `<Property FormalName="CompanyPosition" Value="Director" />`
 - `- <Property FormalName="CompanyExecutive" Value="Meyers, Ph.D., Kevin" />`
 - `<Property FormalName="CompanyPosition" Value="Executive Vice President, Alaska Operations" />`
 - `- <Property FormalName="CompanyExecutive" Value="Lowe, John" />`
 - `<Property FormalName="CompanyPosition" Value="Senior Vice President, Planning and Strategic Transactions" />`
 - `- <Property FormalName="CompanyExecutive" Value="Mulva, J. J." />`
 - `<Property FormalName="CompanyPosition" Value="Chairman of the Board" />`
 - `<Property FormalName="CompanyPosition" Value="Chief Executive Officer" />`
 - `- <Property FormalName="CompanyExecutive" Value="Batchelder, E. L." />`
 - `<Property FormalName="CompanyPosition" Value="Vice President" />`
 - `<Property FormalName="CompanyPosition" Value="Chief Information Officer" />`
 - `- <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 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 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 America" />`
- `<Property FormalName="Sector" Value="Energy" />`
- `<Property FormalName="Industry" Value="Integrated Oil and Gas" />`

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

Customer Needs Driving Innovation

Horizontal Needs

Industry Needs	Content Exploitation	Knowledge Discovery	Competitive Advantage
	<ul style="list-style-type: none">•Understand and leverage siloed data•Increase worker productivity•Better KM across enterprises	<ul style="list-style-type: none">•Access/leverage universe of data•More accurate competitive/threat assessment	<ul style="list-style-type: none">•Outmaneuver competitors•Improve enterprise decision making•Less damage control

- Pharmaceuticals: intelligent literature search/mining, drug discovery
- Government and Intelligence
- Glycomics: semantic annotation of scientific (e.g., mass spectroscopy) data, complex processes with bioinformatics web services

VideoAnywhere and Taalee Semantic Search Engine (2000)

BLENDED BROWSING & QUERYING INTERFACE

The screenshot shows the VideoAnywhere interface. On the left, there's a sidebar with categories like All Assets, Business, Movie Assets, Film Festivals, Interviews, Movies, Reviews, News, and Travel. In the center, there's a search form with fields for Title (Alien), Contents, Classification (Movie Assets-Movies), Film Name, Cast, and Director. Below the form are 'Submit' and 'Clear' buttons. A yellow box labeled 'ATTRIBUTE & KEYWORD QUERYING' is overlaid on the right side of the search form.

SEMANTIC BROWSING

uniform view of worldwide distributed assets of similar type

Details

Title	Alien-Resurrection
Director	
Actors	Sigourney Weaver Winona Ryder
Contents	<p>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)</p>

Search Results				
Title	Details	View	Buy It	
Alien-Resurrection				
Alien Resurrection				

Targeted e-shopping/e-commerce

The screenshot shows the Amazon.com search results for "Alien+Resurrection+". The search bar at the top has "Alien+Resurrection+" entered. The results page shows a list of items, with the first item being "Alien Resurrection (1997) -- Sigourney Weaver, Winona Ryder; VHS". There are filters on the left for VHS, DVD, and All Formats. A yellow box labeled "Top matches for this search:" is overlaid on the right side of the results.

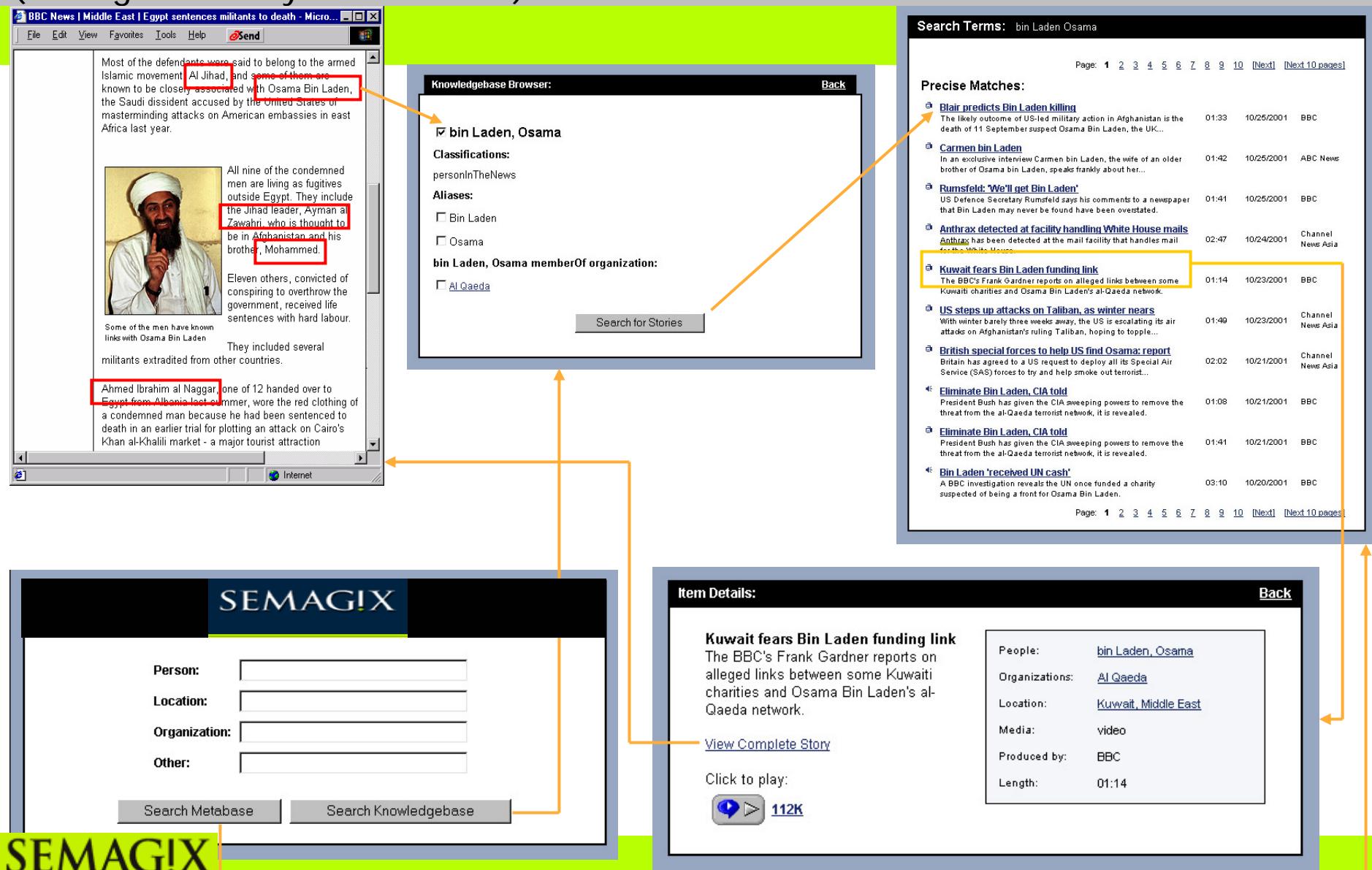
Full Results: 5 items are shown below.

[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

The screenshot shows a window titled "low_alienrest.asx - Windows Media Player". It displays a video frame of the movie "Alien Resurrection". The bottom status bar shows "Playing" and "02:22 / 02:31". The window includes standard media controls (play, stop, forward, backward) and a "Showcase" button. A yellow box labeled "assets access" is overlaid on the top right of the window.

Blended Semantic Browsing and Querying (Intelligence Analyst Workbench): 2002



Visualizer Content: BSBQ Application

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 subcutaneously to patients with solid tumors receiving chemotherapy.

Erythropoietic agents as therapy

Erythropoietin is the primary physiologic regulator of red blood cell production. It exerts its effect by binding to cell surface receptors. Erythropoietin has been shown that both erythropoietin and its receptor are expressed on the surface of normal and malignant cells.

Epoetin alfa: current and future applications

Cancer-related anemia commonly is associated with fatigue and decreased quality of life (QOL). Treatment to achieve hemoglobin levels similar to those of healthy patients receiving chemotherapy can ameliorate symptoms of anemia.

Pure Red-Cell Aplasia and Related Disorders

To the Editor: Casadevall et al. (Feb. 1) report that patients with pure red-cell aplasia and antierythropoietin antibodies who received recombinant erythropoietin (EPO) had a dramatic increase in their red-cell counts.

Role of oral versus IV iron supplementation in cancer patients

BACKGROUND: Preoperative treatment with oral iron (Ferinject, Janssen-Cilag; or PROCRIT, Ortho Biotech Products, L.P.) before surgery increases the erythropoietin response to anemia in patients with cancer.

Erythropoietin (Procrit; EpoGen)

(no description available)

Role of iron in optimizing response to erythropoietin

Approximately 50% of cancer patients have anemia despite an available treatment option for these patients. In the early 1980s, recombinant human erythropoietin (rHuEPO, epoetin alfa [Epogen,

Zoom Find Entity Classes Instances

Content Details

Epoetin alfa: current and future indications and nursing implications.

Cancer-related anemia commonly is associated with fatigue and

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

SEMAGIX
POWER • THROUGH • RELEVANCE

Semantic Information Integration in Portals

The screenshot shows a web browser displaying the KnowledgePlex portal at <http://beta.knowledgeplex.org/topic.html?c=246>. The page is titled "Workforce Development". A red box highlights a news item titled "Trends in Naturalization" from Fix, Passel, and Sucher, dated Sep 17, 2003. Another red box highlights a discussion post titled "New Sources Of Development Finance?" from Affordable Housing Development & Finance, dated Sep 11 20:29. A third red box highlights a document titled "Fundraising: Multiple Mail Appeals" from www.nhi.org, dated Sep 14, 2003. The right sidebar displays a user profile for "Yash Warke" with topics like "Workforce Development" and "Public". A legend on the right side of the slide maps these visual elements to concepts:

- 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

Equity Research Dashboard with Blended Semantic Querying and Browsing

The screenshot shows a Microsoft Internet Explorer window titled "Voquette Equity Research Dashboard - Microsoft Internet Explorer". The main content area displays a stock profile for "Motorola, Inc." (MOT) with the following details:

- Symbol:** MOT
- Change:** +0.61
- Price:** 15.53
- Volume:** 8,935,500
- Equity Indices:** DOW 8,570.15, NASDAQ 1,493.15, S&P 500 1,002.75

Below this are two line charts: "Motorola Incorporated (USD) Price" and "Motorola Incorporated (USD) Volume".

At the bottom of the main content area, there are links for "View competitors..." and "Listen to audio programs...".

The right side of the dashboard features a news feed organized into semantic categories:

- Company News:**
 - 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:**
 - CSFB sees quality in Qualcomm (09/10/2001, CBS Marketwatch)
 - Merrill Dowgrades Motorola to 'Near-...' (09/06/2001, BusinessWeek Online)
 - Motorola (09/06/2001, ON24)
- Earnings News:**
 - 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:**
 - 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:**
 - 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:**
 - 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 Agree... (08/14/2001, COMTEX)

At the bottom right, it says "Powered By VOQUETTE".

Automatic
3rd party
content
integration

Competitive
research
inferred
automatically

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

Semantic Integration of Heterogeneous Data for AML application

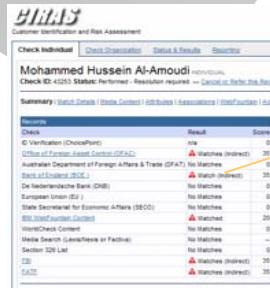


Structure Government Data

Establishing New Account



Scores the entity based on the content and entity relationships



Un-structure text, Semi-structured Data

Snippets is displayed base on best match to entities



User will be able to navigate the ontology using a number of different interfaces

CIRAS KYC Check

Individual Check - Microsoft Internet Explorer

File Edit View Favorites Tools Help

CIRAS
Customer Identification and Risk Assessment

User: Larry Parker | Sign Off

[Check Individual](#) **Check Organization** [Status & Results](#) [Reporting](#)

LMZ Shipping, LLC ENTITY
Check ID: 14092 Status: Pending — Cancel or Refer this Record

Score: 65 | [Print \(PDF\)](#)
Brief explanation of score...

[Summary](#) | [Match Details](#) | [Media Content](#) | [Attributes](#) | [Associations](#) | [WebFountain](#)

Records		
Check	Result	Score
ID Verification (ChoicePoint)	Verified	0
Office of Foreign Asset Control (OFAC)	⚠ Match	20
Australian Department of Foreign Affairs & Trade (DFAT)	No Matches	0
Bank of England (BOE)	⚠ Match	0
De Nederlandsche Bank (DNB)	No Matches	0
European Union (EU)	No Matches	0
State Secretariat for Economic Affairs (SECO)	No Matches	0
CPB Global Name Database ("Swiss Check")	No Matches	0
WorldCheck Content	⚠ Match	--
Media Search (Lexis/Nexis or Factiva)	No Matches	--
Section 326 List	No Matches	0

Perform a New Check

Visualizer Content: BSBQ Application

Semantic Visualizer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Zoom ENRON Find Entity Classes Instances

CITIGROUP INC.

Synonyms Classifications

CITIGROUP INC. Organization

Attributes

name	value
Countries	USA
Location	New York
WC Category	CORPORATE
Display Name	CITIGROUP INC.

Related Content

- ANSCHUTZ, Philip F.
- CITIGROUP INC.
- REED, John Shepard
- WEILL, Sanford I
- CITIGROUP GLOBAL MARKETS INC.

BELDEN, Timothy

Content Details

Title: ANSCHUTZ, Philip F.

Description: Sept - 2002 charged by NY Attorney General with illegally profiting from the now banned practice of spinning in 57 IPOs arranged by Citigroup. May - 2003 agreed to pay \$4.4 million to settle charges without admitting or denying guilt. Californian Supreme Court has rejected his bid to stop an action brought by the California State Teachers Retirement System, which claims to have lost \$200m in stock and bonds invested in Qwest, an Anschutz firm. He is now likely to face trial and could be held personally liable in a securities fraud case. He also

WC Category: INDIVIDUAL

Entity: WC33640

URLs: <http://albany.bizjournals.com/albany/stories/2003/05/12/daily2>

Java Applet Window

SEMAGIX
POWER • THROUGH • RELEVANCE

Copyright © 2004, Semagix Ltd.

CIRAS KYC Check Associations

Individual Check - Microsoft Internet Explorer

File Edit View Favorites Tools Help

User: Larry Parker | Sign Off

CIRAS
Customer Identification and Risk Assessment

[Check Individual](#) **Check Organization** [Status & Results](#) [Reporting](#)

LMZ Shipping, LLC ENTITY
Check ID: 14092 Status: Pending — [Cancel or Refer this Record](#)

Score: 65 | [Print \(PDF\)](#)
Brief explanation of score...

[Summary](#) | [Match Details](#) | [WebFountain](#) | [Media Content](#) | [Attributes](#) | **Associations**

[Launch Associations Visualizer](#)

Is Related to	Wojtech Moroski Wojtech Moroski is related to Rabbita Trust Rabbita Trust appears on FBI Watch List
Undertakes	12 Thompson Av, Athens, Georgia
Is related to	Wojtech Moroski
Active in	Washington
Active in	US

[Perform a New Check](#)

SEMAg!X
POWER • THROUGH • RELEVANCE

[Terms, conditions, caveats and small print](#)

© Copyright 2004. Semagix. All Rights Reserved.

View Risk Scores for a specific company or customer

Screenshot of a Microsoft Internet Explorer window showing the CIRAS system interface.

The main window title is "Transaction and Customer List - Microsoft Internet Explorer".

The URL in the address bar is "http://www.ciras.com/ciras/transactionlist.asp?Company=STATOIL+GAS+TRADING+LTD&Action=View".

The page displays "Client Information" for STATOIL GAS TRADING LTD, with fields for Company, Individual, Location, and Other.

A "Risk Score" section shows a score of 65 for STATOIL GAS TRADING LTD and an Aggregate score of 65. A link "Details" is present.

An "Accept" and "Reject" button are visible.

The "Company Knowledge" section lists STATOIL GAS TRADING LTD [Company] and provides links to "Score Breakdown" and "Know Your Customer Check".

The "Score Breakdown" window title is "Score Breakdown - Microsoft Internet Explorer". It shows a breakdown of the risk score:

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

The "Aggregate Score: 65" is also displayed.

The "Know Your Customer Check" section includes:

- Retrospective Check
- Application Date: 29/01/2003
- Request Outcome: Failed

At the bottom, there are navigation buttons: Retrospective Client Check, All Transactions, Investigate Client, Investigate Transaction, and New.

The status bar at the bottom shows the date and time: 10:43 AM Friday 8/1/2003.



European Law Enforcement Agency – Case Study

To identify and target prolific offenders by creating a holistic view of crime incident data and related entities

Needs / Business Requirements

- Merge and link case data from multiple sources using effective identification and disambiguation of entities associated with cases
- Link cases to a taxonomy of modus operandi based on analysis of unstructured textual information such as witness statements and crime reports
- Ability to use pre-defined or investigation-specific case profiles for search and match, with matching determined by configurable scoring of case attributes and associated entities
- Positive and negative searching of cases based on presence /absence of key characteristics or behaviours
- Ability to explore and browse case data starting from any specific associated entity via link analysis and intelligent entity identification and annotation in supporting unstructured textual documents

System Requirements

- Several hundred users
- Integrate with enterprise single sign-on systems

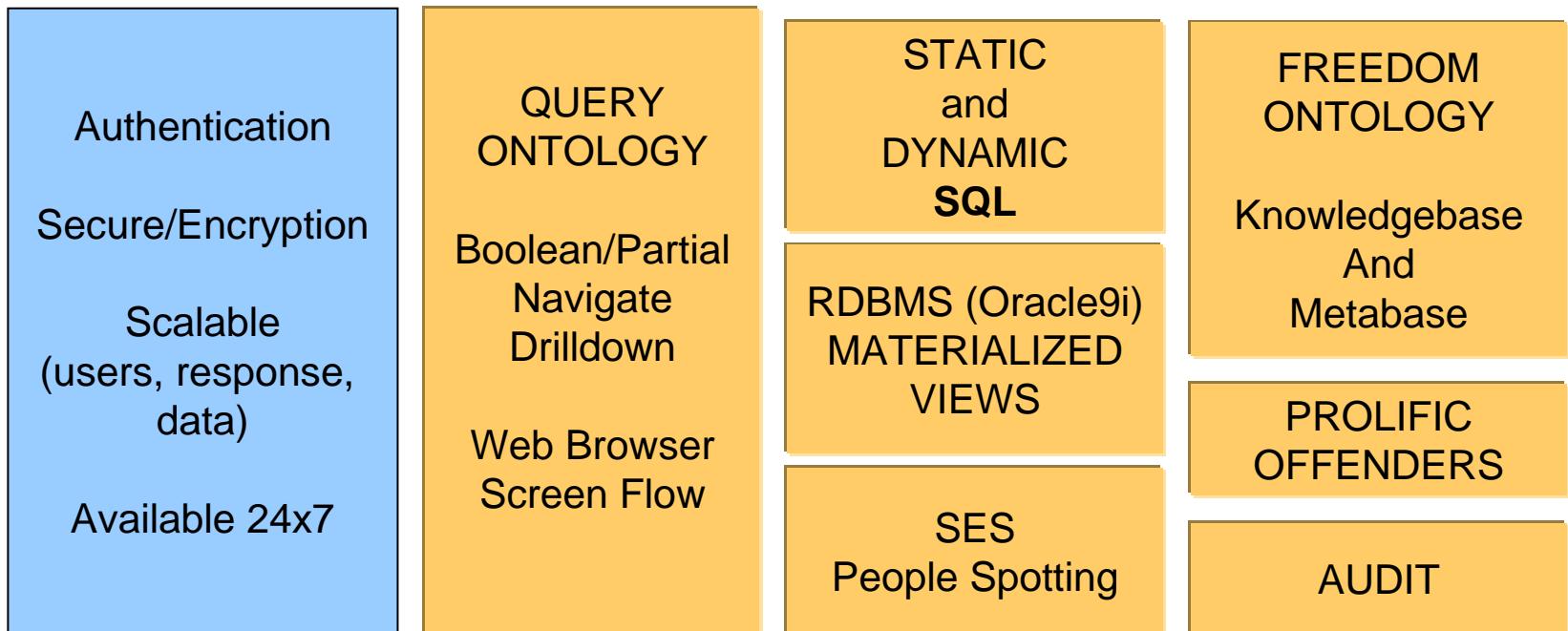
Technical Challenges

- Handle extremely large scale datasets – in excess of 10 million incidents – with a very high rate of daily incremental update
- Enforce visibility and privacy rules appropriate to the user and the available data sets
- Identify case behaviours / modus operandi in unstructured text

Results

- Superior and more timely identification of prolific officers
- Better prioritization of cases
- Greater investigator productivity and effectiveness

European Law Enforcement Agency – Architecture Overview



SEMAGIX FREEDOM

APACHE Web Server, TOMCAT Web/Java App Server, Oracle9i
(DHTML, CSS, JavaScript, JAVA, JSP, XML, XSL, SQL)
(HTTP, HTTPS, SSL, JDBC, OCI, FCGI)

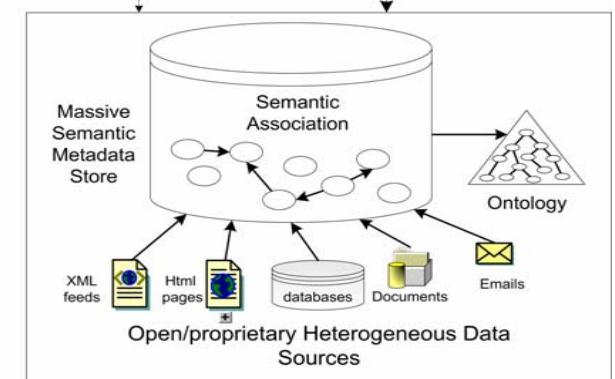
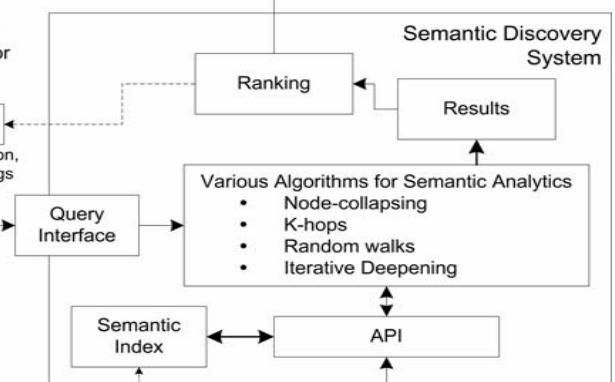
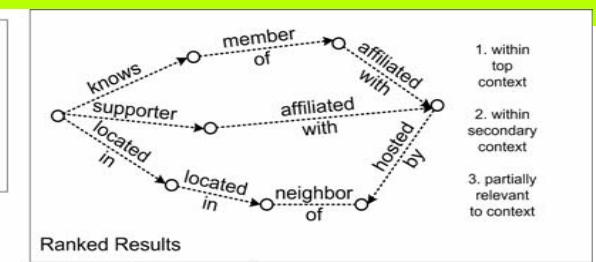
Red Hat Linux Enterprise Server

INTEL 32bit HARWARE



SemDIS Prototype: Relationship Discovery

- SemDIS project
 - Discover relationships between entities in semantic knowledge bases
 - For prototype, use a subset of SWETO as a dataset





Turn Ranking On

Semantic Association Query Engine - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Mail Print Find People

Address http://vader.cs.uga.edu:8080/semdis/matches Go Google Links

LSDIS Large Scale Distributed Information Systems University of Georgia Computer Science Department

Select Entities of Interest

More than one class found.
Select best match from list below:

Entity one: Entity two:

Enter a k value for the search:

Ranking on:

Submit Query

Done Internet



Configure Ranking Schema with Context in ‘Colleges’ and ‘Academic Depts.’

Semantic Association Query Engine - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media Print Links

Address http://vader.cs.uga.edu:8080/semdis/search?psel1=http%3A%2F%2Flsdis.cs.uga.edu%2Fproj%2Fsemdis%2Ftestbed%2F%23SWEET_215003&psel2=http%3A%2F%2Flsdis.cs.uga.edu%2Fproj%2Fser Google Links

Configure Ranking Criteria

Context Specification

Subsumption Adjustment k: .1

Trust Adjustment k: .1

Rarity Adjustment k: .1

Favor Rare Associations:

Popularity Adjustment k: .1

Favor Popular Associations:

Association Length Adjustment k: .1

Favor Long Associations:

Submit Query

University

Mathematics Department

Computer Science Department

Statistics Department

Academic Department

Computer Science Researcher

Researcher

listed author in>

has academic department>

rdfs:subClassOf

faculty member at>

rdi: Expand Node

Hide Node

Select Node

Add to Region 'College Depts.'

Subsumption Adjustment

Trust Adjustment

Rarity Adjustment

Favor Rare Associations:

Popularity Adjustment

Favor Popular Associations:

Association Length Adjustment

Favor Long Associations:

Submit Query

Applet com.touchgraph.linkbrowser.LinkBrowserApplet started

Internet



Returns Ranked Results

Semantic Association Query Engine - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print Mail Links

Address http://vader.cs.uga.edu:8080/semdis/ranker Go Google Links

LSDIS
Large Scale Distributed Information Systems

University of Georgia
Computer Science Department

Associations Found

Results 1 - 10 of 47. Search took: 4.389 seconds

Association	Ranking Score	Context	Association Length	Subsumption	Trust	Rarity	Popularity
1. Chee-Keng Yap <i>faculty_member_at</i> New York University Department of Computer Science <i>has_academic_department</i> New York University <i>located_in</i> New York <i>located_in</i> Columbia University <i>has_academic_department</i> Columbia University Department of Computer Science <i>faculty_member_at</i> Ravi Ramamoorthi	0.3756270205836345	[bar]	[bar]	[bar]	[bar]	[bar]	[bar]
2. Chee-Keng Yap <i>elisted_author_in</i> Refinement Methods for Geometric Bounds in Constructive Solid Geometry. <i>published_in</i> ACM Trans. Graph. <i>published_in</i> Frequency space environment map rendering. <i>elisted_author_in</i> Ravi Ramamoorthi	0.2538365896668301	[bar]	[bar]	[bar]	[bar]	[bar]	[bar]
3. Chee-Keng Yap <i>elisted_author_in</i> Refinement Methods for Geometric Bounds in Constructive Solid Geometry. <i>published_in</i> ACM Trans. Graph. <i>published_in</i> Chromium: a stream-processing framework for interactive rendering on clusters. <i>elisted_author_in</i> Ren Ng <i>elisted_author_in</i> All-frequency shadows using non-linear wavelet lighting approximation. <i>elisted_author_in</i> Ravi Ramamoorthi	0.25343627662676194	[bar]	[bar]	[bar]	[bar]	[bar]	[bar]
4. Chee-Keng Yap <i>elisted_author_in</i> On k-Hulls and Related Problems. <i>published_in</i> SIAM J. Comput. <i>published_in</i> Ranking Algorithms: The Symmetries and Colorations of the n-Cube. <i>elisted_author_in</i> Jay P. Fillmore <i>elisted_author_in</i> Spherical averages and applications to spherical splines and interpolation. <i>published_in</i> ACM Trans. Graph. <i>published_in</i> Frequency space environment map rendering. <i>elisted_author_in</i> Ravi Ramamoorthi	0.2533669312668104	[bar]	[bar]	[bar]	[bar]	[bar]	[bar]
5. Chee-Keng Yap <i>elisted_author_in</i> On k-Hulls and Related Problems. <i>published_in</i> SIAM J. Comput. <i>published_in</i> On Backtracking: A Combinatorial Description of the Algorithm. <i>elisted_author_in</i> Jay P. Fillmore <i>elisted_author_in</i> Spherical averages and applications to spherical splines and interpolation. <i>published_in</i> ACM Trans. Graph. <i>published_in</i> Frequency space environment map rendering. <i>elisted_author_in</i> Ravi Ramamoorthi	0.2533669312668104	[bar]	[bar]	[bar]	[bar]	[bar]	[bar]
6. Chee-Keng Yap <i>elisted_author_in</i> Reversal Complexity. <i>published_in</i> SIAM J. Comput. <i>published_in</i> Ranking Algorithms: The Symmetries and Colorations of the n-Cube. <i>elisted_author_in</i> Jay P. Fillmore <i>elisted_author_in</i> Spherical averages and applications to spherical splines and interpolation. <i>published_in</i> ACM Trans. Graph. <i>published_in</i> Frequency space environment map rendering. <i>elisted_author_in</i> Ravi Ramamoorthi	0.2533669312668104	[bar]	[bar]	[bar]	[bar]	[bar]	[bar]
7. Chee-Keng Yap <i>elisted_author_in</i> Reversal Complexity. <i>published_in</i> SIAM J. Comput. <i>published_in</i> On Backtracking: A Combinatorial Description of the Algorithm. <i>elisted_author_in</i> Jay P. Fillmore <i>elisted_author_in</i> Spherical averages and applications to spherical splines and interpolation. <i>published_in</i> ACM Trans. Graph. <i>published_in</i> Frequency space environment map rendering. <i>elisted_author_in</i> Ravi Ramamoorthi	0.2533669312668104	[bar]	[bar]	[bar]	[bar]	[bar]	[bar]
8. Chee-Keng Yap <i>elisted_author_in</i> Precision-Sensitive Euclidean Shortest Path in 3-Space. <i>published_in</i> SIAM J. Comput. <i>published_in</i> Ranking Algorithms: The Symmetries and Colorations of the n-Cube. <i>elisted_author_in</i> Jay P. Fillmore <i>elisted_author_in</i> Spherical averages and applications to spherical splines and interpolation. <i>published_in</i> ACM Trans. Graph. <i>published_in</i> Frequency space environment map rendering. <i>elisted_author_in</i> Ravi Ramamoorthi	0.2533669312668104	[bar]	[bar]	[bar]	[bar]	[bar]	[bar]
9. Chee-Keng Yap <i>elisted_author_in</i> Precision-Sensitive Euclidean Shortest Path in 3-Space.							

Ontology Quality

- Many real-world ontologies may be described as semi-formal ontologies
 - populated with partial or incomplete knowledge
 - may contain occasional inconsistencies, or occasionally violate constraints (e.g. all schema level constraints may not be observed in the knowledgebase that instantiates the ontology schema)
 - often ontology is populated by many persons or by extracting and integrating knowledge from multiple sources
 - analogy is “dirty data” which is usually a fact of life in most enterprise databases.

Ontology Representation Expressiveness

- Applications vary in terms of expressiveness of representation needed.
- Trade-off between expressive power and computational complexity applies both to knowledge creation/maintenance and to inference mechanisms for such languages. It is often very difficult to capture the knowledge that instantiates the more expressive constructs/constraints.
- Many business applications end up using models/languages that lie closer to less expressive languages.
- On the other hand, we have seen a few applications, especially in scientific domains such as biology, where more expressive languages are needed, and OWL-Full or FOL is not adequate.

Ontology Size / Population / Freshness

- Ontology population is critical. Among the ontologies developed by Semagix or using its technology, a median size of ontology is over 1 million instances/facts and relationship instances each (at least two have exceeded 10 million instances). This level of knowledge makes the system very powerful (as it is applied . Furthermore, in many cases, it is necessary to keep these ontologies current or updated with facts and knowledge on a daily or more frequent basis. Both the scale and freshness requirements dictate that populating ontologies with instance data needs to be automated.

Metadata Extraction

Large scale metadata extraction and semantic annotation is possible. IBM WebFountain [Dill et al 2003] demonstrates the ability to annotate on a Web scale (i.e., over 2.5 billion pages), while Semagix Freedom related technology [Hammond et al 2002] demonstrates capabilities that work for a few million documents per day per server. However, the general trade-off of depth versus scale applies. Storage and manipulation of metadata for millions to hundreds of millions of content items requires database techniques with the challenge of improving performance and scale in presence of more complex structures

Semantic Technology Building Blocks

- A vast majority of the Semantic (Web) Technology Applications that have been developed or envisioned rely on three crucial capabilities: **ontology creation, semantic annotation (metadata extraction) and querying/inferencing**. Enterprise-scale applications share many requirements in these three respects with pan Web applications. All these capabilities must scale to many millions of documents and concepts (rather than hundreds to thousands) for current applications, and applications requiring billions of documents and concepts have also been discussed (esp. in intelligence and government space) but not yet deployed.

Primary Technical Capabilities/Key Research Challenges

- Two of the most basic “semantic” techniques are “named entity identification”, and “semantic ambiguity resolution”. [It would be nice to have relationship extraction too.] A tool for annotation is of little value if it does not support ambiguity resolution. Both require highly multidisciplinary approaches, borrowing from NLP/lexical analysis, statistical and IR techniques and possibly machine learning techniques. A high degree of automation is possible in meeting many real-world semantic disambiguation requirements, although pathological cases will always exist and complete automation is unlikely.

Content Heterogeneity

- Support for heterogeneous content is key – it is too hard to deploy separate products within a single enterprise to deal with structured, semi-structured and unstructured data/content management. New applications involve extensive types of heterogeneity in format, media and access/delivery mechanisms (e.g., news feed in RSS, NewsML news, Web posted article in HTML or served up dynamically through database query and XSLT transformation, analyst report in PDF or WORD, subscription service with API-based access to Lexis/Nexis, enterprise's own relational databases and content management systems such as Documentum or Notes, e-mails, etc). Semi-structured data (XML-based data and RDF based metadata) is growing at an explosive rate.

Processing

- Semantic query processing with the ability to query both ontology and metadata to retrieve heterogeneous content is highly valuable. Consider “Give me all articles on the competitors of Intel”, where ontology gives information on competitors, supports semantics (with the understanding that “Palm” is a company and that “Palm” and “Palm, Inc.” are the same in this case), and metadata identifies the company to which an article refers, regardless of format of the article.
- Analytical applications could require sub-second response time for tens of concurrent complex queries over a large metadata base and ontology, and can benefit from further database research. High performance and highly scalable query processing techniques that deal with more complex representations compared to database schemas and with more explicit roles of relationships, is important. Have not found great use of DL reasoning.



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
 - Standards defined by W3C are very timely and are bringing some level of interoperability
- No longer an exotic technology
 - beyond proof of concept and now facing main stream engineering challenges
 - some industries very open to ontologies
 - in other industries applications hide semantic technologies



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 difficult research and engineering challenges ahead
- Researchers should be mindful of state of the art commercial technologies and real applications

For Further Information

- Article in Data Engineering special issue on Making the Semantic Web Real (Dec. 2003)
<http://wwwt.semagix.com/documents/SemanticWebTechinAction.pdf>
- Commercial Technology: <http://www.semagix.com/download.html>
- Research: Semantic Association and Semantic Discovery Projects:
<http://lsdis.cs.uga.edu/proj/proj.html>
- Publications and Presentations: <http://lsdis.cs.uga.edu/lib/lib.html>