Semantic Web @ W3C: Activities, Recommendations and State of Adoption

Athens, GA, USA, 2006-11-09

Ivan Herman, W3C
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 datasets
- 16+ general development tools (specialized editors, application builders,…)
- etc
Note the large number of large corporations among the tool developers: Adobe, IBM, Software AG, Oracle, HP, Northrop Grumman, … …but the small companies and independent developers also play a major role! 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, or books around

- the wiki page on books lists 20+ (English) textbooks; 19+ proceedings for 2005 & 2006 alone…
- again, some of them good, some of them bad, just as with any other areas…

Active developers’ communities
Large datasets are accumulating

- IngentaConnect bibliographic metadata storage: over 200 million triplets
- UniProt Protein Database: 262 million triplets
- RDF version of Wikipedia: more than 47 million triplets
- RDFS/OWL Representation of Wordnet: 150MB of RDF/XML
  - a good example on how to organize large amount of RDF data
- “Département/canton/commune” structure of France
  published by the French Statistical Institute
- This conference has reported on more!
Ontologies: OWL

- This is also a stable specification since 2004
- Looking at the tool list on W3C’s wiki again:
  - a number programming environments (in Java, Prolog, …) include OWL reasoners (OWL-Lite or OWL-DL)
  - there are also stand-alone reasoners (downloadable or on the Web)
  - ontology editors come to the fore
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
- **UniProt**: protein sequence and annotation terminology and data
- again, look around at the conference… 😊
There are also a number “core vocabularies” (not necessarily OWL based)

- **SKOS Core**: about knowledge systems
  - still to be finalized at W3C
- **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
- …

More are needed; active community participation is important!
Ontologies, vocabularies

- Ontology and vocabulary development is still a complex task, we all know that…
- 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 (new) SW Deployment Working Group, watch this space (and contribute if you can…)!
SPARQL has stirred lots of interest already (though not yet finished)
It is an essential piece of the Semantic Web puzzle
There already a 20+ implementations already(!), either stand alone or part of a system
A number of SPARQL “endpoints” make it possible to experiment with it already!
Some words of warning on SPARQL…

- It is *not* a Recommendation yet
- New issues came up (at the CR phase)
  - *need for a very precise semantics and that is not easy*
  - *these issues forced the Working Group to go “back” to a Working Draft from Candidate Recommendation*
  - *but they should sort his out in 2007*
- Some features are *not* included
  - *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, “to-do”-s
  - 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
  - mapping relational databases to RDF
  - get your data out there!
  - 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…

“SW-aware” tools are around, though more would be good:

- *Photoshop CS* stores metadata in RDF in, say, jpg files *(using XMP)*
- *RSS 1.0* feeds are generated by *(almost) all blogging systems*
- *various works on semantic wikis*
- *etc…*
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
GRDDL is a more systematic way of accessing data and turn it into RDF:

- defines XML attributes to bind a suitable (usually XSLT) script to transform (part of) the data into RDF (also has a variant for XHTML)
  - a “GRDDL Processor” would then run the script and produce RDF on-the-fly
- is a possible link to microformats
- developed in a dedicated W3C Working Group, should be finished in 2007
  - aims at setting a record as the W3C Working Group with the shortest life-span 😊
RDFa (formerly RDF/A) extends XHTML by:

- extending the `link` and `meta` to include child elements
- add metadata to any elements (a bit like the `class` in microformats, but via dedicated properties)
- easier mix of terminologies via namespaces
- a general framework, ie, no extra scripting is necessary, only one general RDFa tool
- aims at a complete representation ("serialization") of RDF embedded in XHTML
  - not bound to XHTML2 (any more); goal is to have an XHTML1 module
- developed in the SW Deployment Working Group
GRDDL & RDFa example

See my public page (marked up both with RDFa and microformats)
SPARQL-ing such data

- http://www.sparql.org/sparql?query=...

  with the query:

```sparql
SELECT DISTINCT ?name ?home ?orgRole ?orgName ?orgHome
# RDFa-ing my home page:
FROM <http://torrez.us/services/rdfa/http://www.w3.org/People/Ivan/>
# GRDDL-ing http://www.w3.org/Member/Mail:
FROM <...online_xslt/xslt?...rdf-in-xhtml-processor&xmlfile=...%2FMember%2FMail%2F...>
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...
SPARQL as a unifying point?
Missing pieces of the SW puzzle…

- 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):
  - $\diamond (P_1 \land P_2 \land \ldots) \rightarrow C$
  - 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)
One would define a full Rule Interchange Format, that *all* rule systems could translate to and from.

That format could also be used to implement rules for (e.g.) RDF data and OWL ontologies.
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
In the real World…

- The Rules Interchange Format WG defines a “core” that is common to all…
  - a “minimal” but useful level that covers a large percentage of use cases (probably “full Horn”) with clear semantics
  - this core will be usable with RDF data, OWL ontologies…
    - details are still to be worked out, a first document might be out before the end of the year…
- … then categorizes the various systems and shows how the core should be extended to cover the various functionalities
- I.e., we will have a set of “variants” to cover various needs
- Hopefully the first results in 2007 (publication of the “core”)
Issues floating around for RDF(S), OWL...

- Extensions/changes of the RDF(S) core
- OWL 1.1
- Alternative serializations (eg, giving Turtle a final form)
- API standardization (eg, having a stable RDF API in Javascript for Web Applications)
- Access control, signatures, provenance

All on the “radar screen” of W3C…
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!
A new “Semantic Web Education and Outreach” Interest Group has just been started

Goals are:

- collect and document real SW use cases, make the available to the World
- improve the general messaging via guidelines to tutorial, teaching materials, FAQ-s

First face-to-face meeting will be next week…
Semantic Web adoption

- 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!
The “corporate” landscape is moving

- Remember the companies’ presence in tools?
- Some of the 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!
Adoption may start with special 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
Some RDF deployment areas (cont)

- Some deployment areas are already very active: health care and life sciences, digital libraries, defense
  - *also at W3C, in the form of an Interest Group for HCLS*
    
    *(participants include Oracle, IBM, Merck, Pfizer, AstraZeneca, …)*
  
  - *look at their workshop proceedings of this week…*

- Others are coming to the fore: eGovernment, energy sector (eg, oil industry), financial services, legal profession, …
Some typical adoption problems

- Change of concepts, habits, traditions
  - adoption of URI-s for one’s data
  - accept (and appreciate!) sharing and reusing (eg, ontologies)
  - privacy, access control, provencance issues
    - applications may be restricted to Intranets today; this is not unlike the early Web applications…
- Access to legacy data (eg, in databases), how to do that, etc
- Efficiency of the tools (eg, triple stores), need for other class of tools (visual representation of data, connection to relational databases, etc)
- Understanding the technology and its advantages (back to the messaging problem…)
Applications are not always very complex...

- Eg: simple semantic annotations of patients’ data greatly enhances communications among doctors
- What is needed: some simple ontologies, an RDFa/microformat type editing environment
- Simple but powerful!
A major adoption area: data integration

- 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*
There has been lots of R&D in the area

- Boeing, MITRE Corp., Elsevier, EU Projects like Sculpteur and Artiste, national projects like MuseoSuomi, DartGrid, …
- Developments are under way at various places in the area
- Issue: is your RDF data visible?
Not only data integration: eg, 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…

- Issue (again): is your RDF data visible?
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
Conclusions

- We have gone a long way (thanks to all of you!)
- We live exciting times in terms of SW Adoption
- Of course lots of work is still to be done (life would be boring otherwise…)
Thank you for your attention!

These slides are publicly available on:

http://www.w3.org/2006/Talks/1109-Athens-IH/

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