A brief introduction to the Semantic Web

Yaron Koren
National Library of Israel
August 11, 2011
About me

• Software developer

• Grew up in Haifa and Massachusetts (U.S.), live in New York City

• Wikipedia enthusiast since 2005

• Semantic MediaWiki enthusiast since 2006

• MediaWiki developer and consultant since 2007
My MediaWiki consulting company:

![WikiWorks MediaWiki consulting](wikiworks.com)

(wikiworks.com)

Via WikiWorks, also run the Semantic MediaWiki-based wiki farm Referata, at referata.com.
What is the Semantic Web?
Semantic = Meaning

(the opposite of syntax/"syntactic", which is the actual words used)
“Semantic Web” means different things to different people. Semantic information on the web can be expressed in three different ways:

1. Inferred
2. Free-form tagging
3. Structured data

These go in order from the consumer of data doing most of the work, to the producer of data doing most of the work.
1. Inferred information

From a website:

“Benny's Shoes is open from 9 AM to 5 PM Monday to Friday, and 10 AM to 7 PM on Sundays. Our phone number is 123-4567.”

Hm... that looks like hours, and a phone number, for a store.
2. Free-form tagging

“Benny's Shoes is open from \(<time itemprop="openingHours" datetime="Mo-Fr 9:00-17:00">9 AM to 5 PM Monday to Friday</time>\), and \(<time itemprop="openingHours" datetime="Su 10:00-19:00">10 AM to 7 PM on Sundays</time>\). Our phone number is \(<span itemprop="telephone">123-4567</span>\).

This is in the “schema.org” microformat.
Microformats are additional tags and attributes in HTM that let you encode meaning.

schema.org is especially important, because it is supported by the Google, Yahoo! and Bing search engines since June 2011.

Similar to microformats is RDFa, which is considered more semantic, and preferred by academics.
3. Structured data

From a database of stores:

<table>
<thead>
<tr>
<th>Name</th>
<th>Product</th>
<th>Telephone</th>
<th>Sunday opening hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Benny’s Shoes</td>
<td>Shoes</td>
<td>123-4567</td>
<td>10 AM</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
For types 2 and 3, the data can be exported as triples.

(For some people, “Semantic Web” = triples.)

**Semantic triple:**

Subject Relationship Object

**Example:**

Benny's shoes Telephone 123-4567
RDF and OWL

RDF = Resource Description Framework
- a framework for storing semantic triples

RDF/XML – a file format for storing RDF data

OWL = Web Ontology Language
- a superset of RDF – used to store both data and information about data structure
1. Inferred information

Advantages:

Untagged information is everywhere! 99.99% of all information is untagged.

Disadvantages:

Much potential for error.
2. Free-form tagging

**Advantages:**

Lets everyone publish any type of semantic data themselves.

**Disadvantages:**

Complicated to do.
3. Structured data

Advantages:
Easy to input data; easy to extract and publish data in a variety of formats.

Disadvantages:
Requires creating a separate system.
For *querying* semantic data, the same basic three options exist:

1) Unstructured language queries
2) Free-form data queries
3) Structured queries
1. Unstructured language queries

(Typed in a search engine:)

“Shoe stores open late on Wednesdays”
2. Free-form data queries

SELECT ?storeName
WHERE {
    ?x storesDB:name ?storeName ;
    storesDB:storeType storesDB:shoeStore ;
    storesDB:closingHoursWednesday > 7 .
}

This is a SPARQL query.
SPARQL

SPARQL = SPARQL (formerly Simple) Protocol and RDF Query Language

- standard language for querying and modifying RDF data
3. Structured queries

In a web/mobile/etc. application:
Wikipedia already contains a lot of data!

Can we query that data?
Yes and no.

Wikipedia cannot be queried directly.

The “DBpedia” site (dbpedia.org) contains information from 3.5 million of the pages in Wikipedia, from English and other languages, in a format that can be queried (RDF).

From their website: “DBpedia is the Semantic Web mirror of Wikipedia.”
In addition, there are ongoing discussions about adding this capability directly into Wikipedia. This may happen in the next few years.
Wikipedia actually has two big roles in the Semantic Web:

1) A huge source of data
2) A set of “canonical” URLs, to define many real-world entities

“In my set of data, 'Yugoslavia' is the entity defined by http://en.wikipedia.org/wiki/Kingdom_of_Yugoslavia”

(RDF lets you express such a thing directly.)
The growing “Linked Open Data” cloud image:

DDpedia is the big circle in the middle – everyone uses its URLs!