Introduction to the Semantic Web

Prof. dr. Jon Atle Gulla

“People are starting to realize that their information outlives their software”
Tim Berners-Lee

“Information meaning is too tightly coupled to its initial use or application. Thus it is very difficult for either (a) machines to reuse information or (b) for people to query on concepts (instead of just on terms).”
Jeffrey. T. Pollock

Outline of Presentation

- Motivational Cases
  - Case 1: Integrated Operations
  - Case 2: Web search
- Semantic Web Vision
  - Vision and Architecture
  - HTML limitations
  - XML limitations
- Semantic Web Machinery
  - Approach
  - Language stack
  - Ontologies
- Future of the Semantic Web
  - Semantic Web Capabilities
  - Future applications
MOTIVATIONAL CASES

1. Integrated Operations Vision

- Introduce a semantic standard that
  - covers all disciplines in subsea petroleum activities
  - is used by all companies
- Use semantic standard to define meaning of data/services & provide uniform access to all data/services

Today: Many databases, need 1-to-1 mappings to improve collaboration and integration

Future: All data in all databases defined with respect to common standard. All services defined with respect to common standard. Standardized semantic interface to all data/services
Current Operational Issues

- **Interoperability:** How to link operations across disciplines and across organizational and geographical boundaries?
- **Semantics:** How to ensure that the composed operation does what it is supposed to do?
- **Adaptability/flexibility:** How to optimize operations in a particular context?
- **Autonomy:** How and when to initiate/execute operations?

Operations performance challenges:

Silly Small Semantic Conflicts

- **Aggregation**
- **Value representation**

Fra SiCoP white paper: "Introducing Semantic Technologies and the Vision of the Semantic Web"
Semantic Conflicts between Applications

<table>
<thead>
<tr>
<th>Semantic conflict</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data type</td>
<td>Different primitives or abstract types for same information</td>
<td>SSN as a VARCHAR vs. a NUM</td>
</tr>
<tr>
<td>Labeling</td>
<td>Synonyms/antonyms have different text labels</td>
<td>When either a NAME and a COMPANY column have data that mean the same thing</td>
</tr>
<tr>
<td>Aggregation</td>
<td>Confused conceptions about the relationships among concepts in similar data sets</td>
<td>Which is a &quot;motorcycle&quot; - 1, 2, 3, 4 or more wheels, how are the constraints modeled in your schema?</td>
</tr>
<tr>
<td>Interrelation</td>
<td>Different abstractions are used to model same domain</td>
<td>Are &quot;cars&quot; and &quot;motorcycles&quot; kinds of &quot;vehicles&quot; or are they top-level concepts themselves?</td>
</tr>
<tr>
<td>Other representation</td>
<td>Consequent choices are made about what concepts are made explicit</td>
<td>&quot;Special&quot; and &quot;MainAssembly&quot; equal &quot;Truck&quot;</td>
</tr>
<tr>
<td>Inference methods</td>
<td>Fundamentally different inference mechanisms are used</td>
<td>Related to Object mappings they emphasize multiplicities, etc.</td>
</tr>
<tr>
<td>Meaning</td>
<td>Information/abstractions exist in unique/multiple concept instance values</td>
<td>&quot;Company&quot;, why so many synonyms? &quot;Corporate&quot;, &quot;Enterprise&quot;, etc., but they refer to the same thing</td>
</tr>
<tr>
<td>Naming and unit</td>
<td>Related data of instances with incompatible units</td>
<td>km vs. English mile</td>
</tr>
<tr>
<td>Concept</td>
<td>Similar concepts with different definitions</td>
<td>&quot;MarginPercent&quot; - object for a NASD application vs. a NYSE system</td>
</tr>
<tr>
<td>Domain</td>
<td>Fundamental incompatibilities in underlying domains</td>
<td>Asset management - object to a fund model system vs. a whole supplier system</td>
</tr>
<tr>
<td>Integrity</td>
<td>Semantically &quot;break&quot; has a primary key that uniquely IDs a passenger?</td>
<td></td>
</tr>
</tbody>
</table>

Interoperability

- Interoperability needed for applications and people to work together

Semantic interoperability:
- Systems share and understand information with respect to mutually accepted domain concepts (e.g. OWL)

Structural interoperability:
- Systems share semantic schemas for common structuring of information (e.g. RDF)

Syntactic interoperability:
- Systems can exchange information by marking up data in a similar fashion (e.g. XML)
2. Web Search

- 75-85% of web users use search engines to traverse the web
- 0.4% of ad banners presented on result page are viewed
- 1/3 of web sessions involve search engines

![Table of search engine statistics](image)

<table>
<thead>
<tr>
<th>Service</th>
<th>Searches Per Day</th>
<th>Searches Per Sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>268 million</td>
<td>1,053</td>
</tr>
<tr>
<td>Overture</td>
<td>167 million</td>
<td>637</td>
</tr>
<tr>
<td>Inktomi</td>
<td>80 million</td>
<td>1,259</td>
</tr>
<tr>
<td>LookSmart</td>
<td>45 million</td>
<td>1,442</td>
</tr>
<tr>
<td>FindWhat</td>
<td>33 million</td>
<td>1,012</td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>20 million</td>
<td>891</td>
</tr>
<tr>
<td>AltaVista</td>
<td>18 million</td>
<td>865</td>
</tr>
</tbody>
</table>

[www.searchenginewatch.com](http://www.searchenginewatch.com)
Retrieved 15 Feb 2005

How to Retrieve Information Today

- Current search technologies on Web:
  - Syntactic search (word frequencies, similarity measures)
  - Natural language preprocessing (lang. detection, lemmatization,..)

- Retrieving information on the Web:
  - Search engines, which index a portion of Web pages as a full-text database;
  - Web directories, which classify selected Web docs by subject;
  - Searching by exploiting hyperlink* structure.

* - hyperlink – a pointer from one page to another.
Where we are Today: the Syntactic Web

- Volume of data:
  - Document explosion
  - Document dynamics (40% changes every month)
  - Distributed over many computers & platforms

- Multitude of languages:
  - Multi-lingual web
  - 40-50 languages used on the web
  - Many text encoding standards

Why is Web Search so Difficult?

<table>
<thead>
<tr>
<th>Search Engine</th>
<th>Reported Size</th>
<th>Page Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>8.1 billion</td>
<td>101k</td>
</tr>
<tr>
<td>MSN</td>
<td>5.0 billion</td>
<td>150k</td>
</tr>
<tr>
<td>Yahoo</td>
<td>4.2 billion (estimate)</td>
<td>500k</td>
</tr>
<tr>
<td>Ask Jeeves</td>
<td>2.5 billion</td>
<td>101k+</td>
</tr>
</tbody>
</table>
Why is Web Search so Difficult?

- **Document Quality:**
  - Misspellings
  - Spam and offensive content
  - Little text
  - All topics
  - Duplicates (30% of web pages)

- **User Behaviour:**
  - Misspellings
  - Query length: 2.4 terms
  - Query session: 8 queries
  - Half of the documents viewed are among top three documents on result page

<table>
<thead>
<tr>
<th>Query</th>
<th>No. of documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>événements</td>
<td>76,000</td>
</tr>
<tr>
<td>événements</td>
<td>420,000</td>
</tr>
<tr>
<td>evenements</td>
<td>35,000</td>
</tr>
<tr>
<td>événements</td>
<td>95,000</td>
</tr>
<tr>
<td>événements</td>
<td>22,000</td>
</tr>
<tr>
<td>événements</td>
<td>9,000</td>
</tr>
</tbody>
</table>

The Sad Truth

- System does not understand what users look for
- System does not understand what the documents are about
- Users depend on the system
Semantic Web vision

- Web was “invented” by Tim Berners-Lee (amongst others), a physicist working at CERN
- His vision of the Web was much more ambitious than the reality of the existing (syntactic) Web:

  “… a plan for achieving a set of connected applications for data on the Web in such a way as to form a consistent logical web of data …”

  “… an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation …”
**Doctor’s appointment**


**Semantic Web Principles**

“Give all information and services well-defined meanings that enable computers and people to work in cooperation”
Background: The Current Web is not Enough

- Homogeneous resources
- Semantically empty
- Needs human interpretation

<table>
<thead>
<tr>
<th>Identified resources</th>
<th>Meaningful links</th>
<th>Machine processable</th>
</tr>
</thead>
</table>

HTML for Interoperability

```html
<H1>
The Rhyme of the Ancient Mariner
</H1>
<i>The Rhyme of the Ancient Mariner</i>, by Samuel Coleridge, is available for the low price of $9.99. This Dover reprint is beautifully illustrated by Gustave Dore.

Julian Schnabel recently directed a movie, <i>Pandemonium</i>, about the relationship between Coleridge and Wordsworth.

Can you devise an algorithm that will retrieve the price and author of the book? **AND** that’s likely to work correctly for **ALL** book descriptions?
HTML for Search

When to retrieve these documents? What is their content?

XML for Interoperability

This is what a web-page in natural language looks like for a machine
XML Adds Structures

XML allows “meaningful tags” to be added to parts of the text

XML is not Machine-Accessible Meaning

But to your machine, the tags look like this….
Schemas is a Step in the Right Direction

Schemas help....

...by relating common terms between documents

But Other People use Other Schemas

Someone else has one like this....
The Semantics is not there

...which don't fit in

XML Conclusion

- Need to know the price? Just look inside the price tag.
- How do I know that you mean the same thing by <price> that I do?
  - Does that include tax? shipping? surcharges?

- XML provides syntactic interoperability.
- There is a need for semantic interoperability.
Whereas XML gives us structure...

XML

presentation-oriented markup

HTML

content-oriented markup

We Need Semantics

- Web Ontology Language
  - W3C Recommendation
  - released Feb. 2004

markup linked to semantics

bookont ontology
Semantic Web Approach

- For a computer program to reason, it must have a conceptual understanding of the world. This understanding is provided by us. That is, we must provide the computer with an ontology.
- An ontology is specified using a semantic markup language

Ontology is the branch of philosophy that answers the question “what is there?”
An ontology is a conceptualization of a domain
An ontology is typically a hierarchical collection of classes, permissible relationships amongst those classes, and inference rules/constraints that is shared by a community

- An ontology reflects a community’s common understanding of a domain
Semantic Markup Languages

- Stack of languages used to represent semantic content of information objects (records, documents, services, etc.)

![Semantic Web LayerCake](Berners-Lee, 99; Swartz-Hendler, 2001)

Ontologies come in Many Flavors

Informal definition of concepts...

- **PLANNING**: an **ACTIVITY** whose purpose is to produce a **STRATEGY**
- **STRATEGY**: a **PLAN** to **ACHIEVE** a high-level **PURPOSE**
- **PLANNING**: an **ACTIVITY** whose major **EFFECT** is to produce a **PLAN**

Pizza OWL ontology from Protege

Tutorial run by Manchester University (see http://www.co-ode.org/resources/tutorials/)

References:

- <http://www.co-ode.org/ontologies/pizza/2005/10/18/classified/pizza.owl#>
- <http://purl.org/dc/elements/1.1/>
- <http://www.w3.org/2002/07/owl#>
- <http://www.w3.org/2001/XMLSchema#>
- <http://protege.stanford.edu/plugins/owl/protege#>
ISO 15926 Petroleum Ontology

Semantic Web is about Communication

- A machine-processable language

<table>
<thead>
<tr>
<th>Means of communication</th>
<th>English</th>
<th>English example</th>
<th>Semantic Web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphabet</td>
<td>Latin letters</td>
<td>a, b, c, d,</td>
<td>Anything</td>
</tr>
<tr>
<td>Morphological structures</td>
<td>Conjugations</td>
<td>valve-s</td>
<td>No</td>
</tr>
<tr>
<td>Syntactic structures</td>
<td>Sentence structures</td>
<td>valves are produced by Y</td>
<td>XML</td>
</tr>
<tr>
<td>Interpretation</td>
<td>Internal representation Dictionary/thesaurus (or internal)</td>
<td>produce(Y, valve)</td>
<td>OWL</td>
</tr>
<tr>
<td>Reasoning</td>
<td>Internal processes</td>
<td>valves are not animate</td>
<td>Ontology</td>
</tr>
</tbody>
</table>

Description logic reasoning engines
FUTURE OF THE SEMANTIC WEB

Key Capabilities

Ontologies available
Data and services annotated with semantic descriptions that refer to relevant ontologies

- **Near term:**
  - Semantic web services
    - Automatically compose services according to user needs
  - Information integration and/or interoperability
    - Application understands other applications’ data
  - Intelligent search
    - Search on concepts, not term frequencies

- **Long-term:**
  - Model-driven applications
  - Adaptive and autonomic computing
  - Intelligent reasoning
1. A Little Search Example

- Your search query:
  "Show me all universities near the beach."

- The NTNU page does not say anything about the beach, but it does say (through semantic markup) that NTNU is near the Trondheimsfjord.

- Search agent makes use of a geography ontology which includes the rule
  
  \[ \text{Fjord}(x) \rightarrow \text{hasBeaches}(x) \]

- When your search agent parses the NTNU page, it loads in the relevant ontologies, deduces that NTNU is near the beach, and returns the NTNU home page.

2. Semantic Approach to Integrated Operations

- **Architectural principles**
  - Common ontology
  - Reasoning over ontology and semantic annotations
  - Data, services and processes given semantic annotations with respect to ontology
  - Uniform semantic access to all resources
  - Uniform treatment of users and agents

- **Current research**
  - Architectural
  - Theoretical
  - Ontological
  - Application
Vision of Semantic Business Processes

- Knowledge implicit in manual labor
- Static fixed structures
- Knowledge explicit in ontologies
- Distinction between process needs and process realization
- System implements process on the fly
- Reasoning with semantically described services, data and process fragments

From Using Fixed Operational Structures to Composing Operations Dynamically

<table>
<thead>
<tr>
<th>Approach</th>
<th>Information</th>
<th>Services</th>
<th>Operations</th>
<th>Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manually based</td>
<td>Users know where data/services are found</td>
<td>Static, limited</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Search-based</td>
<td>System understands external semantics [ONTOLOGY]</td>
<td>Dynamic, growing, interoperable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composition-based</td>
<td>System understands internal semantics &amp; topics [REASONING]</td>
<td>Fragmented, configurable, contextualized</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Semantic Business Process Management)
Dynamically Constructed Operations

- Workflow specified with respect to ontology
- Workflow engine compatible with reasoning engine
- Web services and data given semantic annotations
- Client may be human or agent

Semantic Web Challenges

- Political/human side
  - Do we accept standards?
  - Early wins?
  - Can we agree on the exact content of the ontology?
- Ontology Engineering Process
  - Need experts in both domain and ontology engineering
  - Stakeholder involvement
  - How to deal with complexity?
- Quality
  - One ontology to rule them all?
  - Level of granularity/precision
  - How to assess the correctness/usability/value of ontologies?
  - Tailoring to information retrieval/mining
  - Personalization
- Technology maturity
  - OWL Full reasoning not Turing complete, intractable
  - Performance issues
  - Limited industrial experience
  - Lack of Semantic Web applications
- Practical application of Semantic Web
  - Who & how to maintain ontologies and make them accessible?
  - Who & how to annotate information objects with semantic descriptions?
Conclusions

- HTML & XML limitations
- Semantic Web technologies enable semantic interoperability
  - Semantic markup languages
  - Ontologies
  - Semantic annotations
- Prospects of semantic technologies
  - Intelligent applications/agents
  - Technological challenges