13. The Semantic Web

INFO 202 - 13 October 2008

Bob Glushko

Plan for INFO Lecture #13

Overview of the Semantic Web

RDF

OWL

A Critical Evaluation of the Semantic Web

Semantically-aware systems

The Metadata Questions - Reminder (from 17 September)

What objects have metadata assigned to them?

What is the granularity of the object?

What metadata is being assigned?

What is the format of the metadata?

What standards for the metamodel or controlled vocabulary are being followed?

Who is assigning the metadata?

The Vision of the "Semantic Web"

In a classic 2001 paper, Sir Tim Berners-Lee says:

The Web can reach its full potential only if it becomes a place where data can be shared and processed by automated tools as well as by people...

The Semantic Web will bring structure to the meaningful content of Web pages...

For the Web to scale, tomorrow's programs must be able to share and process data even when these programs have been designed totally independently.

Like the Internet, the Semantic Web will be as decentralized as possible... because central control is stifling

The Semantic Web Scenario (2001 paper)

"...Lucy instructed her Semantic Web agent through her handheld Web browser. The agent promptly retrieved information about Mom's prescribed treatment from the doctor's agent, looked up several lists of providers, and checked for the ones in-plan for Mom's insurance within a 20-mile radius of her home and with a rating of excellent or very good on trusted rating services.

... trying to find a match between available appointment times (supplied by the agents of individual providers through their Web sites) and Pete's and Lucy's busy schedules.

... the agent presented them with a plan. Pete didn't like it

He set his own agent to redo the search with stricter preferences about location and time. ... the new plan was presented: a much closer clinic and earlier times..."

How Semantic Discovery (Supposedly) Works

Many automated Web-based services exist without semantics, but other programs have no way to locate one that will perform a specific function

Service discovery can happen only when there is a common language to describe a service in a way that lets other agents "understand" the function offered and how to take advantage of it

Services and agents can advertise their function by depositing such descriptions in directories analogous to the Yellow Pages

Service discovery enables agents to delegate tasks to create the overall "value chain" in which subassemblies of information are passed from one agent to another

Why Today's Web Isn't Semantic

TBL invented the Web and HTML as a non-proprietary publishing format for the Internet

TBL made conscious decisions NOT to use full SGML (precursor to XML) and state of the art concepts of linking and hypermedia to make it easy to implement Web servers and create Web pages

Designing HTML to be conceptually simple and easy to implement rather than general and powerful led to its rapid adoption after invention of graphical browser in 1994

The Simplicity/Expressiveness Tradeoff - The Upside

"Marking up" a document by surrounding bits of text with "pointy brackets" and tags whose name suggests a structural role or formatting is both conceptually and technically simple

<H1> "I'm a heading"
<i> "Format me in italics"

Most people use HTML tags to create some desired presentation for the content, and "view source" in browsers makes it easy to "cut and paste" existing pages, substituting content in whatever tag "skeleton" that produces the desired look

The Simplicity/Expressiveness Tradeoff - The Downside

The "markup means structure/format" approach works until people want to use the Web as a business platform and publish information that business applications can interpret as prices, quantities, part numbers, addresses, etc.

This changes the problem to be solved by the markup language from presentational formatting to semantic modeling, and HTML isn't designed for that

XML was invented in 1997 to enable richer, content-oriented markup, but millions of non-semantic HTML pages already existed and millions continue to be created

Some of this HTML is "smarter" than others - using , <class>, and embedded "microformats" to create structures to which semantics can more easily be applied ... but not automatically

Getting to the Semantic Web

"The Semantic Web is not a separate Web but an extension of the current one..."

"Adding logic to the Web - the means to use rules to make inferences, choose courses of action and answer questions - is the task before the Semantic Web community"

"A little semantics goes a long way"

How does Today's Web become the Semantic Web?

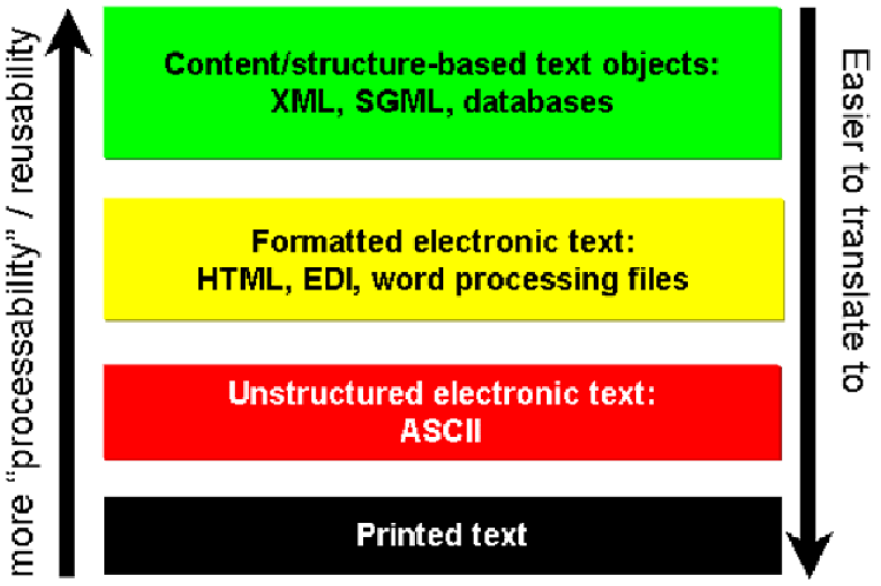
Convert existing Web page content to semantic markup

Annotate existing Web page content with semantic metadata

Create new web page content with semantic content and semantic metadata

Create new pages or resources that are designed from the outset to be "meta-pages" that facilitate semantic processing of all the other metadata

Semantic Markup Can Often Be "Re-applied"



Semantic Markup Can Sometimes be "Extracted"

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Microsoft Corporation
CEO
Bill Gates

Microsoft
Gates

Microsoft

Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation

Extracted Descriptions

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

IE

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

Semantic Annotation

"Annotation" generally means "semantics applied to a document or information resource by a person" -- rather than by NLP

Someone other than the author can sometimes figure out the author's intent and context if:

- they both belong to the same narrow organization, "community of interest" or "social network"
- there are "extraction," "summarization," and other text processing tools that can help them

But being able to apply useful semantic annotation to an arbitrary information resource is more than we expect professional "cataloguers" to do!

Semantic Authoring

But even when an author is creating his own semantically-encoded content or annotation ...

How are semantic descriptors chosen?

What do those descriptors mean?

Can others trust what the author does?

Semantic Authoring: The 2001 Vision

"The clinic's web page will have more than just keywords; it will have computer-processable information about when specific doctors take appointments"

"These semantics were encoded into the Web page when the clinic's office manager (who never took Comp Sci 101) massaged it into shape using off-the-shelf software for writing Semantic Web pages along with resources listed on the Physical Therapy Association's site"

Semantic Authoring: The 2007 Vision

Semantic Authoring: Today's Reality

"Many current semantic web tools still require expertise in semantic technologies and web standards... which can repel web developers"

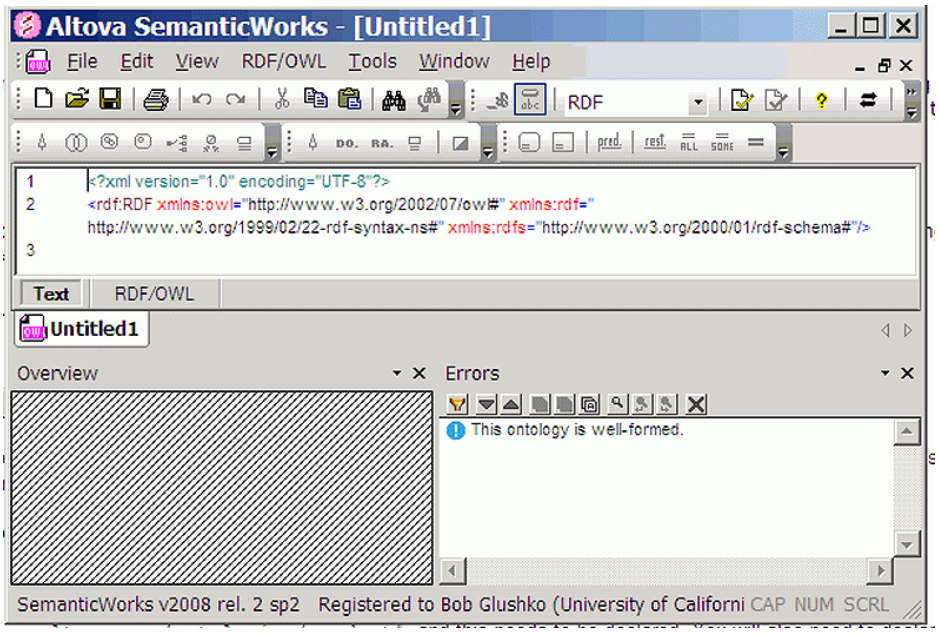
"Or, they burden people cognitively with their own internal semantic models and ontologies"

- Dapper: Semantify Your Site

http://blogs.ischool.berkeley.edu/i202f08/2008/10/01/123/ (Thanks to Nat Wharton)

http://blogs.ischool.berkeley.edu/i202f08/2008/10/01/semantic-apps-and-rdf/ (Thanks to Ben Cohen)

Altova "Semantic Works" Tool



How the Semantic Web Creates Meaning

"URIs ensure that concepts are not just words in a document but are tied to a unique definition that everyone can find on the Web"

- "But two databases may use different identifiers for what is in fact the same concept...
- "The program must have a way to discover such common meanings"

Ontology pages on the Web solve terminology (and other) problems by providing equivalence relations

Often two groups independently develop very similar concepts, and describing the relation between them brings great benefits

Technologies for the Semantic Web

XML

RDF

OWL and other ontology languages

Why XML Alone Isn't Sufficient

You can use XML to create a content-oriented vocabulary rather than the presentation-oriented one in HTML

XML schemas allow you to specify structural, occurrence, and datatyping constraints for instances that must conform to them

You can use XML namespaces to reuse XML constructs across a set of related document types

But the semantics associated with XML constructs are NOT explicitly represented in the instance or the schema

(Element and attribute names, container structures, etc. can suggest semantics to people, but not in a way that is "computable")

RDF: Resource Description Format

RDF is a graph-based model for describing Internet resources and how they relate to each other

RDF can be used to encode metadata, the usual sort of information about an information resource, like its title, author, creation date, etc.

But it can be used to represent information about anything that can be identified on the Web, not just published or retrieved on it

This broader idea about "Web resource" makes it a general mechanism for organizing and integrating information

RDF Data Model -- The Conceptual View

A general way to represent information about something is in three parts:

- The thing (or resource) being described
- The specific property of the thing
- The value of the property

The data model is usually stated as Statement -> (Subject, Predicate, Object) but there are many other ways to say it; don't get confused by the synonyms

Statement (Student, attends, University)

RDF Vocabulary Synonyms

COMMON TERM	SYNONYMS
Resource	Subject, object
Resource identifier	Name, (resource) URI, URL, label, ID, identifier
Properties	Attributes
Statement	Triple, tuple, binding, assertion
Subject	Source, resource, node, root
Predicate	Arc, (statement) URI, property, atom
Object	Value, resource, node, literal

Linking Statements to Create a Graph

When the subject of one statement is the object of another, the statements are conceptually linked

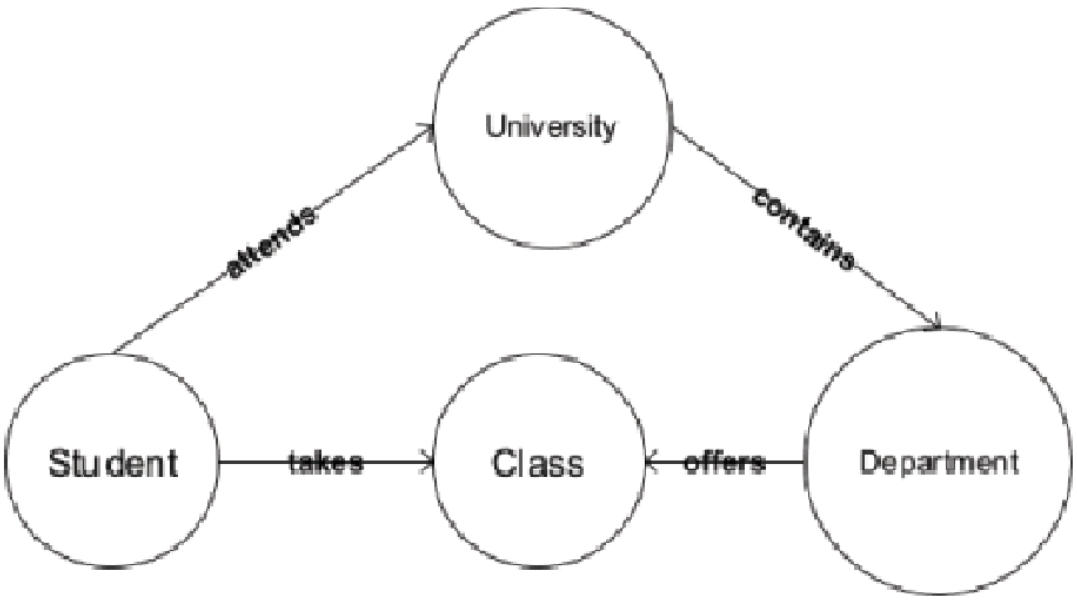
Statement (Student, attends, University)

Statement (University, contains, Department)

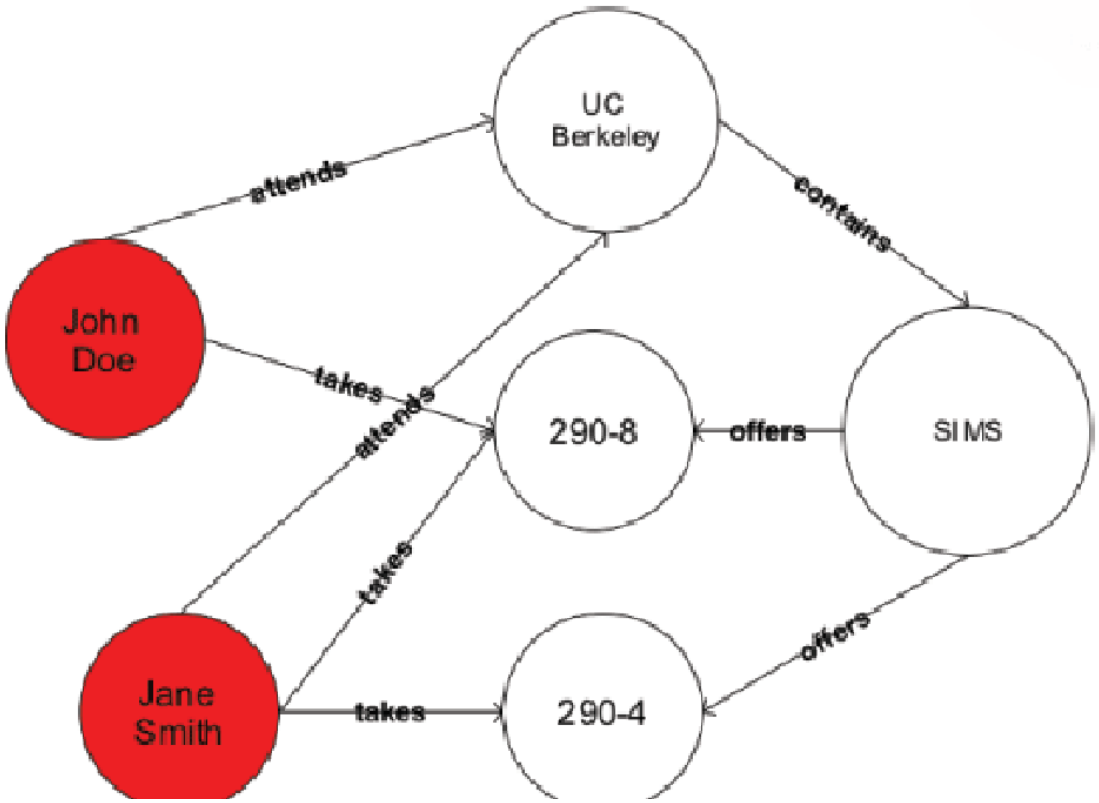
Statement (Department, offers, Class)

Collections or sets of linked statements form a directed, labeled graph

Conceptual Model as Graph



A Graph Instance of the Model



Instance as Set of Statements

John Doe attends UC Berkeley

John Doe takes 290-8

Jane Smith attends UC Berkeley

Jane Smith takes 290-8

Jane Smith takes 290-4

Some Inferences about the Instances?

John Doe takes classes at SIMS

290-4 is offered by UC Berkeley

Jane Smith is a classmate of John Doe

RDF Syntax -- Simplified Implementation View

<Description> describes a resource

Attributes or elements contained in <Description> are properties of the resource

Their content is the value of the property

```
<Description about="some.uri/person/JohnDoe">
   <attends resource="some.uri/institution/UCBerkeley">
   </attends>
</Description>
```

RDF Syntax -- More Realistic Implementation View

If anyone can define or own the subjects, predicates, and objects there must be a way to tell them apart

So we need to declare some namespaces

We can create standalone RDF documents into which we incorporate statements using our own vocabulary

Or we can embed RDF statements into the XML content of a web page, which would be encoded in some other vocabulary with a different namespace

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/TR/rdf-syntax-grammar"
    xmlns:info202="BobBerkeley/info202/examples">

    <rdf:Description about="some.uri/person/JohnDoe">
      <info202:attends resource="some.uri/institution/UCBerkeley">
      </info202:attends>
   </rdf:Description>
</rdf:RDF>
```

But We Still Need Ontologies

Suppose we have RDF statements:

- (Bob Glushko, teaches, INFO202)
- (Information System & Service Design, is-taught-by, Dr. Robert J. Glushko)

No RDF processor can link these into a graph and make the inference that both classes are taught by the same person

Ontology in the Semantic Web

If an ontology defined:

- "Bob Glushko" and "Dr. Robert J. Glushko" to be the same person
- "teaches" and "is-taught-by" to be inverse relationships

Then a processor could make some inferences

OWL

OWL is an XML vocabulary for describing properties and classes in rigorous ways that include relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes.

```
<owl:Class rdf:ID="Wine">
  <rdfs:subClassOf rdf:resource="&food;PotableLiquid"/>
  <rdfs:label xml:lang="en">wine</rdfs:label>
  <rdfs:label xml:lang="fr">vin</rdfs:label>
  ...
</owl:Class>

<owl:ObjectProperty rdf:ID="madeFromGrape">
  <rdfs:domain rdf:resource="#Wine"/>
  <rdfs:range rdf:resource="#WineGrape"/>
</owl:ObjectProperty>
```

Questions at the Heart of the Semantic Web Vision

If resolving semantic ambiguity requires an authority, but anyone can make assertions, how do we know who the authorities are?

Is "defining a new URI somewhere on the Web" the same as "defining a new concept"

You've all developed vocabularies and metamodels... are these things that the ordinary web user can do in a useful and consistent manner?

Have improvements in Web search made a lot of this moot? Has the Web already been "tamed" by Google?

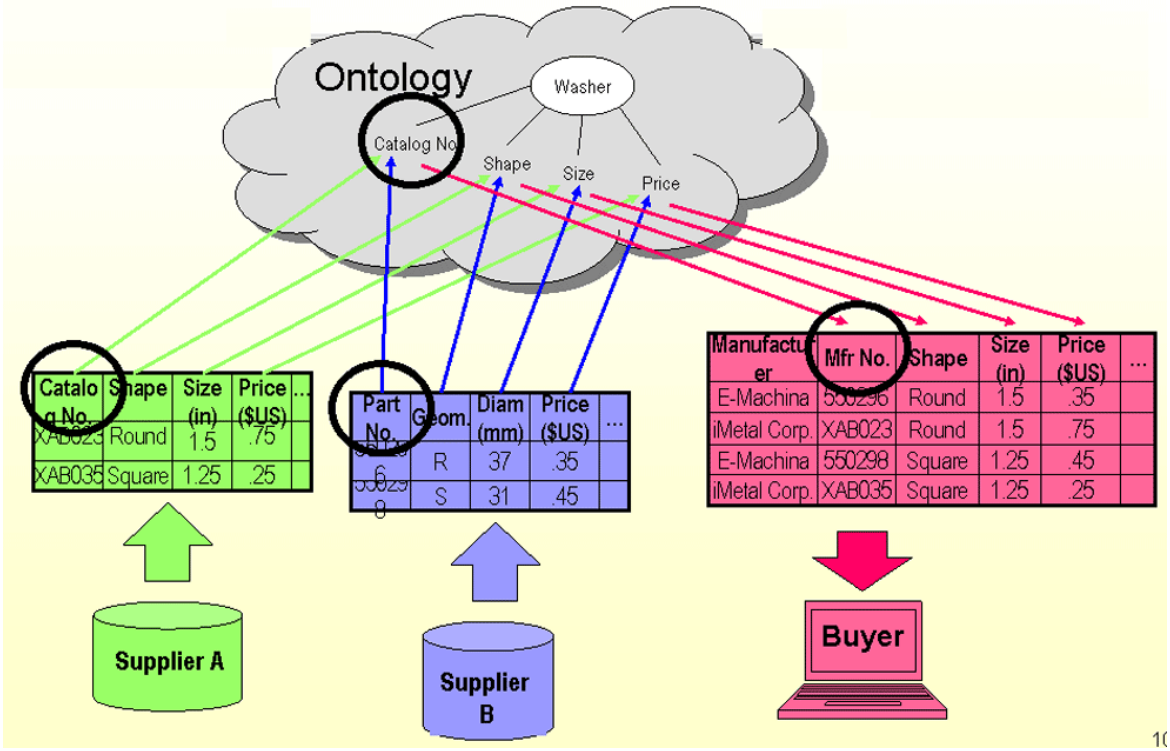
Semantically-Aware Systems

Regardless of the feasibility or likelihood of TBL's Semantic Web vision, it is undeniable that information-intensive systems must become more semantically-aware

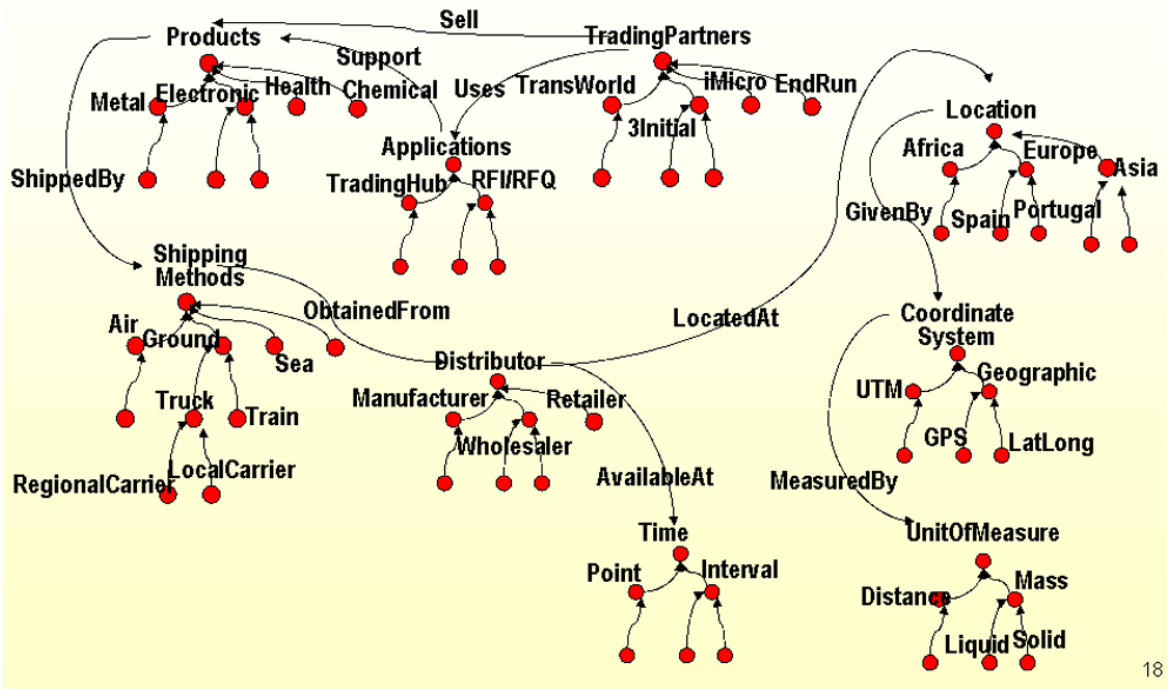
Three kinds of technologies are being developed:

- Ontology editors and mapping tools (Design Time)
- Inference engines that operate on ontologies (Run Time)
- Semantic brokers that operate on messages (Run Time)

Ontology for Semantic Harmonization of Product Catalogs

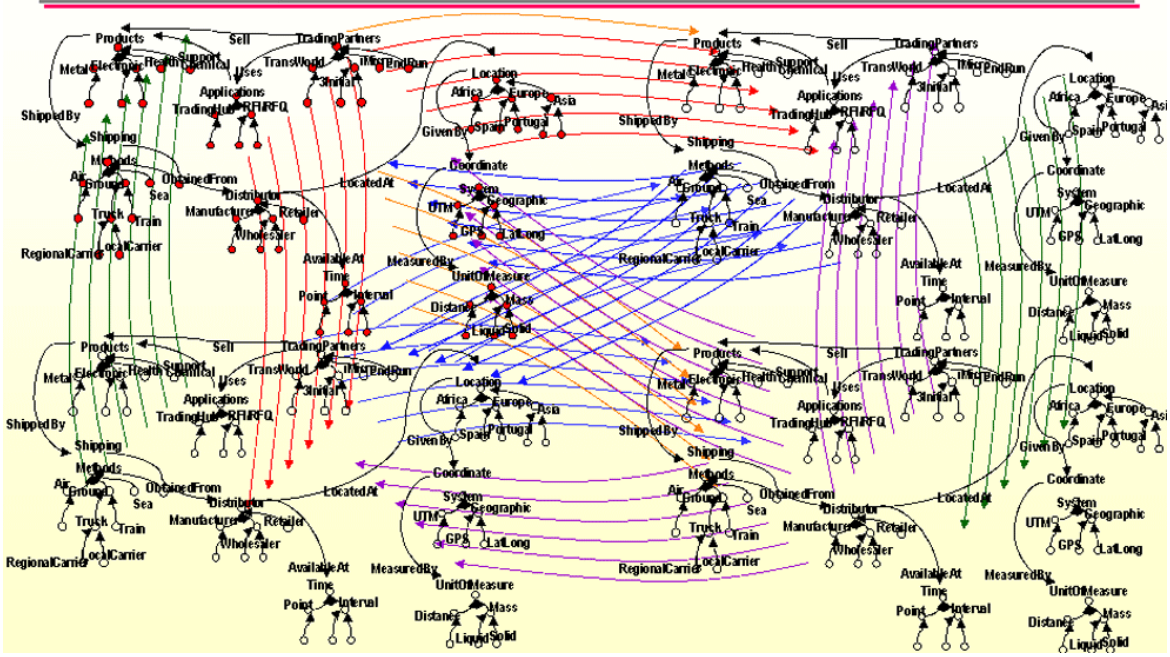


But It's Not that Simple for Each Firm



And As The Number of Firms Grows...

Now Assume Each Company Has Separate Enterprise Semantics, Multiply by the Number of Companies, & Have Them Interoperate and Preserve Semantics



Readings for INFO Lecture #14

Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. "HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read," Proceedings of the seventeenth conference on Hypertext and hypermedia, 2006

Michael G. Noll and Christoph Meinel, "Exploring Social Annotations for Web Document Classification," SAC '08

Tom Gruber, "Collective knowledge systems: Where the social web meets the semantic web," Journal of Web Semantics, 6, 2008