State of the Semantic Web
(at least as seen from W3C…)

3rd Chinese Semantic Web Symposium
Nanjing, China
August 30, 2009

Ivan Herman, W3C
ivan@w3.org
In the past 10 years...

- The Semantic Web started as a vision, then lead to research, then to research and development...
- Today a healthy industry is growing around it
  - big corporations offering tools (IBM, Oracle, HP, Adobe, Profium,…)
  - even more are using it in some way or other (Novartis, Sun, Eli Lily, EDF, Yahoo!, Google, FAO, Bankinter, …)
Lots of tools

• 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!

• We still need more “middleware” tools to properly combine what is already available…

• *Anybody can start developing RDF-based applications today*
Some analysts’ quotes

By 2012, 80% of public Web sites will use some level of semantic hypertext to create SW documents [...] 15% of public Web sites will use more extensive Semantic Web-based ontologies to create semantic databases

Gartner report, May 2007

PWC believes a Web of data will develop that fully augments the document Web of today. You’ll be able to find and take pieces of data sets from different places, aggregate them without warehousing, and analyze them in a more straightforward, powerful way than you can now.

PriceWaterhouseCoopers Spring ‘09 Technology Forecast, May 2009
Standards play an essential role

• The Semantic Web can work because it relies on standards that the community adheres to
  • the essence of the Semantic Web is that one can integrate data from different places using the \textit{same}, i.e., \textit{standard} representation
• W3C has always been one of the major players in this technology
• In what follows, I will give an overview of what is happening at W3C these days
• Lots of things are happening, actually…
  • technology work
    • POWDER, OWL 2, RIF, SPARQL, RDFa, SKOS, media (primarily video) annotations, Linked Open Data community
  • thematic Interest Groups (acting or planned)
    • Health Care and Life Sciences, XBRL, eGovernment
  • incubator groups
    • Semantic Sensor Webs, Social Spaces
What I will do...

- Highlight some technologies that have been under development
- ... and may not have received all the attention they deserve (yet!)
- Unfortunately, there is no time to go into the details of what the thematic groups do 😞
A thread that binds lots of the issues together...
Goal:

- "expose" open datasets via RDF
- set RDF links among the data items from different datasets
  - a typical example is to set an owl:sameAs between two items in different datasets that refer to the same "thing"
- possibly set up query endpoints (usually SPARQL)
- Altogether billions of triples, millions of links…
- The "seed" for a general Web of Data
The LOD “cloud”, September 2008

As of September 2008
The LOD “cloud”, July 2009
A reminder what it means: DBPedia example

@prefix dbpedia <http://dbpedia.org/resource/>.
@prefix dbterm <http://dbpedia.org/property/>.

dbpedia:Amsterdam

dbterm:officialName "Amsterdam" ;
dbterm:longd "4" ;
dbterm:longm "53" ;
dbterm:longs "32" ;
...
dbterm:leaderTitle "Mayor" ;
dbterm:leaderName dbpedia:Job_Cohen ;
...
dbterm:areaTotalKm "219" ;
...
dbpedia:ABN_AMRO

dbterm:location dbpedia:Amsterdam ;
...
Automatic links among open datasets

Processors can switch automatically from one to the other…
Applications using the cloud emerge

- Bookmarking systems, exploration of social graphs, financial reporting
- LOD nodes (eg, DBPedia) provide a set of referenceable URI-s for many things
- Worth looking at the proceedings of the latest workshop, for example
  - April 2009, at WWW2009
  - http://events.linkeddata.org/ldow2009
Faviki: social bookmarking, semantic tagging

- Social bookmarking system (a bit like del.icio.us) but with a controlled set of tags
  - tags are terms extracted from Wikipedia/DBpedia
  - tags are categorized using the relationships stored in DBpedia
  - tags can be multilingual, DBpedia providing the linguistic bridge
- The tagging process itself is done via a user interface hiding the complexities

Courtesy of Vuk Milicic, Faviki, (SWEO Case Study)
That was how one critic was compelled to describe Sviatoslav Richter when he heard Clara Schumann speak of the venerable Franz Liszt. The musical world mourned when a great artist, indisputably one of the greatest pianists of the 20th century, Richter, with his exquisite mastery of the keyboard, ever since the 1960s when he first performed in the bloc countries where he had been renowned for years. He consistently played the music of his generation, and his performances have been responsible for the robust Soviet-American cultural exchange that began in the 1970s.

He was born in Zhitomir in the Ukraine to a family of German ancestry. His father was a respected pianist and piano teacher and his mother an amateur musician who was one of the early admirers of Debussy and Scriabin. He had his first music lessons with his father, becoming a master of the keyboard at the age of 8. The family later moved to Odessa where the young Sviatoslav enrolled at the Odessa Conservatory. In his teens, he was attracted to a career in conducting and at the astoundingly young age of 15 became a conductor for the Odessa Opera and the Ballet Theater, a post he held for four years. He gave his first piano recital at age 19 also in Odessa. Cognizant of his extraordinary talent, his superiors convinced him to study in Moscow with one of Russia's foremost piano teachers, Heinrich Neuhaus. He did so at the age of 22 and soon after completed his studies with the great piano teacher, who later wrote of his star student: "I must say in all honesty that there was nothing more I could teach Richter."
Using the LOD to build Web site: BBC
Using the LOD to build Web site: BBC
Using the LOD to build Web site: BBC
Application areas add their own “sub-cloud”

- Some “bio” related blobs were added by W3C’s HCLS IG
- The eGovernment IG plans to do the same
Challenge: get the data out there!
How to access a database

- Many of the LOD blobs come from relational databases
- Issue: how to “map” a relational database content to RDF
  - different tools exist (Virtuoso’s RDF view, D2RQ, Triplify, R2O, Dartgrid toolkit, Asio, RDBToOnto)
  - the W3C RDB2RDF Incubator Group published a survey:
How to access a database (cont.)

- A new RDB2RDF Working Group is planned
- Goal:
  - “standardize a language for mapping relational data and relational database schemas into RDF and OWL”
  - how to assign public identifiers to database entries
  - group should start in September, watch the news and join!
Data in other formats

- But many data are in XML, HTML and not in databases
- Fortunately, GRDDL and RDFa are already around to easily produce (RDF) data
  - the usual tools begin to adopt GRDDL and RDFa to retrieve RDF automatically
- These data can be added to the cloud easily
Just a few words about GRDDL and RDFa

- GRDDL helps application to convert structural data into RDF
  - eg, microformats to RDF
- With RDFa one can add “meta” data (essentially RDF) to XHTML or to XML dialects
  - automatic tools can then extract RDF
Example: Yahoo’s SearchMonkey

- Search results pages may be customized via small applications
- RDFa/microformats data in pages are reused

Courtesy of Peter Mika, Yahoo! Research, (SWEO Case Study)
Example: Google’s rich sniplet

• Embedded metadata (in microformat or RDFa) is used to improve search result page
  • at the moment only a few vocabularies are recognized, but that will evolve over the years

Drooling Dog Bar B.Q - Colfax, CA
★★★★★ 15 reviews - Price range: $$

Drooling Dog has some really good BBQ. I had the pulled pork sandwich, .... Drooling Dog BBQ is a great place to stop at on your way up the hill to Tahoe ...

www.yelp.com/biz/drooling-dog-bar-b-q-colfax - 75k - Cached - Similar pages
Publication of data: SlideShare

What is the Semantic Web? (In 15 minutes...)

ISOC Nieuwjaarsreceptie 2009

2009-01-15, Amsterdam, The Netherlands

Ivan Herman, W3C
Publication of data: SlideShare

<pre><code>&lt;prefix dc: &lt;http://purl.org/dc/terms/&gt; .
 prefix hx: &lt;http://purl.org/HIN/hinclude&gt; .
 prefix media: &lt;http://search.yahoo.com/searchmonkey/media/&gt; .
 prefix rdf: &lt;http://www.w3.org/1999/02/22-rdf-syntax-ns#&gt; .
 prefix rdfs: &lt;http://www.w3.org/2000/01/rdf-schema#&gt; .
 prefix xhv: &lt;http://www.w3.org/1999/xhtml/vocab#&gt; .
 prefix xsd: &lt;http://www.w3.org/2001/XMLSchema#&gt; .

&lt;http://www.slideshare.net/ivan_herman/what-is-the-semantic-web-in-15-minutes-presentation&gt;
 dc:creator "Ivan Herman"@en ;
 dc:description "Very short introduction to what the Semantic Web is, given at an ISOC/W3C-
 Benelux joint event in Amsterdam, January 2009"@en ;
 media:height "355"@en ;
 media:presentation &lt;http://static.slidesharecdn.com/swf/ssplayer2.swf?doc=whatistheswamsterdam-
 1232029522157652-26stripped_title=what-is-the-semantic-web-in-15-minutes-presentation&gt; ;
 media:thumbnail &lt;http://cdn.slidesharecdn.com/whatistheswamsterdam-1232029522157652-2-
 thumbnail12123240197&gt; ;
 media:title "What is the Semantic Web (in 15 minutes...)?"@en ;
 media:width "425"@en ;
 xhv:alternate &lt;http://www.slideshare.net/rss/latest&gt; ;
 xhv:icon &lt;http://www.slideshare.net/favicon.ico&gt; ;
</code></pre>
RDFa and HTML5

• You may have seen lots of discussion on this
  • both on the HTML5 and the RDFa mailing lists
  • the discussions are, well, passionate…

• There is a general consensus that some variant of RDFa will be available to (X)HTML5
  • the devil is in the details…
How to “assign” RDF data to a collection of resources?

• Instead of spelling out information for each resource, is it possible to “generate” those?

• Some examples:
  • copyright information for all of your photographs
  • is a Web page collection usable on a mobile phone and how?
  • bibliographical data for a series of publications
  • provenance data for a collection of resources
  • annotation of the data resulting from a scientific experiment
  • etc
How to “assign” RDF data to a collection of resources?

• The issue:
  • given the URI of the resource (photograph, publication, etc), how do I find the relevant RDF data?
POWDER
(Protocol for Web Description Resources)

- Lets you define predicates that can be automatically assigned to a set of resources
POWDER scenario: copyright for photos

1. GET .... index

2. Return .... descr.xml

3. GET .... descr.xml

4. GET .... http://ex3.org/img/imgXXX.jpg

5. Deduce triplets

http://ex1.org/index

http://ex2.org/descr.xml

cc:license <http://cp...> for resources:
http://ex3.org/img/*

http://ex3.org/img/imgXXX.jpg

<http://www.ex3.org/img/imgXXX.jpg> cc:license <http://cp...>
The technical details...

- The “description resource” is an XML file:

```xml
<powder xmlns="http://www.w3.org/2007/05/powder#"
    xmlns:cc="http://creativecommons.org/ns#">
  <attribution>
    <issuedby src="http://www.ivan-herman.net/me"/>
  </attribution>
  <dr>
    <iriset>
      <includehosts>www.ex2.org</includehosts>
      <includepathstartswith>/img/</includepathstartswith>
    </iriset>
    <descriptorset>
      <cc:license rdf:resource="http://cp:..."/>
    </descriptorset>
  </dr>
</powder>
```
The technical details...

- Powder processors may then return
  - direct RDF triples, eg:

```
```

- or can transform this XML file into an OWL for more generic processors
  - a canonical processing of the XML file is defined by the POWDER specification
POWDER Service

- Online POWDER service can be set up:
  - a Web service where
    - one can submit a URI and a resource description file
    - the service return the RDF statements for that URI
  - example:
    - http://www.i-sieve.com/cgi-bin/processor.cgi
Challenge: get the data organized
• Just getting the data out there is not enough
• Some sort of organization, categorization of data is necessary
• Ie, the LOD needs various sorts of vocabularies to rely on
SKOS
(Simple Knowledge Organization System)

- Represent and share classifications, glossaries, thesauri, etc
  - for example:
    - Dewey Decimal Classification, Art and Architecture Thesaurus, ACM classification of keywords and terms…
    - classification of Web 2.0 type tags

- Define classes and properties to add those structures to an RDF universe
  - allow for a quick port of this traditional data, combine it with other data
SKOS

- SKOS is based on a simple structure
  - the central concept is, well, a “SKOS concept”
  - concepts can have preferred and alternate labels
  - a concept may be narrower or broader than another one
  - concepts may be related to one another
  - concepts can be collected in “concept schemes”
  - and that is it (well, almost...)
- Other resources can then refer to these concepts as, eg, their subject
Typical example: LC Subject Headings
Typical example: LC Subject Headings

```
<http://id.loc.gov/authorities/sh85061165#concept>
Type: Topical Term

Broader Terms:
- Fiction
```

```
skos:prefLabel
skos:broader
```
Typical example: LC Subject Headings

```xml
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xhv: <http://www.w3.org/1999/xhtml/vocab#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xs: <http://www.w3.org/2001/XMLSchema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<xhv:alternate
    <http://id.loc.gov/authorities/feed/>
    <http://id.loc.gov/authorities/sh85061165.json>,
    <http://id.loc.gov/authorities/sh85061165.nt>,
    <http://id.loc.gov/authorities/sh85061165.rdf> ;
  xhv:icon <http://www.loc.gov/favicon.ico> ;
  dcterms:created "2000-08-21T00:00:00-04:00"^^xsd:dateTime ;
  dcterms:modified "2000-10-04T10:47:15-04:00"^^xsd:dateTime ;
  dcterms:source "GSAFD, 2000 (Historical fiction. UF Fiction, Historical; Historical novelist"
  skos:broader <http://id.loc.gov/authorities/sh85048050#concept> ;
  skos:closeMatch <http://stitch.cs.vu.nl/vocabularies/rameau/ark:/12148/cb119808101> ;
  skos:inScheme <http://id.loc.gov/authorities#conceptScheme>, <http://id.loc.gov/authorities#conceptScheme> ;
  skos:prefLabel "Historical fiction"@en .
```
Using the LCSH terms…
Another Thesaurus example

Payment behaviour

Zahlungsmoral (german)

used for: Payment behavior, Payment practices

Related Terms
- Collection operations
- Corporate liquidity
- Insolvency
- Legal compliance
- Tax compliance
- Willingness to pay

Subject Categories
- B.02 Corporate Finance and Investment Policy

Persistent Identifier (for bookmarking and linking)

http://zbw.eu/stw/descriptor/24859-1

STW Thesaurus for Economics (v 8.04, 2009-02-16) - Suggestions and comments to the thesaurus team
German National Library of Economics (ZBW) / Leibniz Information Center for Economics - Imprint

The STW Thesaurus for Economics is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 Germany License. Permissions beyond the scope of this license are available at ZBW.

Courtesy of Timo Borst and Joachim Neubert, German Nat. Libr. of Economics, (SWEO Case Study)
SKOS and OWL

- SKOS is geared towards specific (though large) use cases, like
  - taxonomies, glossaries, …
  - annotations of complex structures (including ontologies)
- SKOS is based on a very simple usage of OWL
  - using some simple OWL Full constructions
  - the emphasis is on organization and not on logical inferences
- “OWL is a Harley-Davison, SKOS is a mountain bike” — (Tom Baker, co-chair of the relevant WG)
OWL Working Group

- A Working Group has worked on a revision of OWL (a.ka. OWL 2)

- The goal of the group:
  1. add a few extensions to current OWL that are useful, and is known to be implementable
     - many things happened in research since 2004
  2. define “profiles” of OWL that are:
     - smaller, easier to implement and deploy
     - cover important application areas and are easily understandable to non-expert users
The overall structure has not changed
Some new features in OWL 2

- Syntactic sugars
  - eg, disjoint union of classes
- New constructs for properties
  - property chains, reflexive properties
- Richer annotations
- Extended datatype facilities
  - define a numerical interval as an OWL Datatype class
- Profiles
New constructs for properties: chains

Properties, when applied one after the other, may be subsumed by yet another one:

• “if a person «P» was born in city «A» and «A» is in country «B» then «P» was born in country «B»”
• more formally:

\[
\text{ex:born\_in\_country~\text{owl:propertyChainAxiom}}\ (\text{ex:born\_in\_city~ex:city\_in\_country}).
\]

• More than two constituents can be used
• There are some restrictions for DL to avoid “circular” specifications
Keys

“if two persons have the same email and the same homepages then they are identical”

- Identification is based (in this example) on the identical values of two properties
- The rule applies to persons only
Previous rule in OWL 2

:Person rdf:type owl:Class;
   owl:hasKey (:email :homepage) .
Extended datatypes facilities

• OWL 1 just takes over RDF datatypes, but that is it
• But we might want datatypes “restrictions” in a more complex way; eg, numeric intervals
  • “I am interested in a price range between €5 and €15”
• In OWL 1, one has to use XML Schema datatypes
  • but it is very complex, users and reasoners have to understand a whole different syntax
OWL 2 datatype facets

- For each datatype one may have restrictions “facets” (min, max, length, etc)
- New datatypes can be defined as *datatype ranges*
- These can be used via the usual restriction mechanisms
Definition of a numeric interval in OWL 2

:AllowedPrice rdf:type rdfs:Datatype;
   owl:onDatatype xsd:float;
   owl:withRestriction (  
      [ xsd:minInclusive 5.0 ]
      [ xsd:maxExclusive 15.0 ]
   ) .

- The possible facets depend on the datatype:
  xsd:pattern, xsd:length, xsd:maxLength, ...

...
OWL 2 defines “profiles”

- Further restrictions on how terms can be used and what inferences can be expected
- The semantic approaches are identical, but restrictions may ensure even more manageable implementations
OWL 2 profiles

- Classification and instance queries in polynomial time: \textit{OWL-EL}
- Implementable on top of conventional relational database engines: \textit{OWL-QL}
- Implementable on top of traditional rule engines: \textit{OWL-RL}
An example: OWL-RL

- Goal: to be implementable through rule engines
- Usage follows a similar approach to RDFS:
  - merge the ontology and the instance data into a big RDF graph
  - use a rule engine to add new triples (as long as it is possible)
  - then, for example, use SPARQL to query the resulting (expanded) graph
- This application model is very important for RDF based applications
Example: OWL-QL

- The RL “model” has a downside: a large number of extra triples are added to the graph before query
  - big database vendors have means to handle that
  - but simpler implementations may have a problem there…
- QL is different: map a query directly on SQL
  - ie, the data can be queried without “touching” the dataset
Rules

- OWL 2 RL shows the importance of rule engines
- W3C’s Rule work (RIF) is getting to completion
  - b.t.w., OWL 2 RL can be expressed in RIF
- I have no time to go into details here...
Query the data
Querying the data: SPARQL

- Is a W3C Standard since January 2008
  - it has already become one of the essential technologies on the SW
  - many LOD blobs offer a “SPARQL endpoint”
  - there is even a SPARQL endpoint for almost the whole LOD (installed on the Amazon cloud)
SPARQL as a unifying point!
New SPARQL WG: Goals

- To define a small set of extensions to SPARQL
- No complex change, backward compatibility
- Listen to user and implementation experiences of the past few years
- Group started in February 2009
Planned features

- Update, i.e., ability to change the RDF store
- Service description framework
  - what type of extensions, inference possibilities, etc., are available at the endpoint
- Addition to the query language
  - aggregate functions
  - subqueries
  - negation
  - project expressions
Planned features
(tentative syntax examples)

- Aggregate functions and project expressions:

  ```sql
  SELECT AVG(?age) AS average_age WHERE { .... }
  SELECT (?unitprice * ?totalnumber) AS totalprice WHERE { ... }
  ```

- Subqueries:

  ```sql
  SELECT ?person (SELECT ?n WHERE { ?person foaf:name ?n } LIMIT 1)
  WHERE { <http://www.ivan-herman.net/me> foaf:knows ?person. }
  ```

- Negation:

  ```sql
  SELECT *
  WHERE { ?x :p ?v. UNSAID { ?x :q ?v. } }
  ```
Possible features (time permitting)

- Definition of “entailment regimes”
  - RDFS, OWL Profiles, RIF
- Property paths
- Commonly used functions (e.g., string manipulation)
- Basic control for federated queries
- Additional query language syntax
  - commas in select lists, some operators in filters
Are we done?
Certainly not…

- There are a number of issues, problems
  - missing functionalities, unsolved problems
  - misconceptions, messaging problems
  - need for more applications, deployment, acceptance
  - we need more experts
  - etc
Number of issues raised by the LOD…

• Description of datasets
• Semi-automatic generation of links among datasets
  • it is a bit like ontology alignment but on instances
• The “ID Jungle”
  • what to do when the same concept (like a specific person) has a large number of URI-s
  • owl:sameAs may not be the best choice in all cases
Other items…

• Security, trust, provenance
  • combining cryptographic techniques with the RDF model, sign a portion of the graph, etc
  • trust models
  • W3C may start an exploratory group on provenance in autumn
Other items...

- What does reasoning mean on billions of triples?
  - incomplete reasoning: “give me what you can in 2 minutes, even if it is not complete”
Other items...

• How do you interact and present data?
• Currently the result of integration is converted into HTML
• Ideally, the browser would recognize the type of the data and adapt itself automatically
  – eg, displaying foaf data is very different from, say, gene ontology data...
Other items...

- Fuzzy logic
  - look at alternatives of DL 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

- A W3C Incubator Group issued a report on the current status, possibilities, directions, etc
  - report published in April 2008
Revision of the RDF model?

- Some restrictions in RDF may be unnecessary (bNodes as predicates, literals as subject, …)
- Semantics of bNodes
- Named graphs
- Syntax issues (RDF/XML improvements, other serializations)
- …
Conclusions

• Many things are happening at W3C to evolve the Semantic Web
• Many more issues are still to be done…
• So join the club! After all, this is really a community effort…
Thank you for your attention!
And enjoy the conference!

These slides are also available on the Web:

http://www.w3.org/2009/Talks/0830-Nanjing-IH/