Semantic Web Technology Evaluation Ontology (SWETO): A test bed for evaluating tools and benchmarking semantic applications

WWW2004 (New York, May 22, 2004)
Semantic Web Track, Developers Day

Boanerges Aleman-Meza, Amit Sheth,
I. Budak Arpinar, Chris Halaschek
Large Scale Distributed Information Systems Lab
University of Georgia

This material is based upon work supported by the National Science Foundation
under NSF-ITR-IDM Award # 0325464 and NSF-ITR-IDM Award # 0219649.

Motivation for SWETO

- Many new techniques and software tools from emerging Semantic Web (SW) community
  - Need a common infrastructure for testing
- Need of an open and freely available ontology with a very large knowledge base
  - Scalability testing as the most important objective
  - Quality and comparability as other criteria

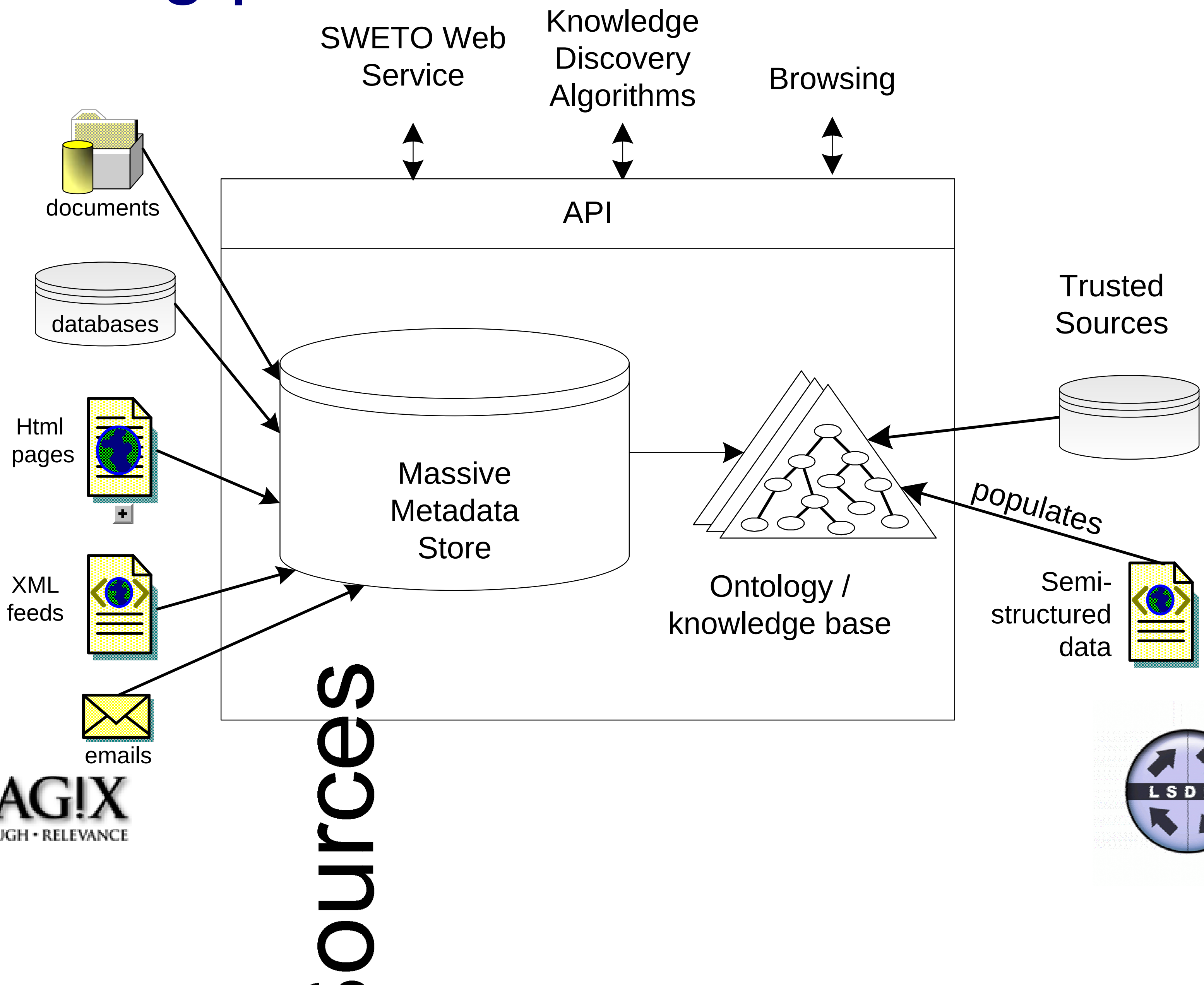
SEMAG!X
POWER • THROUGH • RELEVANCE

LSDIS

SWETO Objectives

- Develop a broad and deep ontology populated with real facts/data from real world heterogeneous sources
  - the instances in the knowledge base should be highly interconnected
- Serve as a test-bed for advanced semantic applications (i.e. business intelligence, national security, etc.)
- Address the requirements of a research benchmark for semantic analytics, and the semantic techniques of:
  - ontology creation
  - semi-automatic extraction
  - entity disambiguation

The big picture

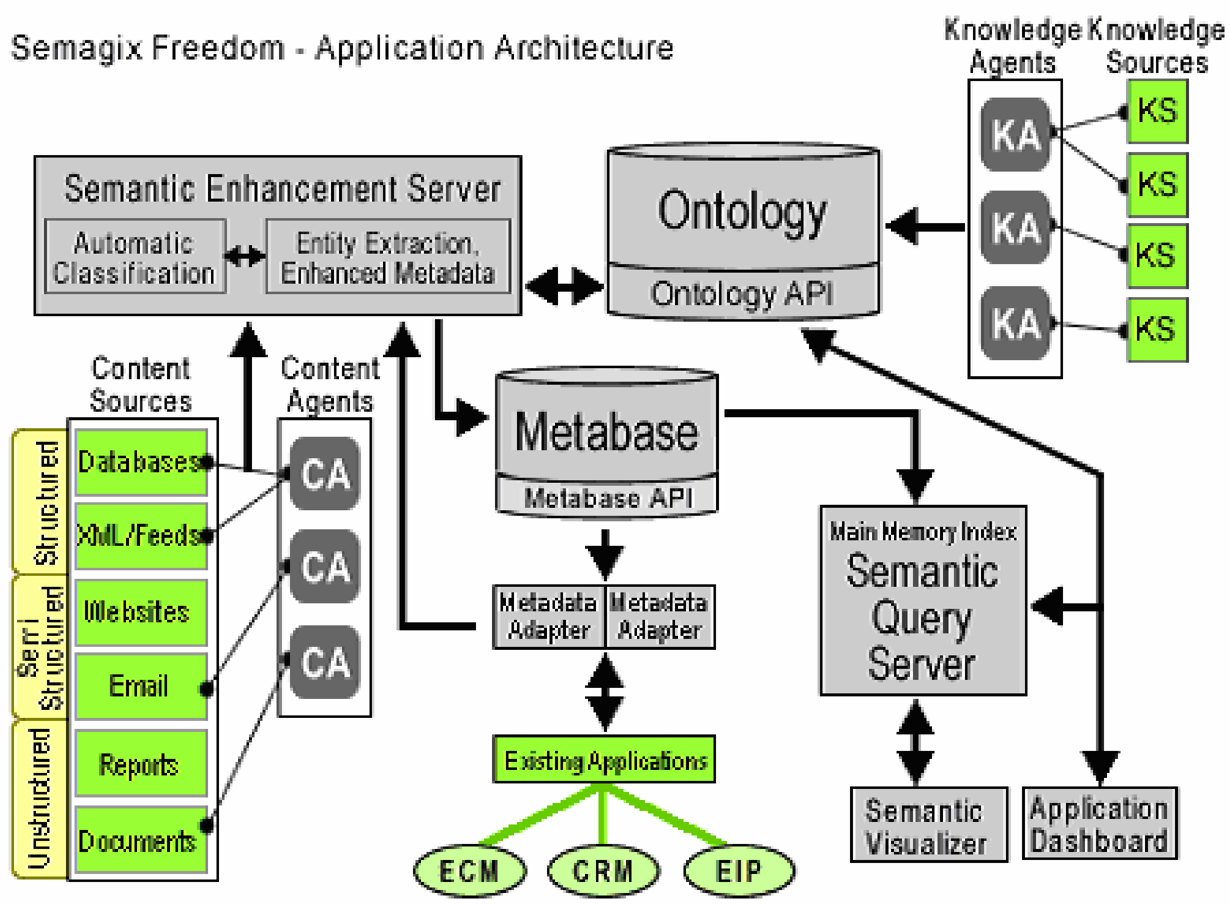


SEMAG!X
POWER · THROUGH · RELEVANCE

LSDIS

Semagix Freedom Architecture

Utilized Semagix Freedom for SWETO ontology design and population



SEMAG!X
POWER · THROUGH · RELEVANCE

LSDIS

Development Framework

- Utilized Semagix Freedom for ontology design and population
- With Freedom, knowledge extractors were created to extract entities from various data sources

Development Framework

- Data sources:
  - Selected sources which were highly reliable Web sites that provide entities in a
    - semi –structured format
    - unstructured data with parse-able structures (e.g.,html pages with tables)
    - dynamic web sites with database back-ends
  - Considered the types and quantity of implicit/explicit relationships
    - preferred sources in which instances were interconnected
  - considered sources whose entities would have rich metadata
  - Public and open sources were preferred
    - due to the desire to make SWETO openly available

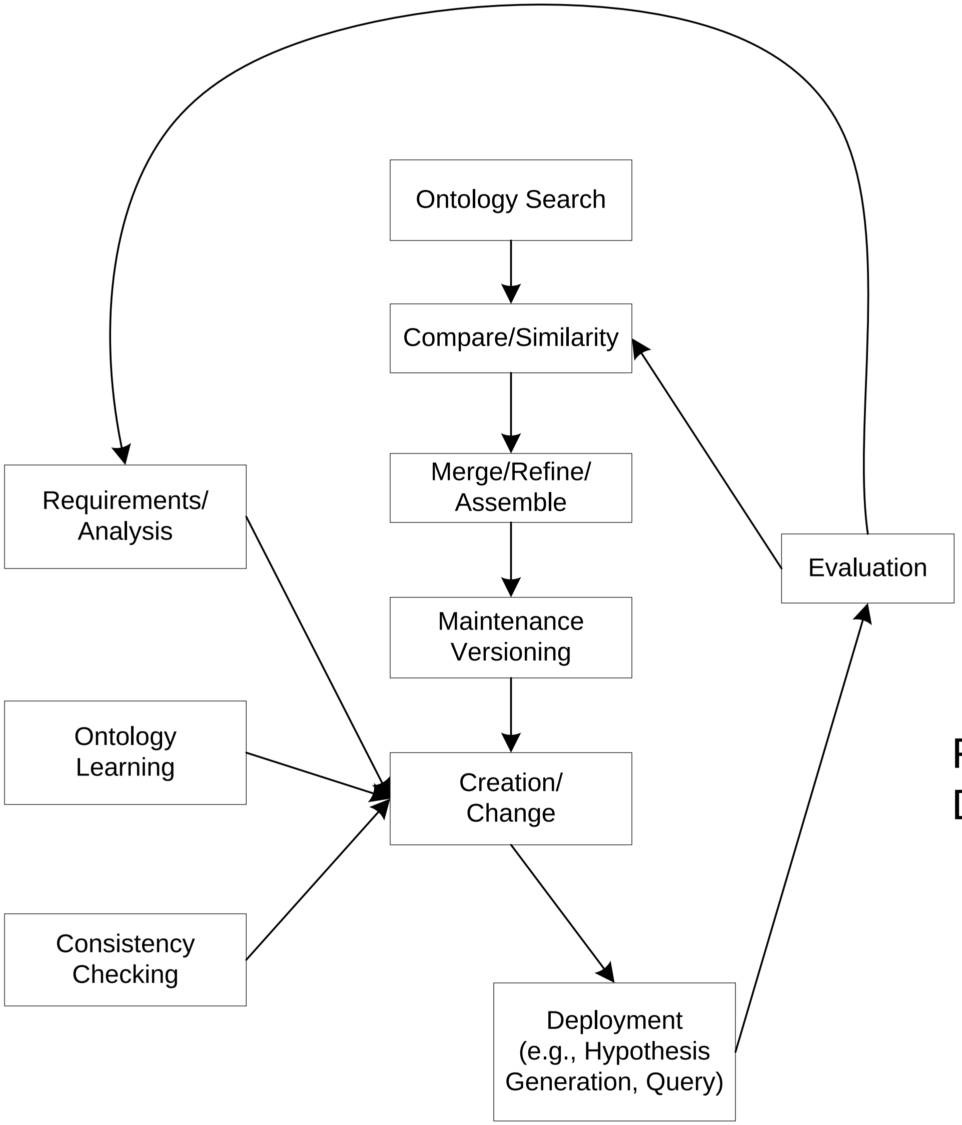
Development Framework

- As the sources are processed by the extractors, entities are extracted and stored in appropriate classes in an ontology
- Due to heterogeneous data sources, entity disambiguation is a crucial step
  - □ Freedom's disambiguation techniques automatically resolved entity ambiguities in 97% of the cases, leaving the rest for human disambiguation (and may be ignored)

Development Framework

- Utilize Freedom's API for exporting both the ontology and its instances in either RDF [5] or OWL [2] syntax
- Extractors are scheduled to rerun for keeping the ontology updated

SWETO Life-cycle



From Amicalola report on DB/IS and SW
[Sheth and Meersman 2002]

SEMAG!X
POWER · THROUGH · RELEVANCE

LSDIS

Current Status

- V.1 population includes 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
- Significant information for provenance/trust support [UMBC partnership]

Current Status – Classes

Subset of classes in the ontology	**# Instances**
Cities, countries, and states	2,902
Airports	1,515
Companies, and banks	30,948
Terrorist attacks, and organizations	1,511
Persons and researchers	307,417
Scientific publications	463,270
Journals, conferences, and books	4,256
TOTAL (as of January 2004)	811,819

SEMAG!X
POWER · THROUGH · RELEVANCE

LSDIS

Current Status – Relationships

Subset of relationships	# Explicit relations
located in	30,809
responsible for (event)	1,425
Listed author in	1,045,719
(paper) published in	467,367

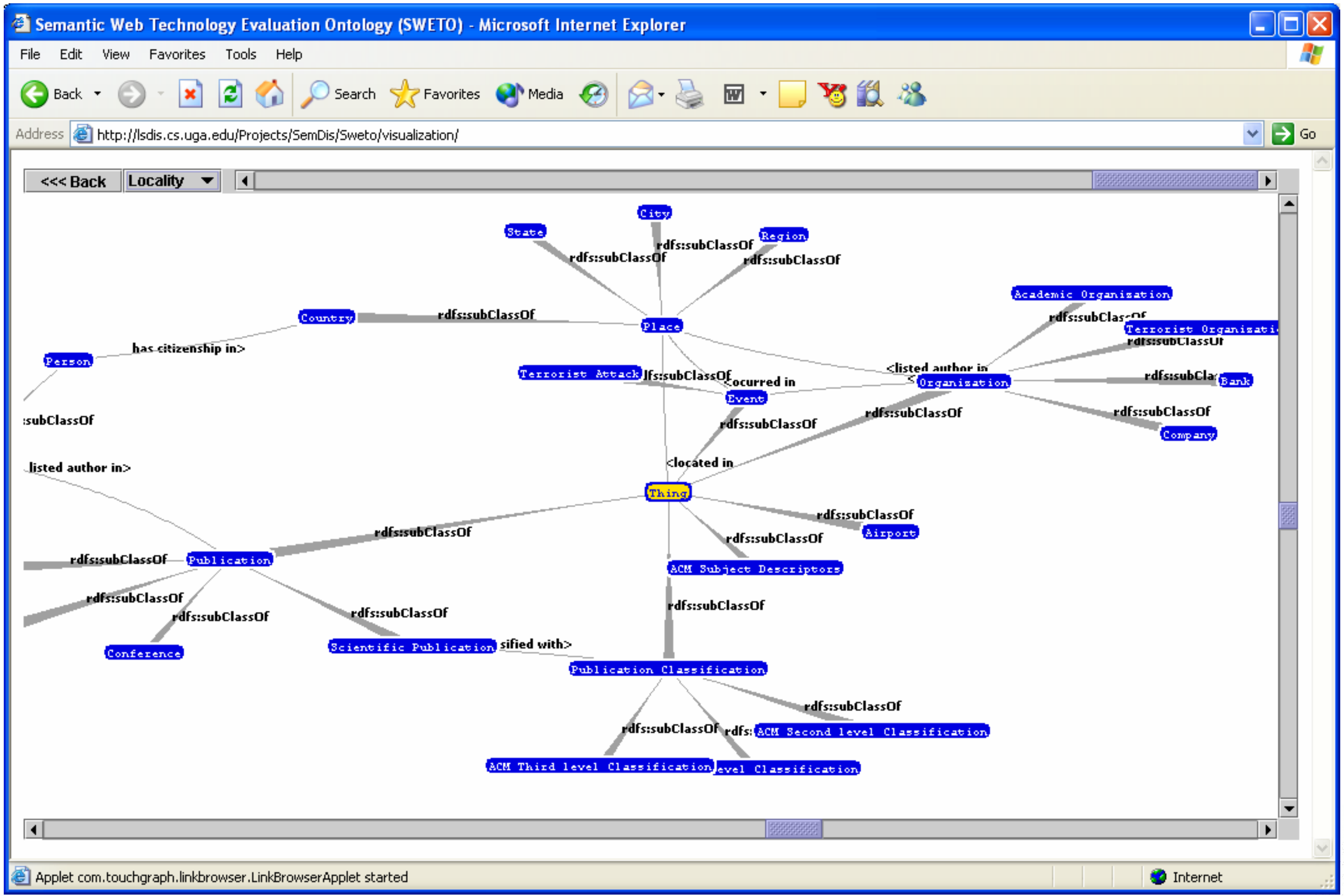
Current Status – Disambiguation

Disambiguation type	# Times used
Automatic (Freedom)	248,151
Manual	210
Unresolved (Removed)	591

SEMAG!X
POWER · THROUGH · RELEVANCE

LSDIS

Browsing of the Schema



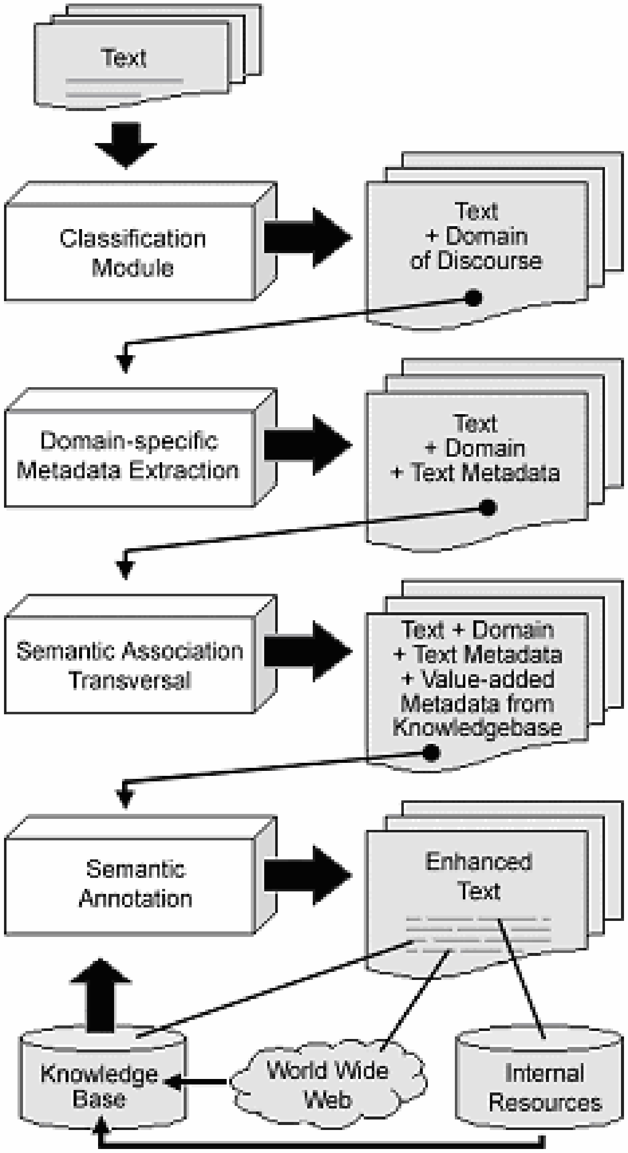
SEMAG!X
POWER · THROUGH · RELEVANCE

LSDIS

Evaluation/ Usage 1: Industry

- Evaluation of Fast Semantic Enhancement

(in Marianas SDK)


Text
Classification Module
Text + Domain of Discourse
Domain-specific Metadata Extraction
Text + Domain + Text Metadata
Semantic Association Transversal
Text + Domain + Text Metadata + Value-added Metadata from Knowledgebase
Semantic Annotation
Enhanced Text
Knowledge Base
World Wide Web
Internal Resources

SEMAG!X
POWER • THROUGH • RELEVANCE

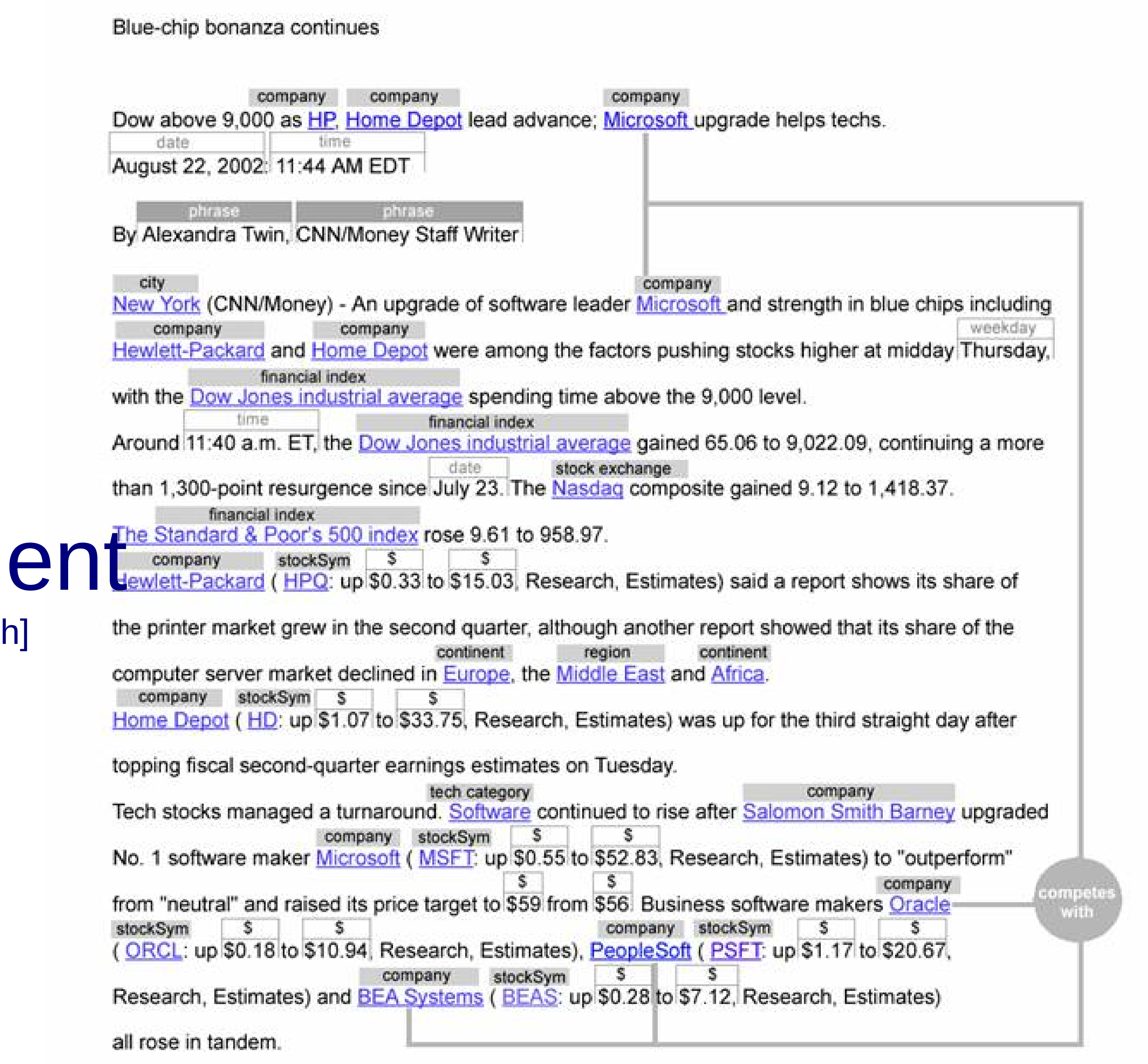
[Hammond, Sheth, Kochut 2002]

LSDIS

Semantic Annotation + Enhancement

[Bancroft, Hammond, Sheth]

SEMAG!X
POWER · THROUGH · RELEVANCE



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.

Application 2: Web of Belief (WOB) by UMBC

- Web Of Belief (WOB) framework that maintains trust and provenance for SWETO
  - □ L. Ding, P. Kolari, A. Joshi, T. Finin, Y. Yesha (UMBC)

Presented at: “Trust on the Web Track” (also at Developers Day)

Ongoing work

- Quality measures of the ontology
- Access to the ontology
  - Web service
  - Filtering, views and versioning
- On-the-fly semantic annotation

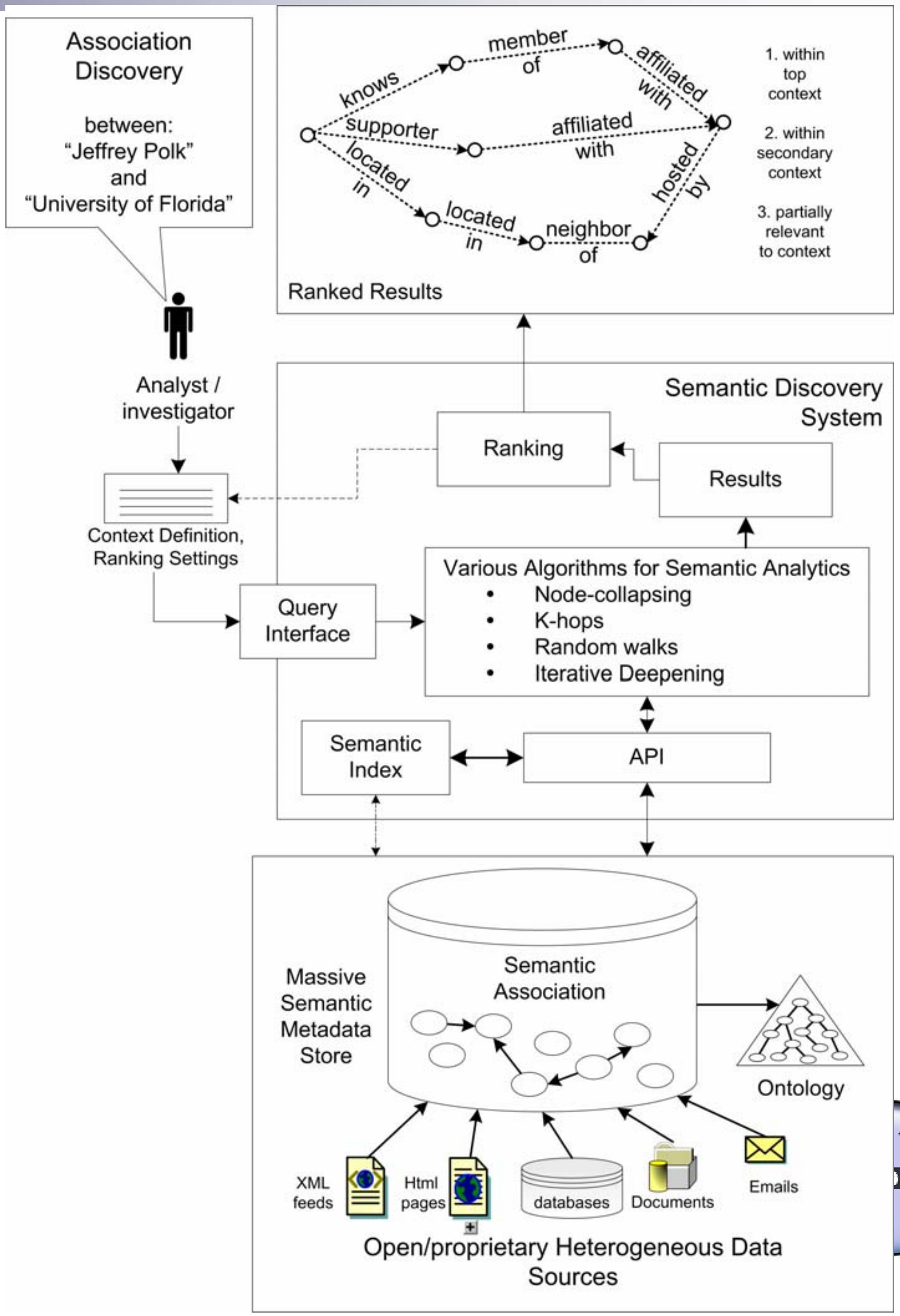
Future plans for benchmarking

- Semantic Search, Browsing and Personalization
- Semantic Portals
  - i.e., *SEMPL* automatically identifies entities
- Semantic Analytics
  - Discovery of *semantic associations* [ρ-operator]
  - Example apps: CIRAS (Semagix), PISTA

SEMAGIX
POWER · THROUGH · RELEVANCE

LSDIS

Approach to developing Semantic Analytics Application Benchmarking



SEMAG!X
POWER · THROUGH · RELEVANCE

Conclusions & Future Work

- Using Semagix Freedom, we have created a broad and deep Semantic Web Evaluation Ontology (SWETO)
  - Public access under Creative Commonsense license;
  - Looking for usage, feedback (of all kinds: schema, population, quality), and partners (for developing bechmarks)

SEMAG!X
POWER · THROUGH · RELEVANCE

LSDIS

Conclusions & Future Work

- More extraction of entities focusing on partners' needs
- Also plan to further investigate the use of semantic similarity for entity disambiguation
- Ontology lifecycle support

SWETO Project Homepage

- http://lsdis.cs.uga.edu/Projects/Semdis/SWETO/
  - □ Google or other search engine: “SWETO”
  - □ Project description, papers, presentations

- Acknowledgements: This work is partially funded by NSF-ITR-IDM Award #0325464 and NSF-ITR-IDM Award # 0219649.

SEMAG!X
POWER · THROUGH · RELEVANCE

LSDIS

References

[1] K. Anyanwu, and A. Sheth. “r-Queries: Enabling Querying for Semantic Associations on the Semantic Web”. Twelfth International World Wide Web Conference, Budapest, Hungary. May 20-24, 2003; pp. 690-699

[2] S. Bechhofer, F. Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneider, et al. (2003). “OWL Web Ontology Language Reference”. W3C Proposed Recommendation, from http://www.w3.org/TR/owl-ref/

[3] R. Guha and R. McCool. “Tap: A Semantic Web Test-Bed”. Journal of Web Semantics, 1(1), Dec. 2003, pp. 81-87

[4] B. Hammond, A. Sheth, K. Kochut. “Semantic Enhancement Engine: A Modular Docu-ment Enhancement Platform for Semantic Applications over Heterogeneous Content in Real World Semantic Web Applications”. V. Kashyap & L. Shklar, Eds., IOS Press, 2002

[5] O. Lassila, & R. Swick. “Resource Description Framework (RDF) Model and Syntax Specification”. W3C Recommendation, from http://www.w3.org/TR/REC-rdf-syntax/

SEMAG!X
POWER · THROUGH · RELEVANCE

LSDIS