State of the Semantic Web

Ivan Herman, W3C

Sunday, 18 May, 2008
> So where are we with the Semantic Web?
We have the basic technologies

- Stable specifications for the basics since 2004: RDF, OWL
- Work is being done to properly incorporate rules
- We have a standard for query since 2008: SPARQL
- We have some additional technologies to access/create RDF data: GRDDL, RDFa, POWDER, …
- Some fundamental vocabularies became pervasive (FOAF, Dublin Core, …)
> Lots of Tools (not an exhaustive list!)

- **Triple Stores**
  - RDFStore, AllegroGraph, Tucana
  - RDF Gateway, Mulgara, SPASQL
  - Jena’s SDB, D2R Server, SOR
  - Virtuoso, Oracle11g
  - Sesame, OWLIM, Tallis Platform
  - …

- **Reasoners**
  - Pellet, RacerPro, KAON2, FaCT++
  - Ontobroker, Ontotext
  - SHER, Oracle 11g, AllegroGraph
  - …

- **Converters**
  - flickurl, TopBraid Composer
  - GRDDL, Triplr, jpeg2rdf
  - …

- **Search Engines**
  - Falcon, Sindice, Swoogle
  - …

- **Middleware**
  - IODT, Open Anzo, DartGrid
  - Ontology Works, Ontoprise
  - Profium Semantic Information Router
  - Software AG’s EII, Thetus Publisher, Asio, SDS
  - …

- **Semantic Web Browsers**
  - Disco, Tabulator, Zitgist, OpenLink Viewer
  - …

- **Development Tools**
  - SemanticWorks, Protégé
  - Jena, Redland, RDFLib, RAP
  - Sesame, SWI-Prolog
  - TopBraid Composer, DOME
  - …

- **Semantic Wiki and CMS systems**
  - Semantic Media Wiki, Platypus
  - Visual knowledge
  - Drupal 7

Inspired by “Enterprise Semantic Web in Practice”, Jeff Pollock, Oracle. See also W3C’s Wiki Site.
Lots of tools (cont.)

- Significant speed, store capacity, etc; improvements are reported every day
- Some of the tools are open source, some are not; some are very mature, some are not: it is the usual picture of software tools, nothing special any more!
- **Anybody can start developing RDF-based applications today**
There is a great community

- There are *lots* of tutorials, overviews, and books around
  - again, some of them good, some of them bad, just as with any other areas…
- Active developers’ communities
  - blogs, IRC channels, mailing lists, various fora: more than what one person can oversee…
- Some measures claim that there are over $10^7$ Semantic Web documents on the Web
> Some deployment communities

- Major communities pick the technology up: digital libraries, defense, eGovernment, energy sector, financial services, health care, oil and gas industry, life sciences …

- Semantic Web also appears in the “Web 2.0/Web 3.0” world (whatever that means 😊)
  - exchange of social data
  - personal “space” applications
  - multimedia asset management (video, photos, audio, …)
  - etc
> So what is the Semantic Web?

- There is a growing number of application patterns referring to the Semantic Web:
  - data integration using RDF, SKOS, OWL, …
  - knowledge engineering with complex ontologies
  - better data management, archiving, cataloging, digital libraries, …
  - managing, coordinating, combining Web services
  - intelligent software agents
  - improving search (usually using domain specific vocabularies…)
  - etc
  - and, of course, mixtures of these…

- But: what binds these all together?
> Is this where we are?
So what is the Semantic Web? (cont.)

- Maybe, but being an elephant is not necessary bad!

- it shows that the Semantic Web is a mature technology
- that there is lots of interest, applications
- various application areas pick what they need…
  - e.g., some need sophisticated knowledge management, so they go for complex ontologies…
  - some concentrate on semantically simpler vocabularies but large volume of data

  ...and that is fine, there is room for many!
But it is good to (re-)emphasize some principles

The Semantic Web:

- a unique way of specifying data and data relationships
- extend principles of the Web from documents to data; create a Web of data
- it is the Semantic Web, and not only Semantics
  - data, ontologies, vocabularies, etc, should be shared, reused, potentially on Web scale
  - e.g., one can use the Web infrastructure to denote “things”…
    - Eg: http://www.ivan-herman/me denotes, well, me (not my home page, not my foaf file, but me!)
  - ... and add relationships for those, too!
A few words about “newer” technologies
Querying RDF: SPARQL

- Querying RDF graphs is essential (can you imagine Relational Databases without SQL?)
- SPARQL is
  - a query language based on graph patterns
  - a protocol layer to use SPARQL over, eg, HTTP
  - an XML return format for the query results
- Numerous implementations are already available (eg, built in triple stores)
There are also SPARQL “endpoints” on the Web:

- send a query and a reference to data over HTTP GET, receive the result in XML or JSON
  - some of those can be easily installed on any machine
- big datasets often offer “SPARQL endpoints” to their local data
  - applications may not need any direct RDF programming any more, just use a SPARQL processor

SPARQL can also be used to construct graphs!
The power of CONSTRUCT

CONSTRUCT {
}
WHERE {
}

- SPARQL endpoint
- returns RDF/XML

SELECT *
FROM <http://dbpedia.org/sparql/?query=CONSTRUCT+%7B++…>
WHERE {
        rdf:type          ?type;
        foaf:name         ?foaf_name.
    FILTER regex(str(?type),"foaf")
}

- Data reused in a query elsewhere…
A word of warning on SPARQL…

- Some features are missing
  - control and/or description on the entailment regimes of the triple store (RDFS? OWL-DL? OWL-Lite? …)
  - modify the triple store
  - querying collections or containers may be complicated
  - no functions for sum, average, min, max, …
  - ways of aggregating queries
  - …

- Delayed for a next version…
> But: how do you get the data on the SW?
Public datasets are accumulating

- IgentaConnect bibliographic metadata storage: over 200 million triplets
- RDFS/OWL Representation of WordNet: also downloadable as 150MB of RDF/XML
- “Département/canton/commune” structure of France published by the French Statistical Institute
- Geonames Ontology and Data: 6 million (and growing) geographical features
- RDF Book Mashup: book data from, eg, Amazon,
- “dbpedia”: infobox data of Wikipedia into RDF
- Note the “Billion Triple Challenge 2008”!
How to get RDF data?

- Of course, one could create RDF data manually…
- … but that is unrealistic on a large scale
- Goal is to generate RDF data automatically (when possible) and “fill in” by hand only when necessary
- Different data formats should be considered
  - databases (relational or otherwise)
  - data in XML, HTML, in pictures, videos, etc
- Details of the process is still subject of very active R&D!
Bridge to relational databases

- Huge amount of data are stored in (relational) databases
  - "RDFying" them is impossible
- “Bridges” are being defined:
  - a layer between RDF and the relational data
    - RDB tables are “mapped” to RDF graphs, possibly on the fly
  - systems can now be used as relational database as well as triple stores (e.g., Oracle, OpenLink, …)
- Work for a survey on mapping techniques has just started at W3C
- SPARQL is becoming the tool of choice to query the data (via “SPARQL Endpoints”)
Linking Open Data Project

- Goal: “expose” open datasets in RDF
- Set RDF links among the data items from different datasets
- Set up SPARQL endpoints to query the data, too

- billions of triples
- millions of “links”
Example data source: DBpedia

- **DBpedia** is a community effort to
  - extract structured (“infobox”) information from Wikipedia
  - provide a SPARQL endpoint to the dataset
  - interlink the DBpedia dataset with other datasets on the Web
Structured data from Wikipedia

http://en.wikipedia.org/wiki/San_Jose%2C_California

<http://dbpedia.org/resource/San_Jose%2C_California>

dbpedia:areaTotalSqMi "178.2"^^xsd:double;
dbpedia:areaTotalKm "461.5"^^xsd:double;
dbpedia:populationTotal "929936"^^xsd:double;
dbpedia:leaderName
dbpedia:Chuck_Reed, "Debra Figone";
...

City of San Jose

Flag

Government
- Type charter city, mayor-council
- Mayor Chuck Reed
- Vice Mayor Dave Cortese
- City Manager Debra Figone
- Senate List of Senators
- Assembly Assembly List

Area [1]
- City 178.2 sq mi (461.5 km²)
- Land 174.9 sq mi (452.9 km²)
- Water 3.3 sq mi (3.6 km²)
- Urban 260.11 sq mi (673.68 km²)
- Metro 2,694.7 sq mi (6,979.4 km²)

Elevation [2] 95 ft (29 m)

Population [2006][3][4][5]
- City 920,038 (10th)
- Density 5,216.3/sq mi (2,014.4/km²)
- Urban 1,611,000
- Metro 7,264,887
- Demonym San Josean
Automatic links among open datasets

Linking DBPedia and the US Census data:

```
<http://dbpedia.org.../San_Jose%2C-California>
  owl:sameAs
    <http://.../usgov/geo/.../santa_clara_county/san_jose>;
...
```

```
<http://.../usgov/geo/.../santa_clara_county/san_jose>
  owl:sameAs <http://DBpedia.org/.../San_Jose%2C-California>
  census:households "559949"^^xsd:integer;
  wgspos:lat "37.318892"^^xsd:double;
...
```

Processors can switch automatically from one to the other…
This is a **major** community project

- anybody can participate; to subscribe to the list:
  - http://lists.w3.org/Archives/public/public-lod/
- or look at the project site:
  - http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
- if you know of open data sets: contact the project to incorporate it with the rest!

Applications using this set of data in real-life setting should come to the fore soon
Data may be around already...

- Part of the (meta)data information is present in tools ... but thrown away at output
  - e.g., a business chart can be generated by a tool: it “knows” the structure, the classification, etc. of the chart, but, usually, this information is lost
  - storing it in web data would be easy!
- “SW-aware” tools are around (even if you do not know it...), though more would be good:
  - Photoshop CS stores metadata in RDF in, say, jpg files (using XMP)
  - ...
Data may be extracted (a.k.a. “scraped”)

- Different tools, services, etc, come around:
  - get RDF data associated with images, for example:
    - service to get RDF from flickr images
    - service to get RDF from XMP
  - scripts to convert spreadsheets to RDF
  - etc

- Many of these tools are still individual “hacks”, but show a general tendency

- Hopefully more tools will emerge
  - there is a separate wiki page collecting references to existing ones
Getting structured data to RDF: GRDDL

- GRDDL is a way to access structured data in XML/XHTML and turn it into RDF:
  - defines XML attributes to bind a suitable script to transform (part of) the data into RDF
  - a “GRDDL Processor” runs the script and produces RDF on–the–fly
- A way to access existing structured data and “bring” it to RDF
  - eg, a possible link to microformats
  - exposing data from large XML use bases, like XBRL
GRDDL example: Dan’s homepage...

Dan Connolly
Research Scientist, MIT/CSAIL
connolly@w3.org
32 Vassar Street
Room 32-G508
Cambridge, MA, 02139 USA
+1-617-395-0241

Standards: HTML WG, TAG, GRDDL WG, RDF
Calendar, QA, DAWG/SPARQL, Semantic Web IG,
OWL, HTML 2, ESW
Research: breadcrumbs journal/weblog, cwm, N3,
tabulator, PAW, TAMI, microformats open source
Life: family, volleyball, guitar

Dan Connolly is a research scientist at the MIT Computer Science and
Artificial Intelligence Laboratory (CSAIL) in the Decentralized Information
Group (DIG) and a member of the technical staff of the World Wide Web
Consortium (W3C). His research interest is investigating the value of formal
descriptions of complex systems like the Web, especially in the
consensus-building process.

Dec 12 - 13: to San Jose, CA
for the W3C Video on the Web Workshop

Feb 10-Feb 14: to New York, NY
for Tools of Change (TOC)
> ... behind the scenes ...
>...yielding, through the GRDDL transform

```xml
<http://www.w3.org/People/Connolly/#_6768>
a <http://www.w3.org/2002/12/cal/icaltzd#Vevent>;
<http://www.w3.org/2002/12/cal/icaltzd#organizer>
  <http://www.w3.org/People/Connolly/#me>;
<http://www.w3.org/2002/12/cal/icaltzd#summary>
  "Tools of Change (TOC)";
<http://www.w3.org/2002/12/cal/icaltzd#dtstart>
  "2008-02-10"^^<http://www.w3.org/2001/XMLSchema#date>;
<http://www.w3.org/2002/12/cal/icaltzd#dtend>
  "2008-02-14"^^<http://www.w3.org/2001/XMLSchema#date>;
<http://www.w3.org/2002/12/cal/icaltzd#url>
  <http://www.w3.org/People/Connolly/#_6768>;
<http://www.w3.org/2002/12/cal/icaltzd#location>
  "New York, NY" .
```
Getting structured data to RDF: RDFa

- RDFa extends XHTML with a set of attributes to include structured data into XHTML
- Makes it easy to “bring” existing RDF vocabularies into XHTML
  - uses namespaces for an easy mix of terminologies
> RDFa example: Ivan’s homepage...

Ivan Herman

My Work at W3C

I am Semantic Web Activity Lead; that is my main work at W3C. I am member of IW3C2 (International World Wide Web Conference Committee) (the committee coordinating the yearly WWW conference series), serving as a liaison for W3C, and of SWSA (Semantic Web Science Association), the committee responsible for the International Semantic Web Conferences series.

As part of my work, I also participate in lots of outreach activities, and I regularly make presentations, tutorials, etc. You can consult my list of presentations for further details.

Contact information

Email: ivan@w3.org
(sha1sum: 5ac8032d5f6012aa1775ea2f63e1676bafd5e80b)
Postal address
> … behind the scenes …

```
<h2><a id="Meh2">My Work at W3C</a></h2>

I joined the <a rel="foaf:workInfoHomepage" href="http://www.w3.org">W3C</a> Team as Head of <a rel="foaf:pastProject" href="http://www.w3.org/Consortium/Offices" title="W3C Offices">W3C Offices</a> in January 2001 while maintaining my position at <abbr title="Centrum voor Wiskunde en Informatica" xml:lang="nl">CWI</abbr>.
```

Leiden, in the Netherlands.
> …yielding, by extracting RDF

```xml
<Overview.html#me> a foaf:Person;
   foaf:currentProject <http://www.w3.org/2001/sw>;
   foaf:workInfoHomepage
       <http://www.iw3c2.org>,
       <http://www.w3.org>;
   foaf:pastProject
       <http://www.w3.org/Consortium/Offices>;
   ...

<http://www.w3.org/Consortium/Offices> dc:title "W3C Offices".
```
> Such data can be SPARQL-ed

```sql
SELECT DISTINCT ?name ?home ?orgRole ?orgName ?orgHome
# Get RDFa from my home page:
FROM <http://www.w3.org/People/Ivan/>
# GRDDL-ing http://www.w3.org/Member/Mail:
FROM <http://www.w3.org/Member/Mail/>
WHERE {
  ?foafPerson   foaf:mbox  ?mail;
  foaf:homepage ?home.
?individual   contact:mailbox  ?mail;
  contact:fullName  ?name.
?orgUnit   ?orgRole  ?individual;
  org:name  ?orgName;
  contact:homePage  ?orgHome.
}```
SPARQL as a unifying point!
> How to “assign” RDF data to resources?

- Some examples:
  - copyright information for your photographs
  - is a Web page usable on a mobile phone and how?
  - bibliographical data for a publication
  - annotation of the data resulting from a scientific experiment
  - etc

- The issue: if I have the URI of the resource (photograph, publication, etc), how do I find the relevant RDF data?
The data might be embedded

- Some data formats allow the direct inclusion of (RDF) metadata:
  - SVG (Scalable Vector Graphics)
  - XHTML+RDFa
  - microformats+GRDDL
  - JPG files using the comment area and, eg, Adobe’s XMP technology
- That can include all the information, or link to further data
POWDER (Protocol for Web Description Resources) provides for more elaborate scenarios with POWDER:

1. Define a set of resources by constraints on the URIs; e.g.,
   - URIs must begin with `http://www.example.com/bla/`
   - The port number in the URI-s should be `XYZW`

2. Define description resources, binding each resource in the set to additional information
   - This must be attributed and may be open to authentication

3. Get such description resources, e.g., via a link statements, via HTTP, via SPARQL from a depository, …

A technology under development at W3C
A POWDER scenario: copyrights

Client

GET .../index

http://www.ex.org/index

Return Descr. Res. URI

GET Descr. Res.

GET http://.../img/imgXXX.jpg


cc:license <http://cp.../> for resources:

http://www.ex2.org/img/*

http://www.ex2.org/img/...
Everything has not been solved…

- Integration of rules and Semantic Web
  - subject of the RIF Working Group at W3C
- Updating OWL
  - in a renewed OWL Working group at W3C:
    - add new features that practice has shown to be important
    - define better “profiles” to adapt to various user communities
> Everything has not been solved...

- A number of open issues, problems
  - open technical issues
  - misconceptions, messaging problems
  - need for more applications, deployment, acceptance
  - etc
Some open technical issues

- Security, trust, provenance
  - combining cryptographic techniques with the RDF model, sign a portion of the graph, etc
  - trust models
- Ontology merging, alignment, term equivalences, versioning, development, …
- Uncertainty (fuzzy and/or probabilistic relations, reasoning, …)
- etc
> Other items: naming

- The SW infrastructure relies on unique naming of "things" via URI-s
- Lots of discussions are happening that also touch upon general Web architecture:
  - HTTP URI-s or other URN-s?
    - using non-HTTP unnecessarily complicates the general infrastructure
  - URI-s for "informational resources" and "non informational resources"
  - how to ensure that URI-s used on the SW are dereferencable
  - what inferences can be drawn upon an HTTP session?
  - etc
> Other items: naming (cont)

- A different aspect of naming: **what** is the URI for a specific resource
  - what is the unique URI for, eg, Bach’s Well-Tempered Clavier?
    - obviously important for, eg, music ontologies and data
    - who has the authority or the means to define and maintain such URI-s?
    - the traditional library community may be of a big help in this area
  - what is the URI of time-dependent entity (e.g., a specific point **within** a video)?
A major problem: messaging

- Some of the messaging on Semantic Web has gone terribly wrong over the years
- This has created lots of (unnecessary) controversies
Some of the usual misconceptions...

- The Semantic Web is simply Artificial Intelligence on the Web
- It is just an ugly application of XML
- One has to add metadata to all Web pages, convert all databases and XML data to use the Semantic Web
  - which is obviously unrealistic...
- One has to learn formal logic, knowledge representation techniques, etc, to use it (i.e., it has nothing to do with the Web)
> … and some more …

- It relies on a centrally controlled super-ontology for “meaning”
  - and this is impossible, because people will never agree on all terms
  - as opposed to a democratic, bottom-up control of terms
  - we do not need Semantic Web, Web 2.0 approaches solve all our needs...

- It is, essentially, an academic project, of no interest to industry and the real World
> Just to pick some of these…
> SW Ontologies ≠ a central, big ontology!

- The “ethos” of the Semantic Web is on sharing, i.e., sharing ontologies (small or large)
- A huge, central ontology would be unmanageable
- The practice:
  - SW applications using ontologies always mix large numbers of ontologies and vocabularies (FOAF, DC, and others)
  - the real advantage comes from this mix: that is also how new relationships may be discovered
  - applications in an area can agree on their vocabularies: and that is enough; a global consistency may not be necessary
A good example...

Courtesy of Susie Stephens, Eli Lilly
Web 2.0 and SW are no enemies…

- Web 2.0 recognized the importance of data to be processed, mashed-up, mixed
  - this is at the heart of the Semantic Web
  - SW provides a set of consistent tools and definitions to help that

- Sometimes the simplicity (eg, in tagging, microformats) pays off; sometimes more rigor is necessary
  - GRDDL is a good example for a “bridge”
  - SPARQL can be used for more complex mash-ups

- Let us forget about a turf/ego war; it is unnecessary and counterproductive
> Semantic Web ≠ academic research only

- SW has indeed a strong foundation in research results
- But a large number of applications are now available
- See the collection at W3C: http://www.w3.org/2001/sw/swoe/public/UseCases
- But, mainly: look at this conference and enjoy the variety!
Thank you for your attention!

- These slides are publicly available on:

  http://www.w3.org/2008/Talks/0518-SanJose-IH/