Semantics in the Semantic Web—
the implicit, the formal and the powerful

(with a few examples from Glycomics)

Amit Sheth
Large Scale Distributed Information Systems (LSDIS) lab, Univ. of Georgia

October 26, 2004, 1:30pm to 2:30pm
Berkeley Initiative in Soft Computing (BISC) Seminar

Special thanks to Christopher Thomas & Satya Sanket Sahoo
NIH Integrated Technology Resource for Biomedical Glycomics

Complex Carbohydrate Research Center
The University of Georgia

Biology and Chemistry
- Michael Pierce – CCRC (PI)
- Al Merrill - Georgia Tech
- Kelley Moremen - CCRC
- Ron Orlando - CCRC
- Parastoo Azadi – CCRC
- Stephen Dalton - UGA Animal Science

Bioinformatics and Computing
- Will York - CCRC
- Amit Sheth, Krys Kochut, John Miller; UGA Large Scale Distributed Information Systems Laboratory
Central thesis

• Machines do well with “formal semantics”; but current mainstream approach for “formal semantics” based on DLs and FOL is not sufficient

• Incorporate ways to deal with raw data and unorganized information, real world phenomena, and complex knowledge humans have, and the way machines deal with (reason with) knowledge

• Need to support “implicit semantics” and “powerful semantic” which go beyond prevalent “formal semantics” based Semantic Web
Towards the semantic web

- One goal of the semantic web is to facilitate the communication between machines
- Based on this, another goal is to make the web more useful for humans
The Semantic Web

- capturing “real world semantics” is a major step towards making the vision come true.
- These semantics are captured in **ontologies**
- Ontologies are meant to express or capture
  - Agreement
  - Knowledge
- Current choice for ontology representation is primarily Description Logics
What are formal semantics?

• Informally, in formal semantics the meaning of a statement is unambiguously burned into its syntax
• For machines, syntax is everything.
• A statement has an effect, only if it triggers a certain process.
Description Logics

- The current paradigm for formalizing ontologies is in form of bivalent description logics (DLs).
- A subset of First Order Logics (FOL)
Real World?
The world is informal

Even more than humans, machines have a hard time understanding the "real world" of Semantics.

Hiya Dad! I'm having a great time. But unfortunately, it's my last half day; I'm going back to London.

The solution: Formal Semantics.

There's a significant visit from our "French window doors." At the top when you look up, you can see sheets of snow but of your eyes "shift counting."

Anyway, I'm XXX Martin.
The world can be incomprehensible

Sometimes we only see a small part of the picture

We need to be able to see the big picture
The world is complex

• Sometimes our perception plays tricks on us
• Sometimes our beliefs are inconsistent
• Sometimes we can not draw clear boundaries
• → we need more Powerful Semantics
William Woods

“Over time, many people have responded to the need for increased rigor in knowledge representation by turning to first-order logic as a semantic criterion. This is distressing, since it is already clear that first-order logic is insufficient to deal with many semantic problems inherent in understanding natural language as well as the semantic requirements of a reasoning system for an intelligent agent using knowledge to interact with the world.” [KR2004 keynote]
Lotfi Zadeh

- Lotfi Zadeh identifies a lexicon of World Knowledge and a sophisticated deduction system as the core of a question answering system
- World Knowledge cannot be adequately expressed in current bivalent logic formalisms
- Much of Human knowledge is perception based
- “Perceptions are intrinsically imprecise”*

Michael Uschold’s semantic categories

**Pump**: “a device for moving a gas or liquid from one place or container to another”

- **Shared human consensus.**
- **Text descriptions.**
- **Semantics hardwired; used at runtime.**
- **Semantics processed and used at runtime.**

**Implicit**

**Informal** (explicit)

**Formal** (for humans)

**Formal** (for machines)
Metadata and Ontology: Primary Semantic Web Enablers

- **Ontology**
  - Example: Anatomy, Diagnostics, ...

- **Semantic Metadata**
  - Example ontology-driven metadata:
    - Region: Upper Abdomen
    - Organ: Liver
    - Pathological Structure: Abscess, Abscess located in Liver

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

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

- **Data**
  - (Structured, semi-structured and unstructured)
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 price that enables
  - resolution of semantic heterogeneity
  - semantic integration
  - semantically correlating/associating objects and documents
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
Expressiveness Range:
Knowledge Representation and Ontologies

Ontology Dimensions After McGuinness and Finin

- Simple Taxonomies
- Expressive Ontologies

Thesauri
“narrower term” relation

Terms/glossary

Formal is-a
Frames (properties)
Value Restriction

Disjointness, Inverse, part of…

General Logical constraints

Catalog/ID DB Schema Wordnet UMLs RDF RDFS DAML OWL IEEE SUO CYC

BioPAX

KEGG TAMBIS

BioPAX

GO SWETO GlycO Pharma EcoCyc

Large Scale Distributed Information Systems
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
What are implicit semantics?

- Every collection of data or repositories contains hidden information
- We need to look at the data from the right angle
- We need to ask the right questions
- We need the tools that can ask these questions and extract the information we need
How can we get to implicit semantics?

- Co-occurrence of documents or terms in the same cluster
- A document linked to another document via a hyperlink
- Automatic classification of a document to broadly indicate what a document is about with respect to a chosen taxonomy.
- Use the implied semantics of a cluster to disambiguate (does the word “palm” in a document refer to a palm tree, the palm of your hand or a palm top computer?)
- Bioinformatics applications that exploit patterns like sequence alignment, secondary and tertiary protein structure analysis, etc.
- Techniques and Technologies: Text Classification/categorization, Clustering, NLP, Pattern recognition, ...
Implicit semantics

• Most knowledge is available in the form of
  – Natural language $\rightarrow$ NLP
  – Unstructured text $\rightarrow$ statistical

• Needs to be extracted as machine processable semantics/ (formal) representation
Taxaminer

- Since ontology creation is an expensive and time-consuming task, the Taxaminer project at the LSDIS lab aims at semi-automatically creating a Labeled hierarchical topic structure as a basis for automatic classification and semi-automated ontology creation
  - Knowledge about a domain can be found in documents about the domain.
  - How can a machine identify sub-topics and give them an adequate label?
Taxaminer

Document collection

Build vector space model

Build hierarchical cluster

Select the “best” clusters and assign the most pertaining words as labels of the nodes in the resulting hierarchy.
A document collection is “flat”. The topic hierarchy is implicit.

Select the “best” clusters and assign the most pertaining words as labels of the nodes in the resulting hierarchy.

**Document collection**

**Build vector space model**

**Build hierarchical cluster**

**Taxaminer**
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.

KB, statistical and linguistic techniques
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
GlycO statistics: Ontology schema can be large and complex

- 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.
Large Scale Distributed Information Systems

Query and visualization
A biosynthetic pathway

UDP-N-acetyl-D-glucosamine + G00020 ⇄ UDP + G00021
The impact of GlycO

- GlycO models classes of glycans with unprecedented accuracy.
- Implicit knowledge about glycans can be deductively derived.
- Experimental results can be validated according to the model.
Identification and Quantification of N-glycosylation

Cell Culture

Glycoprotein Fraction

Glycopeptides Fraction

Separation technique I

Glycopeptides Fraction

PNGase

Peptide Fraction

Separation technique II

Peptide Fraction

Mass spectrometry

ms data

ms peaklist

Data reduction

Data correlation

N-dimensional array

Peptide identification and quantification

ms/ms data

ms/ms peaklist

Data reduction

Peptide list

Peptide identification
Four structural components†:

- Sample Creation
- Separation (includes chromatography)
- Mass spectrometry
- Data analysis

†: pedrodownload.man.ac.uk.Domains.shtml
ms/ms peaklist data

830.9570 194.9604 2
580.2985 0.3592
688.3214 0.2526
779.4759 38.4939
784.3607 21.7736
1543.7476 1.3822
1544.7595 2.9977
1562.8113 37.4790
1660.7776 476.5043

Annotated ms/ms peaklist data
Annotated ms/ms peaklist data
Beyond Provenance…. Semantic Annotations

- Data provenance: information regarding the ‘place of origin’ of a data element

- Mapping a data element to concepts that collaboratively define it and enable its interpretation – Semantic Annotation

- Data provenance paves the path to repeatability of data generation, but it does not enable:
  - Its interpretability
  - Its computability

Semantic Annotations make these possible.
The inference: instances of the class collection of Biosynthetic enzymes (GNT-V) are involved in the specific cellular process (metastasis).
Ontologies – many questions remain

- How do we design ontologies with the constituent concepts/classes and relationships?
- How do we capture knowledge to populate ontologies?
- Certain knowledge at time $t$ is captured; but real world changes
- Imprecision, uncertainties and inconsistencies
  - what about things of which we know that we don’t know?
  - What about things that are “in the eye of the beholder”?
- Need more powerful semantics
Large Scale Distributed Information Systems

**Dimensions of expressiveness**

- **Complexity**
  - Bivalent
  - Multivalued discrete
  - Continuous

- **Degree of Agreement**
  - Informal
  - Semi-Formal
  - Formal

- **Expressiveness**
  - Current Semantic Web Focus
  - Future research

**Current Semantic Web Focus**
- FOL with functions
  - RDFS/OWL
- FOL w/o functions
  - XML
  - RDF

Cf: Guarino, Gruber
The downside

- That a structure is not valid according to the ontology could just mean that it is a new kind of structure that needs to be incorporated.
- That a substance can be synthesized according to one pathway does not exclude the synthesis through another pathway.
Man$_9$GlcNAc$_2$ is a Glycosyl Transferase that synthesizes Glycan. May Synthesize Lipid $\beta$-mannosyl transferase which transfers Mannose and contains Mannose in Glycan.
What we want

• Validate pathways with experimental evidence. Many pathways still need to be verified.

• Reason on experimental data using statistical techniques such as Bayesian reasoning

• Are activities of iso-forms of biosynthetic enzymes dependent on physiological context? (e.g. is it a cancer cell?)
What we need

• We need a formalism that can
  – express the degree of confidence that e.g. a glycan is synthesized according to a certain pathway.
  – express the probability of a glycan attaching to a certain site on a protein
  – derive a probability for e.g. a certain gene sequence to be the origin of a certain protein
Protein Classes

Protein
Enzyme
Hydrolase
Transferase
...
Regulatory Protein
DNA-Binding Protein
Receptor
...

Large Scale Distributed Information Systems
Protein Classes

- The classification of proteins according to their function shows that there are proteins that play more than one role
- There are no clear boundaries between the classes
- Protein function classes have fuzzy boundaries
The consequence

• Both in future main stream applications such as question answering systems and in scientific domains such as BioInformatics, reasoning beyond the capabilities of bivalent logic is indispensable.

• We need more powerful semantics
Powerful Semantics

• Fuzzy logics and probability theory are “complementary rather than competitive”*.
• Fuzzy logic allows us to blur artificially imposed boundaries between different classes.
• The other powerful tool in soft computing is probabilistic reasoning.

*Lotfi A. Zadeh. Toward a perception-based theory of probabilistic reasoning with imprecise probabilities.
Powerful Semantics

• In order to use a knowledge representation formalism as a basis for tools that help in the derivation of new knowledge, we need to give this formalism the ability to be used in abductive or inductive reasoning.

*Lotfi A. Zadeh. Toward a perception-based theory of probabilistic reasoning with imprecise probabilities.
Powerful Semantics

• The formalism needs to express probabilities and fuzzy memberships in a meaningful way, i.e. a reasoner must be able to meaningfully interpret the probabilistic relationships and the fuzzy membership functions

• The knowledge expressed must be interchangeable, hence a suitable notation, following the layer architecture of the Semantic Web, must be used.
How to power the semantics

- A major drawback of logics dealing with uncertainties is the assignment of prior probabilities and/or fuzzy membership functions.
- Values can be assigned manually by domain experts or automatically.
- Techniques to capture implicit semantics:
  - Statistical methods
  - Machine Learning
What are powerful semantics?

- Powerful semantics can be formal
- Powerful semantics can capture implicit knowledge
- Powerful semantics can cope with inconsistencies
- Powerful semantics can formalize our perceptions
- Powerful semantics can deal with imprecision
The long road to more power

Implicit Semantics
+
Formal Semantics
+
Soft Computing Technologies
=
Powerful Semantics
For more information

- http://lsdis.cs.uga.edu
  - Especially see Glycomics project
- http://www.semagix.com