The Semantic Web

(Slides by Fabian M. Suchanek)
Motivation

Australia's scientists visit Brisbane
The National Science Education Unit invites Australian scientists to gather in Brisbane
www.nsceu.au/brisbane

Today's state of the art

<HTML>
   Sam Smart is a scientist from Brisbane.
</HTML>

Vision of the Semantic Web

bornIn

Brisbane

label

„Sam Smart“
The Semantic Web

The Semantic Web is the project of creating a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.

**Goals:**

- make computers „understand“ the data they store
- allow them to answer „semantic“ queries
- allow them to share information across different systems

**Techniques:** (= this talk)

- defining semantics in a machine-readable way (RDFS)
- identifying entities in a globally unique way (URIs)
- defining logical consistency in a uniform way (OWL)
- linking together existing resources (LOD)

http://www.w3.org/2001/sw/
The Resource Description Framework (RDF)

RDF is a format of knowledge representation that is similar to the Entity-Relationship-Model.

RDF is used as the only knowledge representation language.

=> All information is represented in a simple, homogeneous, computer-processable way.

http://www.w3.org/TR/rdf-primer/
n-ary relationships

n-ary relationships can always be reduced to binary relationships by introducing a new identifier.
Uniform Resource Identifiers (URIs)

A URI is similar to a URL, but it is not necessarily downloadable. It identifies a concept uniquely.

```
SamSmart: http://brisbane-corp.au/people/SamSmart
bornIn:  http://mpii.de/yago/resource/bornIn
Brisbane: http://brisbane.au
```

URIs are used as globally unique identifiers for resources.

=> Knowledge can be interlinked. A knowledge base on one server can refer to concepts from another knowledge base on another server.
Namespaces

A namespace is a shorthand notation for the first part of a URI.

Without namespaces, our statement is a triple of 3 URIs -- quite verbose.


Namespaces make our statement much less verbose.

bsco:SamSmart  yago:bornIn  <http://brisbane.au>

Namespaces are used to abbreviate URIs

=> Namespaces with useful concepts can become popular.
   This facilitates a common vocabulary across different knowledge bases.
Popular Namespaces: Basic

rdf:    The basic RDF vocabulary
        http://www.w3.org/1999/02/22-rdf-syntax-ns#

rdfs:   RDF Schema vocabulary (predicates for classes etc., later in this talk)
        http://www.w3.org/1999/02/22-rdf-syntax-ns#

owl:    Web Ontology Language (for reasoning, later in this talk)
        http://www.w3.org/2002/07/owl#

dc:     Dublin Core (predicates for describing documents, such as „author“, „title“ etc.)
        http://purl.org/dc/elements/1.1/

xsd:    XML Schema (definition of basic datatypes)
        http://www.w3.org/2001/XMLSchema#

Standard namespaces are used for basic concepts
=> The basic concepts are the same across all RDF knowledge bases
Popular Namespaces: Specific

**dbp:** The DBpedia ontology (real-world predicates and resources, e.g. Albert Einstein)
http://dbpedia.org/resource/

**yago:** The YAGO ontology (real-world predicates and resources, e.g. Albert Einstein)
http://mpii.de/yago/resource/

**foaf:** Friend Of A Friend (predicates for relationships between people)
http://xmlns.com/foaf/0.1/

**cc:** Creative Commons (types of licences)
http://creativecommons.org/ns#

.... and many, many more

There exist already a number of specific namespaces
=> Knowledge engineers don't have to start from scratch
We are using standard RDF vocabulary here. The objects of statements can also be literals. The literals can be typed. Types are identified by a URI.

Popular types: xsd:string xsd:date xsd:nonNegativeInteger xsd:byte

 Literals are can be labeled with pre-defined types => They come with a well-defined semantics.

http://www.w3.org/TR/xmlschema-2/
Classes

A class is a resource that represents a set of similar resources.

```
example:SamSmart         yago:bornIn         <http://brisbane.au>
example:SamSmart         rdf:type                example:scientist
example:scientist             rdfs:subclassOf   example:person
```

Due to historical reasons, some vocabulary is defined in RDF, other in RDFS.

http://www.w3.org/TR/rdf-schema/
„Meta-Data“

Meta-Data is data about classes and properties

Properties themselves are resources in RDF

yago:bornIn     rdf:type     rdf:Property
yago:bornIn     rdfs:domain  example:person
yago:bornIn     rdfs:range   example:city
example:person  rdf:type     rdfs:Class
rdfs:Class      rdf:type     rdfs:Class

RDFS can be used to talk about classes and properties, too
=> There is no concept of „meta-data“ in RDFS

http://www.w3.org/TR/rdf-schema/
The OWL vocabulary can be used to express properties of classes and predicates
=> We can express logical consistency

http://www.w3.org/TR/owl-guide/
Reasoning: Flavors of OWL

There exist 3 different flavors of OWL that trade off expressivity with tractability.

OWL Full is very powerful, but undecideable.

OWL DL has the expressive power of Description Logics.

OWL Lite is a simplified subset of OWL DL.

http://www.w3.org/TR/owl-guide/
Formats of RDF data

RDF is just the model of knowledge representation, there exist different formats to store it.

1. In a database („triple store“) with the schema

   \[ \text{FACT(resource, predicate, resource)} \]

2. As triples in plain text („Notation 3“, „Turtle“)

   ```
   @prefix yago http://mpii.de/yago/resource
   yago:SamSmart yago:bornIn <http://brisbane.au>
   ```

3. In XML

   ```
   <?xml version="1.0"?>
   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:yago="http://mpii.de/yago/resource">
      <rdf:Description rdf:about="http://mpii.de/yago/resource/SamSmart">
         <yago:bornIn rdf:resource="http://brisbane.au"/>
      </rdf:Description>
   </rdf:RDF>
   ```
There exist already a number of knowledge bases in RDF.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>URL</th>
<th>#Statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freebase</td>
<td><a href="http://www.freebase.com">http://www.freebase.com</a></td>
<td>2.5m</td>
</tr>
<tr>
<td>(community collaboration)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OpenCyc</td>
<td><a href="http://www.opencyc.org">http://www.opencyc.org</a></td>
<td>60k</td>
</tr>
<tr>
<td>(spin-off from commercial ontology Cyc)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DBpedia</td>
<td><a href="http://www.dbpedia.org">http://www.dbpedia.org</a></td>
<td>270m</td>
</tr>
<tr>
<td>(extraction from Wikipedia, focus on coverage)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YAGO</td>
<td><a href="http://mpii.de/yago">http://mpii.de/yago</a></td>
<td>20m</td>
</tr>
<tr>
<td>(extraction from Wikipedia, focus on accuracy)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Existing OWL/RDF knowledge bases: Specific

<table>
<thead>
<tr>
<th>Dataset</th>
<th>URL</th>
<th>#Statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>MusicBrainz</td>
<td><a href="http://www.musicbrainz.org">http://www.musicbrainz.org</a></td>
<td>23k</td>
</tr>
<tr>
<td>(Artists, Songs, Albums)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geonames</td>
<td><a href="http://www.geonames.org">http://www.geonames.org</a></td>
<td>85k</td>
</tr>
<tr>
<td>(Countries, Cities, Capitals)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DBLP</td>
<td><a href="http://www4.wiwiss.fu-berlin.de/dblp/">http://www4.wiwiss.fu-berlin.de/dblp/</a></td>
<td>15m</td>
</tr>
<tr>
<td>(Papers, Authors, Citations)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>US Census</td>
<td><a href="http://www.rdfabout.com/demo/census">http://www.rdfabout.com/demo/census</a></td>
<td>1000m</td>
</tr>
<tr>
<td>(Population statistics)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...and many more....

=> The Semantic Web has already a reasonable number of knowledge bases

http://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/DataSets
The Linking Open Data Project

yago:AlbertEinstein  owl:sameAs  dbpedia:Albert_Einstein

http://esw.w3.org/topic/SweolG/TaskForces/CommunityProjects/LinkingOpenData
Querying the knowledge bases: SPARQL

SPARQL is a query language for RDF data. It is similar to SQL.

Example:
Which scientists are from Brisbane?

```
PREFIX rdf:http://www.w3.org/1999/02/22-rdf-syntax-ns#
PREFIX example:....

SELECT ?x WHERE {
  ?x      rdf:type                  example:scientist .
  ?x      example:bornIn     example:Brisbane .
}
```

http://www.w3.org/TR/rdf-sparql-query/
Sample Query on YAGO

Which scientists are from Brisbane?

http://mpii.de/yago
References

Specifications

RDF  http://www.w3.org/TR/rdf-primer/
RDFS  http://www.w3.org/TR/rdf-schema/
URIs  http://www.ietf.org/rfc/rfc3986.txt
Literals  http://www.ietf.org/rfc/rfc3986.txt
OWL  http://www.w3.org/TR/owl-guide/
SPARQL  http://www.w3.org/TR/rdf-sparql-query/

Projects

YAGO  Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum
„YAGO - A Core of Sematic Knowledge“ (WWW 2007)

DBpedia  S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, Z. Ives
„DBpedia: A Nucleus for a Web of Open Data“ (ISWC 2007)

LOD  Christian Bizer, Tom Heath, Danny Ayers, Yves Raimond
„Interlinking Open Data on the Web“ (ESWC 2007)