Defining a Semantic Web Initiative for HEP

Bebo White
bebo@slac.stanford.edu

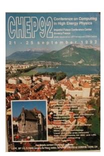
CHEP'04 INTERLAKEN

SLAC STANFORD LINEAR ACCELERATOR CENTER

Many Thanks to Ken Peach!

- For verifying that this BOF was a viable idea
- For establishing some of the concepts and definitions

CHEP and Web Technology



- Annecy - CHEP92
  - WWW introduced to HEP community
  - Most sites in the world were HEP
  - Pre-Mosaic
- San Francisco – CHEP94
  - Tom Nash says "HEP needs to take back WWW!"
- Grid is often viewed as "son of the Web"
- What about Semantic Web?

"Current Web technology is clearly insufficient for the needs of interdisciplinary science and comes up short when it comes to supporting the needs of the collaborative and interdisciplinary "e-Science." Fortunately, new Web technologies are emerging with the potential to revolutionize the ability of scientists to do collaborative work. However, to realize this potential, scientists and information technologists must forge new models of cooperation, and new thinking must go into the funding and dissemination of this next generation of scientific tools on the Web.

A new generation of Web technology, called the Semantic Web, is designed to improve communications between people using differing terminologies, to extend the interoperability of databases, to provide tools for interacting with multimedia collections, and to provide new mechanisms for the support of "agent-based" computing in which people and machines work more interactively."

-- Jim Hendler, Univ. of Maryland

"Unfortunately, most scientists are unaware of the Semantic Web effort, and most of the current development is going on separate from the scientific enterprise. This situation parallels that of the development of the original Web, where scientists largely served as customers and users of Web technology, rather than helping to evolve the technology toward the needs of their fields. In fact, much of the information technology research investment for science has gone into technologies that could not compete with the Web and that ended up less used than the commercially available Web technology. Scientific Web site development is often done by publishers or students in their spare time, and being good at bringing science to the Web is typically not seen as a major career enhancer."

-- Hendler

Towards the Chemical Semantic Web. An introduction to RSS.

Peter Murray-Rust[a] and Henry S. Rzepa[b]

[a]Unilever Centre for Molecular Informatics, University of Cambridge, UK,
[b]Department of Chemistry, Imperial College London, SW7 2AY.

Towards a Semantic Web for Bioinformatics

Dr. Andreas Doms
Biotechnological Centre (BioZ)
Technological University Dresden
adoms@web.de

Much of the biological data currently online is intended for humans: Web pages e.g. contain extensive rendering information beside their contents and do not formally describe the structure of their contents. As a result, it is difficult to automatically process and integrate this information. The semantic web aims to address this

Main Ontologies Browse Project CVS Subscribe Contact

OBO
open biological ontologies

Open Biological Ontologies is an umbrella web address for well-structured controlled vocabularies for shared use across different biological domains.

This site contains ontologies and points to some other efforts within the community. Ideally we see a range of ontologies being designed for biological domains. Some of these will be generic and apply across all organisms and others will be more restricted in scope, for example to specific taxonomic groups.

W3C Technology and Society domain

Web Links: W3C Workshop on Semantic Web for Life Sciences

27-28 October 2004, Cambridge, Massachusetts USA

Nearby: Workshop Home Page | Call For Participation | Public-Semweb-Lifesci Archives | Position Papers Archive

Possible Agenda Items…

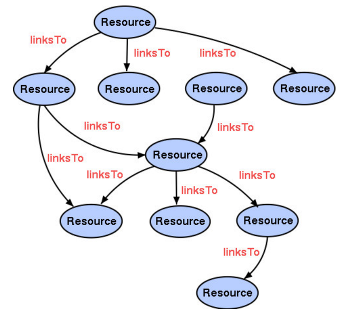
- (To provide a Semantic Web overview)
- To identify interest in Semantic Web
- To share experience with Semantic Web (if any)
- Semantic Web examples
- To discuss whether SW and HEP are a good fit
- If so, a SW HEP initiative
- Identify HEP candidates for SW resources to get started
- Semantic Web can lead to Semantic Grid

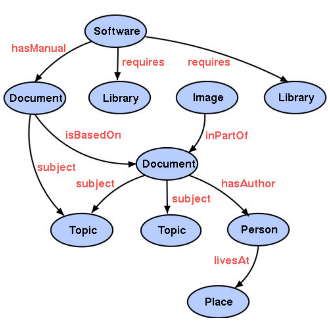
The World Wide Web versus The Semantic Web

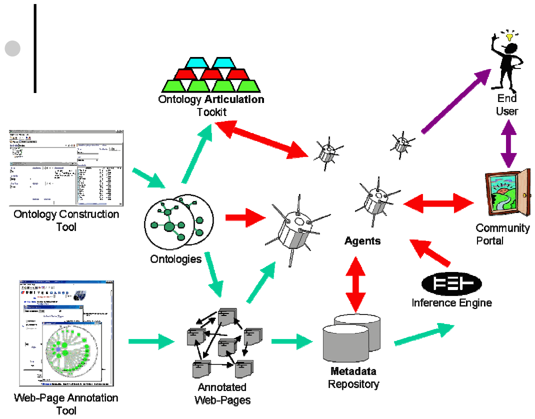
- The World Wide Web is the Web for people
  - Information is predominantly textual
  - Technologies include URI, HTTP, XML, HTML
- The Semantic Web is the Web for machines
  - Information needs to be structured
  - Technologies include RDF, RDFS, OWL (in addition to those for the Web)

Machine Readable Versus Machine Understandable

- In the World Wide Web, information needs humans to give it interpretation
  - Information is predominantly natural language
  - Difficult to mediate by software agents
- In the Semantic Web, information is structured so that it can be interpreted by machines
  - Humans need not interact directly with Semantic Web information – mediation through agents
- Formal meaning is critical to understanding







Metadata

- The origins of the Semantic Web lie in metadata
- Metadata is data about data
  - A webpage is data
  - A description of the webpage is metadata
  - Metadata for a webpage could include
    - author
    - date of publication
    - file size
    - …
- Library cataloguing = metadata

Beyond Metadata

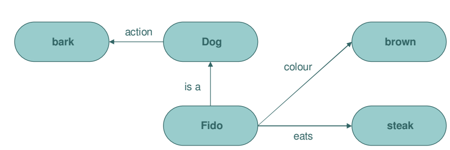
- The scope of the modern Semantic Web goes beyond bibliographic metadata for webpages
- Metadata is still just data
- If we have an infrastructure for metadata, we can use it for data in general

Knowledge Representation

- Long-standing discipline within Artificial Intelligence
  - (the Semantic Web has a strong heritage!)
- Knowledge representation languages should:
  - Handle qualitative knowledge
  - Allow new knowledge to be inferred
  - Represent both the general and the specific
  - Capture complex meaning
  - Allow meta-level reasoning
- RDF, RDF Schema and OWL are knowledge representation languages

Network Knowledge Representation

- "Traditional" knowledge representation is formal logic
- Network knowledge representation originated in 1960s with psychologists and linguists
- Knowledge is represented as a graph
  - Nodes are objects or concepts
  - Edges are relations or associations



Network Knowledge Representation

- Many types of network KR
  - Conceptual graphs
  - Semantic networks
  - Conceptual dependency theory
  - (and the Semantic Web…)
- Close correspondence with other KR techniques (logic, frames, scripts, etc)
  - A different way of viewing knowledge

Vocabularies and Ontologies

- A knowledge representation language by itself is of little use
- We need to be able to tailor the language to our application domain
  - The bibliographic domain needs to be able to talk about works and authors
  - The e-commerce domain needs to be able to talk about orders and prices
  - …
- We need domain-specific vocabularies and ontologies

What is an Ontology?

- "a specification of a conceptualization"
  - specification: a formal description
  - conceptualization: a model of the objects and structures in some application domain
- Set of concepts (e.g. entities, classes, attributes, processes), their definitions and inter-relationships
- Vocabulary for describing this set of concepts

What is an Ontology?

- Shared understanding
- Facilitate communication
  - Establish a joint terminology for a community of interest
  - Normative models…
- Inter-operability: sharing and reuse
  - Inter-lingua…
- Ontologies are predominantly designed artifacts
  - But, substantial research on automatic construction

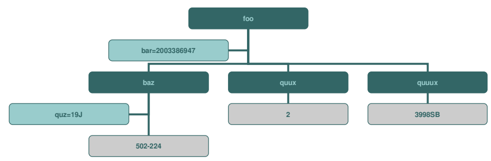
Machine Readable Versus Machine Understandable

- XML is a machine readable format
  - It can be parsed to give an unambiguous document structure

  but
  - It has no formal meaning
  - Meanings of XML interchange formats must be explicitly agreed

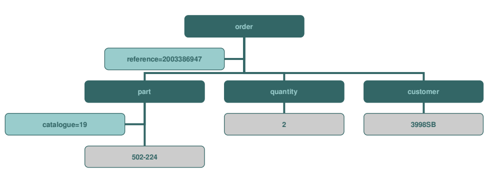
Machine Readable: XML

```
<foo bar="2003386947">
  <baz qux="19J">502-224</baz>
  <quux>2</quux>
  <quuux>3998SB</quuux>
</foo>
```



Machine Readable: XML

```
<order reference="2003386947">
  <part catalogue="19J">502-224</part>
  <quantity>2</quantity>
  <customer>3998SB</customer>
</order>
```



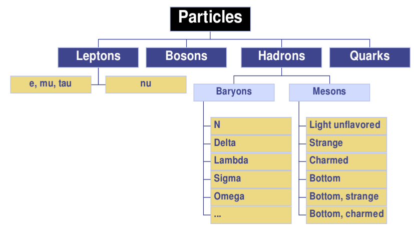
Example of Particle Properties in XML

- YAPPI – Yet Another Particle Property Interface
- Part of FreeHEP Java Library
- Uses Java API for reading and writing
- JAS – Java Analysis Studio

Particles and Properties

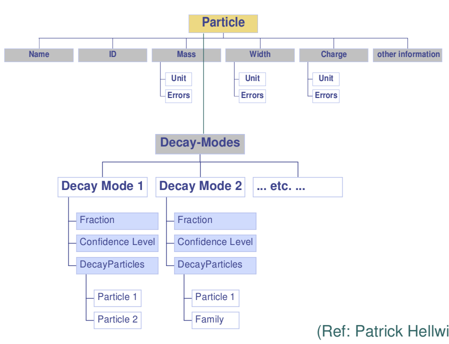
- Properties of Particles are needed for analysis
- Standard source of data: Particle Data Group (PDG)
  - But no database !
  - Each software has its own database
- Two areas:
  - Particles
    - General Information (Mass, Width, etc.)
    - Decay Modes
  - Family structure

Data Model: Particle-Families



(Ref: Patrick Hellwig)

Data Model: Particle



(Ref: Patrick Hellwig)

XML Technology

http://www.w3.org

- Multiple parts:
  - Data-file (XML)
  - Schema-file (XSD/DTD)
  - Style-file (XSL)
- Elements
- Attributes

XML:

```
<Particle Name="Proton">
    <ID>2212</ID>
    <Mass Unit="MeV">
        938.272
    </Mass>
</Particle>
```

Resource Description Framework

- RDF is a language for representing information about resources on the World Wide Web and beyond
- RDF uses Uniform Resource Identifiers to identify things and the relations between them

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"
  xmlns:dc="http://purl.org/DC/>
  <rdf:Description
about="http://slac.stanford.edu/xml/>
    <dc:Creator>Bebo White</dc:Creator>
    <dc:Title>The Dutch Goose</dc:Title>
  </rdf:Description>
</rdf:RDF>
```

Machine Readable Versus Machine Understandable

- RDF is a machine understandable format
  - The structures generated by an RDF parser have a formal meaning
  - RDF is a framework for interchange formats that provides a base level of common understanding
  - RDF provides basic notions of classes and properties
  - RDF enables simple inference
    - RDF permits certain types of deduction to be made from existing knowledge

Limitations of RDF

- RDF lets us make assertions about resources using a given vocabulary
- RDF does not let us define these domain vocabularies by itself
- RDF Schema is an RDF vocabulary which we can use to define other vocabularies

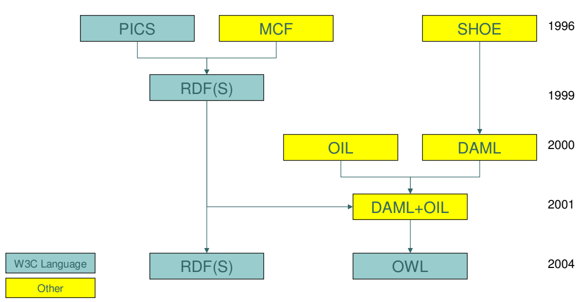
Defining an RDF Vocabulary

- Define classes of objects and their relationship with other classes
  - "there is a class called Employee which is a subclass of the class called Person"
- Define properties that relate objects together and their characteristics
  - "there is a property called worksFor which relates objects of class Employee to objects of class Company"

Fundamental Principles

- Anyone can make assertions about anything
- Entities are referred to using Uniform Resource Identifiers
  - URI-based vocabularies
- Based on XML technologies
- Formal semantics

Languages of the Semantic Web



The Triple

- Underlying model of *triples* used to describe the relations between entities in the Semantic Web
- (*subject, predicate, object)*
  - e.g. "RDF Semantics", "edited by", "Pat Hayes"

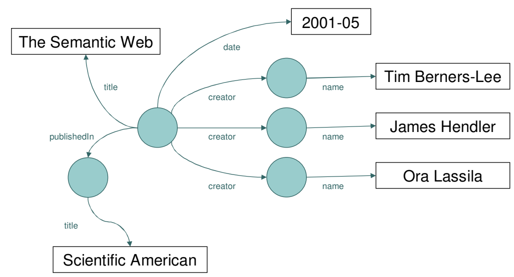
- Network knowledge representation
  - Labelled, directed graph
  - Entities as nodes, relations as edges

Example: Scientific American article

- Take a citation:
  Tim Berners-Lee, James Hendler and Ora Lassila. The Semantic Web. *Scientific American*, May 2001
- We can identify a number of distinct statements in this citation:
  - There is an article titled “The Semantic Web”
  - One of its authors is a person named “Tim Berners-Lee” (etc)
  - It appeared in a publication titled “Scientific American”
  - It was published in May 2001

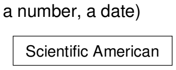
Example: Scientific American article

- We can represent these statements graphically

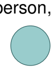


Example: Scientific American article

- There are two types of node in this graph:
  - Literals, which have a value
    (a string, a number, a date)



  - Resources, which represent objects with identity
    (a web page, a person, a journal)

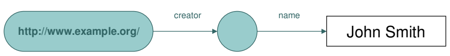


Example: Scientific American article

- Resources are identified by URIs
- Property labels are identified by URIs, and are drawn from a vocabulary or ontology

Blank nodes (bNodes)

- Sometimes we have resources which we do not wish to identify with a URI
- These are *blank nodes* or *anonymous resources*



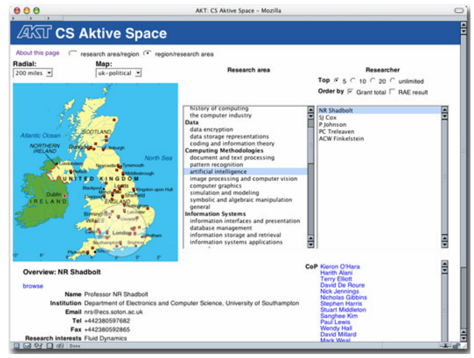
How Might Semantic Web Be Used in HEP?

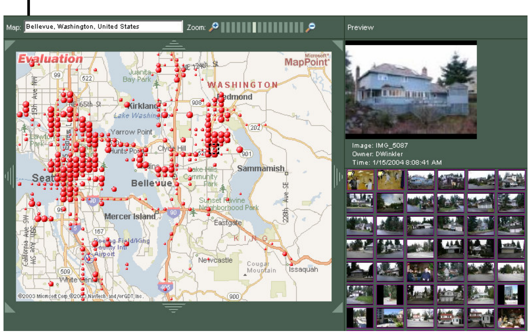
- PDG – extend the work of YAPPI
- HEP Physicist Info (e.g., FOAF)
- HEP Image Annotation (e.g., W3Photo)
- Archives/Preprints/Citations (e.g., Scientific American example)
- Etc.,etc.

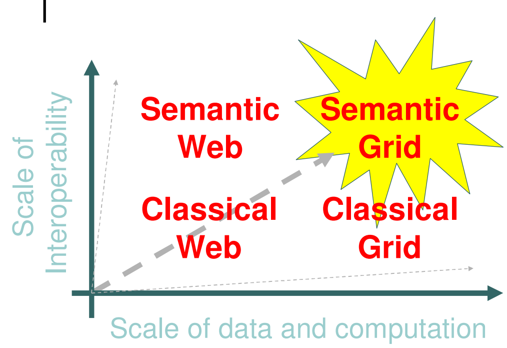
Examples

- RSS
- FOAF - http://www.ldodds.com/foaf/foaf-a-matic.html
- W3 Photo Project - http://w3photo.org/photos/
- CS Aktive Space - http://www.aktors.org/technologies/csaktivespace/
- World-Wide Media Exchange - http://wwmx.org/

```
<dc:creator>
  <foaf:Agent>
    <foaf:homepage rdf:resource='http://duhblog.com'/>
    <foaf:name>Bebo White</foaf:name>
    <foaf:mbox_sha1sum>22d778f17d97f4fdea46c47423dc3a6ed4541fec</foaf:mbox_sha1sum>
  </foaf:Agent>
 </dc:creator>
 <dc:date></dc:date>
 <foaf:depicts>
 <conf:Event>
  <rdf:type rdf:resource='http://www.w3.org/2002/12/cal/ical#Vevent'/>
   <ical:url rdf:resource='http://www.w3photo.org/photos/www1'/>
 </conf:Event>
 </foaf:depicts>
</foaf:Image>
<foaf:Image rdf:about="http://w3photo.org/photos/www1/photos/Closingpanel5.JPG">
<fotonotes:ID>.photos.www1.Closingpanel5.JPG</fotonotes:ID>
<fotonotes:datecreated>2004-Jun-01T00:02:06EDT</fotonotes:datecreated>
<fotonotes:serial>3</fotonotes:serial>
<fotonotes:selectionCounter>1</fotonotes:selectionCounter>
<fotonotes:authorization  rdf:parseType='Resource'>
 <fotonotes:username>bebo</fotonotes:username>
<fotonotes:shareLevel>2</fotonotes:shareLevel>
</fotonotes:authorization>

 <imreg:hasRegion>
 <imreg:Rectangle rdf:ID='p0'>
  <dc:description>Answering questions at the first conference.</dc:description>
  <imreg:coords>291,107 387,219</imreg:coords>
  <imreg:boundingBox>291,107,96,112</imreg:boundingBox>
<dc:creator>
```



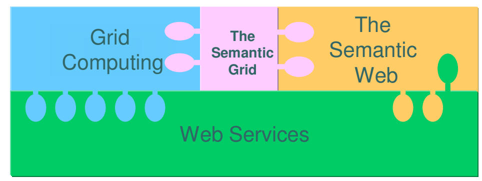


Based on an idea by Norman Paton

Semantic Grid Definition

The Semantic Grid is an extension of the current Grid in which information and services are given well-defined meaning, better enabling computers and people to work in cooperation

Semantics in and on the Grid



Acknowledgements

- Tim Berners-Lee
- David de Roure
- Nick Gibbins
- Carole Goble
- Peter Hellwig

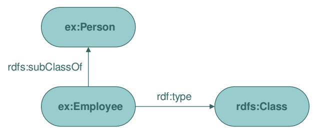
RDF Schema Class Definitions

- We wish to define the class Person


```
<rdf:Description rdf:about="#Person">
  <rdf:type rdf:about="&rdfs;Class"/>
</rdf:Description>


<rdfs:Class rdf:about="#Person"/>
```

RDF Schema Class Definitions

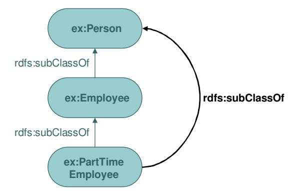
- Employee is a subclass of Person




```
<rdfs:Class rdf:about="#Employee">
  <rdfs:subClassOf rdf:resource="#Person"/>
</rdfs:Class>
```

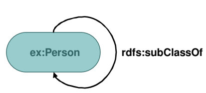
RDF Schema Class Semantics

- rdfs:subClassOf is transitive
  - (A rdfs:subClassOf B) and (B rdfs:subClassOf C) implies (A rdfs:subClassOf C)



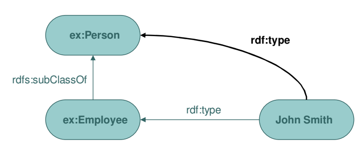
RDF Schema Class Semantics

- rdfs:subClassOf is reflexive
  - All classes are subclasses of themselves



RDF Schema Class Semantics

- rdf:type distributes over rdf:subClassOf
  - (A rdfs:subClassOf B) and (C rdf:type A) implies (C rdf:type B)



RDF Schema Property Definitions

- We wish to define the property worksFor

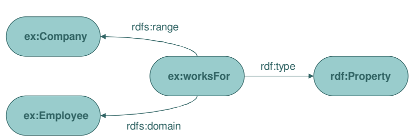

```
<rdf:Description rdf:about="#worksFor">
  <rdf:type rdf:resource="&rdf;Property"/>
</rdf:Description>


<rdf:Property rdf:about="#worksFor"/>
```

RDF Schema Property Definitions

- Important difference between RDF and object oriented programming languages
  - OO languages define classes in terms of the properties they have
  - RDF defines properties in terms of the classes whose instances they relate to each other
- The *domain* of a property is the class that the property runs *from*
- The *range* of a property is the class that a property runs *to*

RDF Schema Property Definitions

- The property worksFor relates objects of class Employee to objects of class Company




```
<rdf:Property rdf:about="#worksFor">
  <rdfs:domain rdf:resource="#Employee"/>
  <rdfs:range rdf:resource="#Company"/>
</rdf:Property>
```