State of the Semantic Web

Tampere, 4 April, 2007

Ivan Herman, W3C
What will I talk about?

- The history of the Semantic Web goes back to several years now
- It is worth looking at what has been achieved, where we are, and where we might be going…
Let us look at some results first!
We have a solid specification since 2004: well defined (formal) semantics, clear RDF/XML syntax.

Lots of tools are available. Are listed on W3C’s wiki:
- RDF programming environment for 14+ languages, including C, C++, Python, Java, Javascript, Ruby, PHP,… (no Cobol or Ada yet 😐)
- 13+ Triple Stores, ie, database systems to store (sometimes huge!) datasets
- converters to and from RDF
- etc

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 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

Large datasets are accumulating. E.g.:
- IngentaConnect bibliographic metadata storage: over 200 million triplets
- RDF access to Wikipedia: more than 27 million triplets
- tracking the US Congress: data stored in RDF (around 25 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 associated RDF data: 6 million (and growing) geographical features
- RDF Book Mashup, integrating book data from Amazon, Google, and Yahoo

Some measures claim that there are over $10^7$
Semantic Web documents… (ready to be integrated…)
Ontologies: OWL

- This is also a stable specification since 2004
- Separate layers have been defined, balancing expressibility vs. implementability (OWL-Lite, OWL-DL, OWL-Full)
  - quite a controversial issue, actually…
- Looking at the tool list on W3C’s wiki again:
  - a number of programming environments (in Java, Prolog, …) include OWL reasoners
  - there are also stand-alone reasoners (downloadable or on the Web)
  - ontology editors come to the fore
- OWL-DL and OWL-Lite relies on Description Logic, i.e., can use a large body of accumulated research knowledge
Ontologies

- Large ontologies are being developed (converted from other formats or defined in OWL)
  - eClassOwl: eBusiness ontology for products and services, 75,000 classes and 5,500 properties
  - the Gene Ontology: to describe gene and gene product attributes in any organism
  - BioPAX, for biological pathway data
  - UniProt: protein sequence and annotation terminology and data
Vocabularies

- There are also a number “core vocabularies” (not necessarily OWL based)
  - **SKOS Core**: about knowledge systems
  - **Dublin Core**: about information resources, digital libraries, with extensions for rights, permissions, digital right management
  - **FOAF**: about people and their organizations
  - **DOAP**: on the descriptions of software projects
  - **MusicBrainz**: on the description of CDs, music tracks, …
  - **SIOC**: Semantically-Interlinked Online Communities
  - **vCard in RDF
  - …

- One should *never* forget: ontologies/vocabularies must be shared and reused!
A mix of vocabularies/ontologies (from life sciences)…
Ontologies, Vocabularies

- Ontology and vocabulary development is still a complex task
- The W3C SW Best Practices and Deployment Working Group has developed some documents:
  - “Best Practice Recipes for Publishing RDF Vocabularies”
  - “Defining N-ary relations”
  - “Representing Classes As Property Values”
  - “Representing "value partitions" and "value sets"”
  - “XML Schema Datatypes in RDF and OWL”

the work is continuing in the SW Deployment Working Group
Querying RDF: SPARQL

- Querying RDF graphs becomes essential
- SPARQL is almost here
  - query language based on graph patterns
  - there is also a protocol layer to use SPARQL over, eg, HTTP
  - hopefully a Recommendation end 2007
- There are a number of implementations already
- 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
  - applications may not need any direct RDF programming any more, just a SPARQL endpoint
SPARQL as the *only* interface to RDF data?


  with the query:

```
SELECT ?translator ?translationTitle ?originalTitle ?originalDate
FROM <http://.../TR_and_Translations.rdf>
WHERE {
  ?trans rdf:type trans:Translation;
    trans:translationFrom ?orig;
    trans:translator [ contact:fullName ?translator ];
    dc:language "fr";
    dc:title ?translationTitle.
  ?orig rdf:type rec:REC;
    dc:date ?originalDate;
    dc:title ?originalTitle.
}
ORDER BY ?translator ?originalDate
```

- yields...
A word of warning on SPARQL…

- It is *not* a Recommendation yet
- New issues may pop up at the last moment via reviews
  - *a query language needs very precise semantics and that is not that easy 😞*
- 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*
  - *
  - postponed to a next version…*
Of course, not everything is so rosy…

- There are a number of issues, problems
  - how to get RDF data
  - missing functionalities: rules, “light” ontologies, fuzzy reasoning, necessity to review RDF and OWL,…
  - misconceptions, messaging problems
  - need for more applications, deployment, acceptance
  - etc
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
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)
  - RSS1.0 feeds are generated by (almost) all blogging systems (a huge amount of RDF data!)
  - ...
- There are a number of projects “harvesting” and linking data to RDF (e.g., “Linking Open Data on the Semantic Web” community project)
Data may be extracted (a.k.a. “scraped”)

- Different tools, services, etc, come around every day:
  - *get RDF data associated with images, for example:*
    - service to *get RDF from flickr images* (see example)
    - service to *get RDF from XMP* (see example)
  - *XSLT scripts to retrieve microformat data from XHTML files*
  - *scripts to convert spreadsheets to RDF*
  - *etc*

- Most of these tools are still individual “hacks”, but show a general tendency
- Hopefully more tools will emerge
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*
    - script is usually XSLT but not necessarily
    - has a variant for XHTML
  - *a “GRDDL Processor” runs the script and produces RDF on–the–fly*
- A way to access existing structured data and “bring” it to RDF
  - *a possible link to microformats*
Getting structured data to RDF: RDFa

- RDFa (formerly RDF/A) extends XHTML with a set of attributes to include structured data into XHTML
  - an XHTML1 module is being defined
- Makes it easy to “bring” existing RDF vocabularies into XHTML
- Uses namespaces for an easy mix of terminologies
- It can be used with GRDDL but RDFa aware systems can manage it directly, too
  - no need to implement a separate transformation per vocabulary
I graduated as mathematician at the Eötvös Loránd University of Budapest, Hungary, in 1979. After a brief scholarship at the Université Paris VI, I joined the Hungarian research Institute in computer science (SZTAKI) where I worked for 8 years. I left Hungary in 1988 and, after a few years in industry, I joined the Centre for Mathematics and Computer Sciences (CWI) in Amsterdam where I have a tenure position since 1988. I received a PhD degree in Computer Science in 1990 at the University of Leiden, in the Netherlands. I joined the W3C team as Head of Offices in January 2001 while maintaining my position at CWI. I served as Head of Offices until June 2006, when I was asked to take the Semantic Web Activity Lead position, which is now my principal work at W3C.

Before joining W3C I worked in quite different areas (distributed and database programming, language design, system programming), but I spend most of my research years in computer graphics and visualization. I also participated in various graphics related ISO standardization activities and software developments. My homepage at CWI contains a list of my publications and details of the various projects I participated in the past. (e.g. ASTER, based on my publications, my Erdős number is 4.)

I was member of the Executive Committee of the Eurographics Association for 15 years, and I was vice-chair of the Association between 2000 and 2002. I was the co-chair of the 6th World Wide Web Conference in Amsterdam, May 2000. I am also member of W3C, the committee responsible for the World Wide Web Conference series. Finally, I am on the management board of Accessibility.nl, a non-profit Dutch organization dedicated to the accessibility of Web sites.

I have also developed some software (in Python) that might be of interest, particularly a SPARQL (RDF Query Language) API implementation on the top of the RDFLib package. This package has recently been added to the latest release of RDFLib; a SPARQL language parser on the top of it is being developed by Chinmaya Ogale.

See my separate page if you are curious about more private things, including some of my photos that put on the Web...

**Upcoming Trips**

(See these also in .ics and .rdf formats)

- November 28-December 2, Tokyo, Japan: W3C AC, W3C AC Meeting and W3C Team day (fly to Tokyo on the 28th of November with an arrival on the 29th, return to Amsterdam on the 2nd of December)
- January 14-23, Cambridge, MA, USA, visit to MIT
- January 31, February 3, Girona, Spain, European W3C Symposium on eGovernment
marked up with GRDDL headers...
...and hCard microformat tags...
...yielding; ...
(see the full file if interested…)
...marked up with RDFa tags...

---

I graduated as mathematician at the [School of Mathematics and Computer Science](http://www.w3.org/People/Ivan/) in 1979. After a brief scholarship at the Université Paris (1979/80) I joined the Hungarian research science (a [degree](http://www.istok.hu)) in 1981. I left Hungary in 1986 and, after a few years in industry, I joined the [Center for Computer Science and Mathematics](http://www.ewi.tudelft.nl) in 1989. I have been there as a tenure position since 1988. I received a PhD degree in Computer Science in 1998.

I joined the [W3C](http://www.w3.org) team in 2001 as Head of [Consortium/Offices](http://www.w3.org) in January 2001 while maintaining my position at [Center for Computer Science and Mathematics](http://www.ewi.tudelft.nl) in Amsterdam. I served as Head of Offices until June 2006, when I was asked to take the position of [Semantic Web](http://www.w3.org/2001/sw/) Activity Lead position, which is now...
...yielding; ...

(see the full file if interested...)
SPARQL-ing such data

- [http://www.sparql.org/sparql?query=...](http://www.sparql.org/sparql?query=...) with the query:

```sparql
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.
}
```

- yields...
Linking to SQL

- A huge amount of data in Relational Databases
- Although tools exist, it is not feasible to convert that data into RDF
- Instead: SQL ⇔ RDF “bridges” are being developed:
  - a query to RDF data is transformed into SQL on-the-fly
  - the modalities are governed by small, local ontologies or rules
- An active area of development, on the radar screen of W3C!
SPARQL as a unifying point?
Missing features, functionalities…

- Everybody has a favorite item, ie, the list tends to infinite…
- W3C is a *standardization* body, and has to look at where a consensus can be found
OWL-DL and OWL-Lite are based on Description Logic; there are things that DL cannot express
• a well known examples is Horn rules (eg, the “uncle” relationship):
  ○ \( (P_1 \land P_2 \land \ldots) \rightarrow C \)
  ○ e.g.: for any «X», «Y» and «Z»: “if «Y» is a parent of «X», and «Z» is a brother of «Y» then «Z» is the uncle of «X»”
• there are a number of attempts to combined these: RuleML, SWRL, cwm, …

There is also an increasing number of rule-based system that want to interchange rules
• a new type of data (potentially) on the Web to be interchanged…
Some typical use cases

- Negotiate eBusiness contracts across platforms: supply vendor-neutral representation of your business rules so that others may find you
- Describe privacy requirements and policies, and let clients “merge” those (e.g., when paying with a credit card)
- Medical decision support, combining rules on diagnoses, drug prescription conditions, etc,
- Extend RDFS (or OWL) with rule-based statements (e.g., the uncle example)
In an ideal World…
In the real World...

- Rule based systems can be very different
  - *different rule semantics (based on various type of model theories, on proof systems, etc)*
  - *production rule systems, with procedural references, state transitions, etc*
RIF “core”: only partial interchange

- Specification of the “core” is the first step
- It also forms a logic language to be used, e.g., with OWL, RDF, XML data, …
Possible variants: F-logic, production rules, fuzzy logic systems, …; none of these have been finalized yet
Role of variants
“Light” ontologies

- For a number of applications RDFS is not enough, but even OWL Lite is too much
- There may be a need for a “light” version of OWL, just a few extra possibilities v.a.v. RDFS
- There are a number of proposals, papers, prototypes around: RDFS++, OWL Feather, pD*,…
  - pD*, for example, has property characterization (symmetric, transitive, inverse), class and property equivalence, and property restrictions with some or all values
- This might consolidate in the coming years
Revisions of RDF and OWL?

- Such specifications have their own life
- Missing features come up, errors show up
- There may be a next version at some point
  - *but: it is always a difficult decision; introducing a new version creates uncertainty in the developers’ community* 😞
Revision of the RDF model?

- Some restrictions in RDF may be unnecessary (bNodes as predicates, literals as subject,...)
- Issue of “named graph”: possibility to give a URI to a set of triplets and make statements on those
- Syntax issues in RDF/XML (eg, QNames in properties)
- Alternative XML serializations?
- Add a time tag to statements? A probability value? A measure of “fuzzyness”?
- Internationalization issues with literals (how do I set “bidi” writing?)

These are just ideas floating around…
There is a (non-W3C) group working on this

**Small additions** to the current OWL:

- "qualified cardinality restrictions" (i.e., "class instance must have two black cats")
- disjoint properties
- reflexive, irreflexive properties
- property composition
- own datatype constructs instead of complex XML Schema datatypes
- "light" ontologies (called “tracable fragments”)
- some syntactic sugar (eg, disjoint union)
- …

At this moment not yet decided how, if, and when this would become a W3C document
Other items…

- Fuzzy logic
  - look at alternatives of Description Logic based on fuzzy logic
  - alternatively, extend RDF(S) with fuzzy notions
- Probabilistic statements
  - have an OWL class membership with a specific probability
  - combine reasoners with Bayesian networks
- Security, trust, provenance
  - combining cryptographic techniques with the RDF model, sign a portion of the graph, etc
- Ontology merging, alignment, term equivalences, versioning, development, …
- etc

(Need a new PhD topic? 😊)
A major problem: messaging

- Some of the messaging on Semantic Web has gone terribly wrong. See these statements:
  - "the Semantic Web is a reincarnation of Artificial Intelligence on the Web"
  - "it relies on giant, centrally controlled ontologies for "meaning" (as opposed to a democratic, bottom-up control of terms)"
  - "one has to add metadata to all Web pages, convert all relational databases, and XML data to use the Semantic Web"
  - "it is just an ugly application of XML"
  - "one has to learn formal logic, knowledge representation techniques, description logic, etc, to use it"
  - "it is, essentially, an academic project, of no interest for industry"
  - ...

- Some simple messages should come to the fore!
RDF ≠ RDF/XML!

- RDF is a model, and RDF/XML is only one possible serialization thereof
  - lots of people prefer, for example, Turtle
  - a good percentage of the tools have Turtle parsers, too!
- The model is, after all, simple: interchange format for Web resources. That is it 😊!
RDF ≠ RDF/XML! (cont.)

- RDF/XML is indeed a very complex serialization format
- Certainly not the nicest possible XML application
  - *good to know that it was created when XML was not yet final…*
- Again: it is only syntactic sugar!
- One has to emphasize: RDF is *not* an XML application!
RDF is not *that* complex…

- Of course, the formal semantics of RDF *is* complex
- But the average user should not care, it is all “under the hood”
  - *how many users of SQL have ever read its formal semantics?*
  - *it is not much simpler than RDF…*
- *People should “think” in terms of graphs,* the rest is syntactic sugar!
Formal ontologies (like OWL) are important, but use them *only when necessary*

- you can be a perfectly decent citizen of the Semantic Web if you do not use Ontologies, not even RDFS…
- remember the “light ontologies” issue?
Web 2.0 and Semantic Web are not antagonistic…

- Web 2.0 recognized the importance of *data* to be processed, mashed-up, mixed
  - *this is at the center of the Semantic Web*
  - *SW provides a set of consistent tools and definitions to achieve that*
- Sometimes the simplicity of Web 2.0 (eg, in tagging, microformats) pays off; sometimes more rigor is necessary in which case Semantic Web technologies come to the fore
  - *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
The “ethos” of the Semantic Web is on sharing, ie, sharing ontologies (small or large)
A huge, central ontology would be unmanageable
OWL includes statements for versioning, for equivalence and disjointness of terms
- a revision of those may be necessary, but the goal is clear
The practice:
- SW applications using ontologies always mix large number of ontologies and vocabularies (FOAF, DC, and others)
- the real advantage comes from this mix: that is also how new relationships may be discovered
Remember?
Semantic Web ≠ an academic research only!

- SW has indeed a strong foundation in research results
- But remember:
  - (1) the Web was born at CERN…
  - (2) …was first picked up by high energy physicists…
  - (3) …then by academia at large…
  - (4) …then by small businesses and start-ups…
  - (5) “big business” came only later!
- network effect kicked in early…
- Semantic Web is now at #4, and moving to #5!
May start with small communities

- The needs of a deployment application area:
  - have serious problem or opportunity
  - have the intellectual interest to pick up new things
  - have motivation to fix the problem
  - its data connects to other application areas
  - have an influence as a showcase for others

- The high energy physics community played this role for the Web in the 90’s
<table>
<thead>
<tr>
<th>Problem to solve?</th>
<th>Library metadata</th>
<th>Defense</th>
<th>Life sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-domain integration</td>
<td>yes, serious data integration needs</td>
<td>yes, connections among genetics, proteomics, clinical trials, regulatory,…</td>
<td></td>
</tr>
<tr>
<td>Willingness to adopt?</td>
<td>yes: OCLC push and Dublin Core Initiative(*)</td>
<td>yes: funded early DAML (OWL) work</td>
<td>yes: intellectual level high, much modeling done already.</td>
</tr>
<tr>
<td>Motivation</td>
<td>light</td>
<td>strong</td>
<td>very strong</td>
</tr>
<tr>
<td>Links to</td>
<td>other library data</td>
<td>phone calls records, etc</td>
<td>chemistry, regulatory, medical, etc</td>
</tr>
</tbody>
</table>

(*) note that the Dublin Core Initiative’s work go way beyond digital libraries these days
Some RDF deployment areas (cont)

- These are just examples
- Others are coming to the fore: eGovernment, energy sector (oil industry), financial services,…
- Health care and life science sector is now very active
  - also at W3C, in the form of an Interest Group
Major companies offer (or will offer) Semantic Web tools or systems using Semantic Web: Adobe, Oracle, IBM, HP, Software AG, webMethods, Northrop Gruman, Altova,…

Some of the names of active participants in W3C SW related groups: ILOG, HP, Agfa, SRI International, Fair Isaac Corp., Oracle, Boeing, IBM, Chevron, Siemens, Nokia, Merck, Pfizer, AstraZeneca, Sun, Citigroup,…

“Corporate Semantic Web” listed as major technology by Gartner in 2006

The Semantic Technology Conference series also attract lots of participants
- speakers in 2006: from IBM, Cisco, BellSouth, GE, Walt Disney, Nokia, Oracle, …
- not all referring to Semantic Web (eg, RDF, OWL,…) but semantics in general
- but they might come around!
Data integration comes to the fore as one of the SW Application areas. Very important for large application areas (life sciences, energy sector, eGovernment, financial institutions), as well as everyday applications (e.g., reconciliation of calendar data).

Life sciences example:

- data in different labs…
- data aimed at scientists, managers, clinical trial participants…
- large scale public ontologies (genes, proteins, antibodies, …)
- different formats (databases, spreadsheets, XML data, XHTML pages)
- etc
General approach

1. Map the various data onto RDF
   - assign URI-s to your data
   - “mapping” may mean on-the-fly SPARQL to SQL conversion, “scraping”, etc
2. Merge the resulting RDF graphs (with a possible help of ontologies, rules, etc, to combine the terms)
3. Start making queries on the whole!
   - Remember the role of SPARQL?
Example: antibodies demo

- Scenario: find the known antibodies for a protein in a specific species
- Combine ("scrape"…) three different data sources
- Use SPARQL as an integration tool (see also demo online)
There has been lots of R&D

- Pfizer, MITRE Corp., Elsevier, EU Projects like Sculpteur and Artiste, national projects like MuseoSuomi, UN FAO’s MeteoBroker, DartGrid, …
- Developments are under way at various places in the area
- A general question: can I access your (RDF) data directly?
Example: ontology controlled annotation
Portals

- Vodafone's Live Mobile Portal
  - search application (e.g. ringtone, game, picture) using RDF
    - page views per download decreased 50%
    - ringtone up 20% in 2 months
- A number of other portal examples: Sun’s White Paper Collections and System Handbook collections; Nokia’s S60 support portal; Harper’s Online magazine linking items via an internal ontology; Oracle’s virtual press room; Opera’s community site, Yahoo! Food, FAO’s Food, Nutrition and Agriculture Journal portal,…
- A general question again: can I access your (RDF) data directly?
Improved Search via Ontology: GoPubMed

- Improved search on top of pubmed.org
  - search results are ranked using the specialized ontologies
  - extra search terms are generated and terms are highlighted
- Importance of domain specific ontologies for search improvement
- Center of information for the treatment of premature babies
- Provides an OWL service as a Web Service
  - combines disparate vocabularies like medical, insurance, etc
  - users can add new entries to ontologies
  - complex questions can be asked through the service
Other Application Areas Come to the Fore

- Knowledge management
- Business intelligence
- Linking virtual communities
- Management of multimedia data (e.g., video and image depositories)
- Content adaptation and labeling (e.g., for mobile usage)
- etc
Thank you for your attention!

These slides are publicly available on:

http://www.w3.org/2007/Talks/0403-Tampere-IH/

in XHTML and PDF formats; the XHTML version has active links that you can follow.