12. The Semantic Web

INFO 202 - 8 October 2007

Bob Glushko
Plan for INFO Lecture #12

Overview of the Semantic Web
Metadata Mechanisms for the Semantic Web
RDF
OWL
A Critical Evaluation of the Semantic Web
Semantically-aware systems
The Metadata Questions - A Re-Reminder (from 12 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"

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
"...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..."
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 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.

Short-term solution of writing business applications that "scrape" HTML pages by assuming fixed formatting is too brittle.

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

How does today's Web become the Semantic Web?
Some Alternatives for Making the Web Semantic

- 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 Authoring in the Scientific American Scenario

"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"

What might the office manager have learned in Comp Sci 101 (or IS 202) that is now somehow unnecessary?

What specifically is the "off-the-shelf" software going to do?
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
How 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.
Three Technologies for the Semantic Web

- XML
- RDF
- OWL and other ontology languages
XML

The Scientific American article says that:

- "XML lets everyone create their own tags"
- "XML allows users to add arbitrary structure to their documents but says nothing about what the structures mean"
The Scientific American article says that:

- "Meaning is expressed by RDF"
- Naming every concept simply by a URI lets anyone express new concepts that they invent with minimal effort
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.
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

<table>
<thead>
<tr>
<th>COMMON TERM</th>
<th>SYNONYMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource</td>
<td>Subject, object</td>
</tr>
<tr>
<td>Resource identifier</td>
<td>Name, (resource) URI, URL, label, ID, identifier</td>
</tr>
<tr>
<td>Properties</td>
<td>Attributes</td>
</tr>
<tr>
<td>Statement</td>
<td>Triple, tuple, binding, assertion</td>
</tr>
<tr>
<td>Subject</td>
<td>Source, resource, node, root</td>
</tr>
<tr>
<td>Predicate</td>
<td>Arc, (statement) URI, property, atom</td>
</tr>
<tr>
<td>Object</td>
<td>Value, resource, node, literal</td>
</tr>
</tbody>
</table>
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

- Student attends University
- Class offers Department
- Student takes Class

Diagram: [Diagram Image]
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

```xml
<rdf:RDF
   xmlns:rdf="http://www.w3.org/TR/rdf-syntax-grammar"
   xmlns:is202="BobBerkeley/is202/examples">

   <rdf:Description about="some.uri/person/JohnDoe">
       <is202:attends resource="some.uri/institution/UCBerkeley"/>
   </rdf:Description>
</rdf:RDF>
```
But We Still Need Ontologies

Suppose we have RDF statements:

- (Bob Glushko, teaches, IS202)
- (Information & Services Economy, 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
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.

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

McComb (CIO's Guide to Semantics) lists three kinds of technologies:

- Ontology editors and mapping tools (Design Time)
- Inference engines that operate on ontologies (Run Time)
- Semantic brokers that operate on messages (Run Time)
Opportunities / Business Problems that Need Semantic Awareness

- Transaction processing
- Knowledge management
- Decision support
- Unstructured activity
- Environment
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 #13

Adam Mathes, "Folksonomies - Cooperative Classification and Communication Through Shared Metadata


Morgan Ames & Mor Naaman, "Why We Tag: Motivations for Annotation in Mobile and Online Media," CHI 2007

Shilad Sen, Shyong Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, & John Riedl, "tagging, communities, vocabulary, evolution," CSCW'06