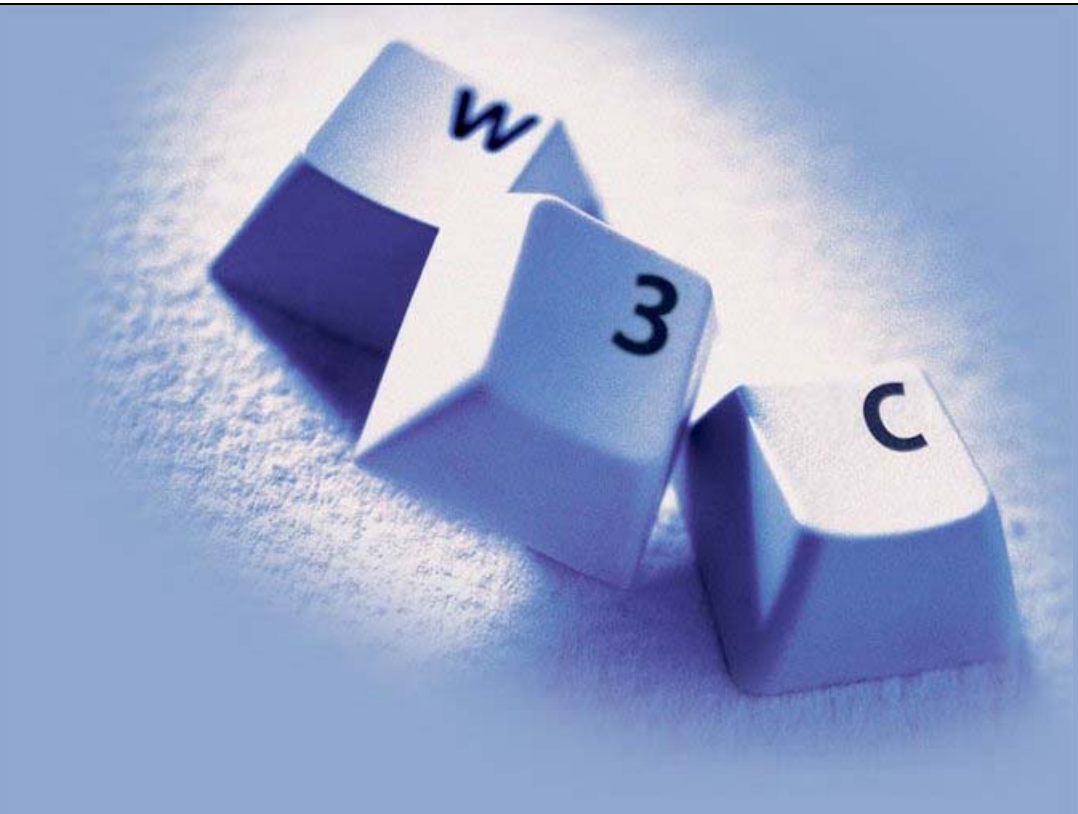


State of the Semantic Web

Stavanger, Norway, 2007-04-24

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!

The basics: RDF(S)

- 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*

The basics: RDF(S) (cont.)

- There are lots of tutorials, overviews, and books around
- Active developers' communities
- Large datasets are accumulating
- Some mesaures 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)
- Looking at the tool list on W3C's wiki again:
  - *a number 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, ie, 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, thesauri, glossaries*
  - *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*
  - *Music Ontology: 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!

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 can also be used to construct graphs!

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!)*
  - *…*

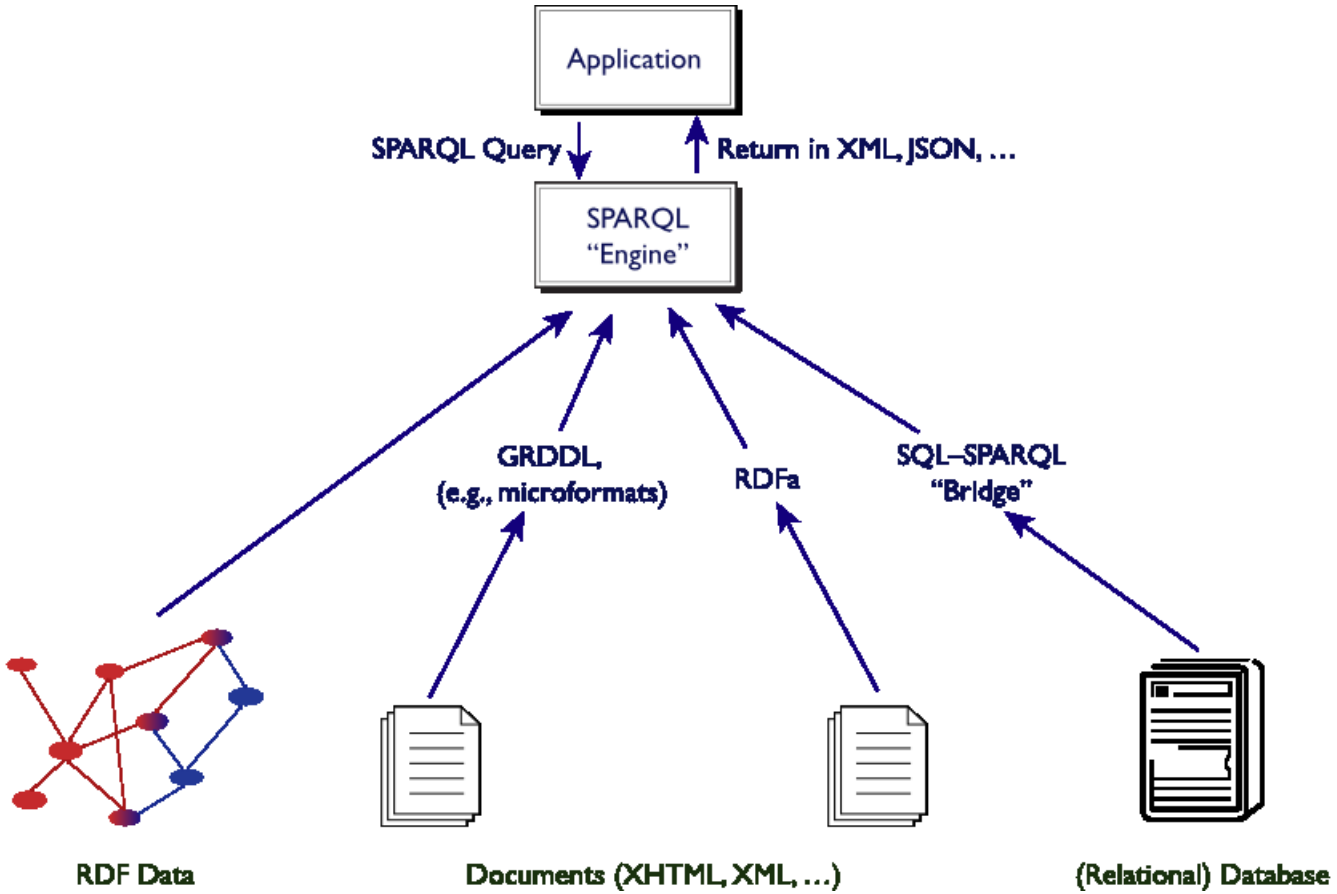
Data may be extracted (a.k.a. “scraped”)

- 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
- W3C’s new GRDDL technology is a formal way of doing this for XML/XHTML

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!
- There are a number of projects “harvesting” and linking data to RDF (e.g., “Linking Open Data on the Semantic Web” community project)

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

Rules

- OWL-DL and OWL-Lite are based on Description Logic; there are things that DL cannot express
  - *a well known examples is Horn rules:*
    - $(P_1 \wedge P_2 \wedge ...) \rightarrow C$
  - *there are a number of attempts to combine 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…*

Rules (cont)

- 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)*
- The "Rule Interchange Format" Working Group is working on this problem as we speak…

“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: EL++, RDFS++, OWL Feather, pD*, DL Lite,…
- This might consolidate in the coming years

New versions 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*

Other items…

- Revision of the RDF model (eg, no restriction on predicates and literals)
- Revision of OWL (you may have heard of OWL1.1…)
- Fuzzy logic
  - *look at alternatives of Description Logic based on fuzzy logic*
  - *alternatively, extend RDF(S) with fuzzy notions*
- Probabilistic statements
- 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

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 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!

Semantic Web ≠ Ontologies on the Web!

- 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?*

SW Ontologies ≠ some *central*, big ontology!

- 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*

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!

Some Semantic Web deployment communities

- The technology is picked up by specialized communities
  - *just like the high energy physics community did for the original Web…*
- Some examples: digital libraries, defence, eGovernment, energy sector, financial services, health care, life sciences…
- Health care and life science sector is now very active
  - *also at W3C, in the form of an Interest Group*

The “corporate” landscape is moving

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

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 (eg, 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*

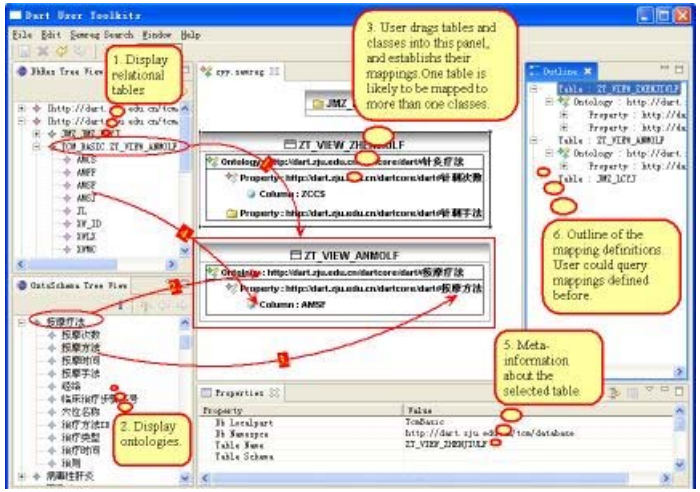
General approach

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!

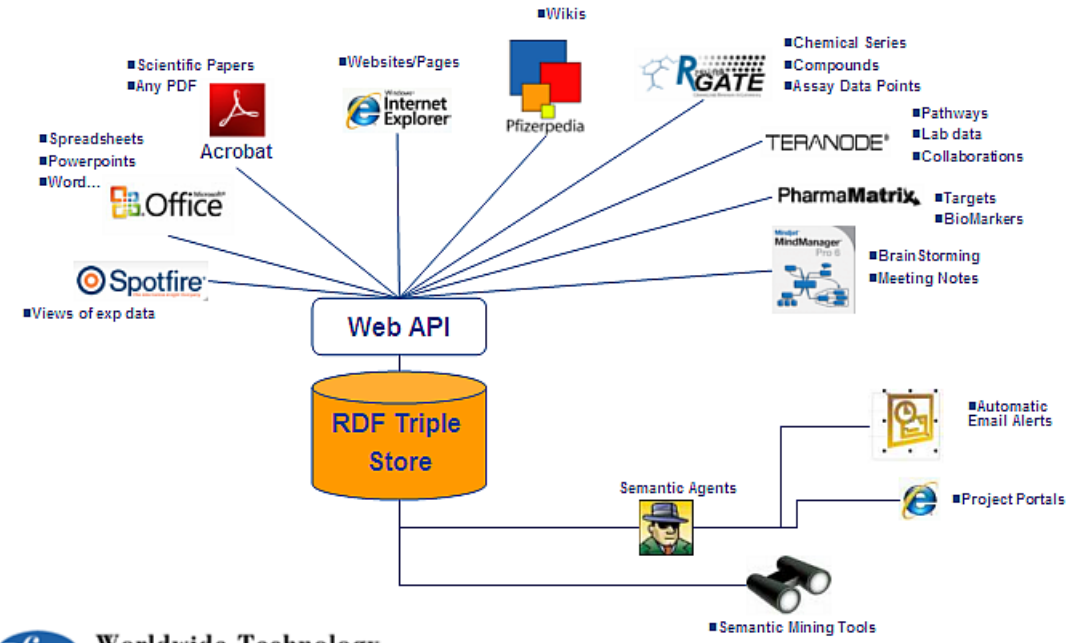
Remember the role of SPARQL?

A number of projects in the area

- Pfizer, NASA, Eli Lilly, MITRE Corp., Elsevier, EU Projects like Sculpteur and Artiste, UN FAO's MeteoBroker, DartGrid, …
- Developments are under way at various places in the area



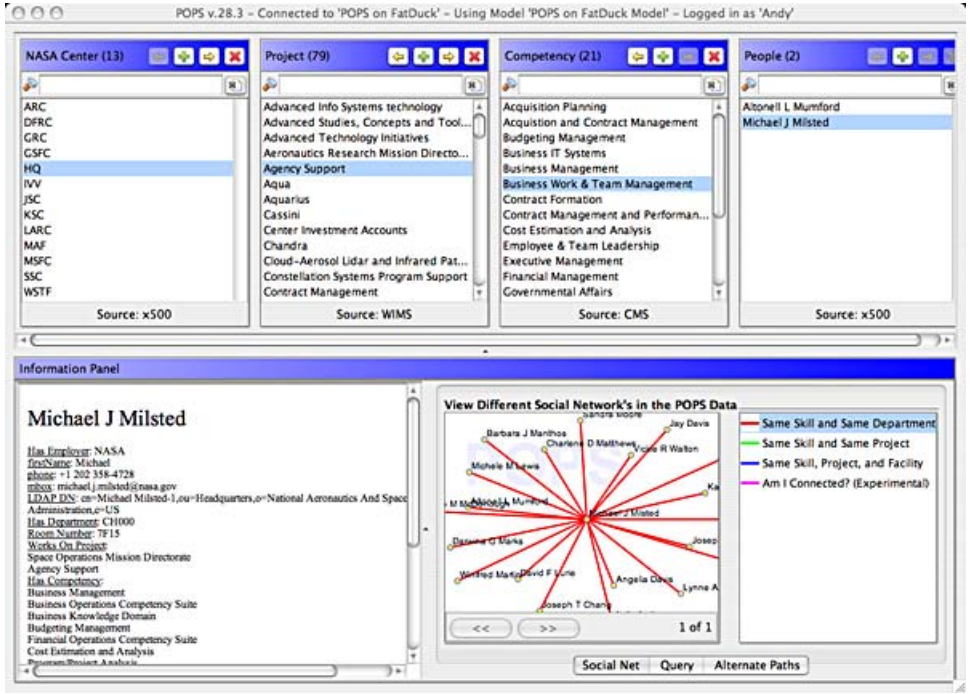
Example: ontology controlled annotation



Worldwide Technology
R&D Informatics

Example: find the right experts at NASA

Expertise locator for nearly 20,000 NASA civil servants using RDF integration techniques over 6 or 7 geographically distributed databases, data sources, and web services…



(Courtesy of Clark & Parsia, LLC)

Portals

- Vodafone's Live Mobile Portal
  - *search application (e.g. ringtone, game) using RDF*
    - page views per download decreased 50%
    - ringtone up 20% in 2 months
- 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,…

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/0424-Stavanger-IH/
```

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