Introduction and Applications of the Semantic Web

Slides by: Ivan Herman, W3C

http://www.w3.org/2009/Talks/05-Oz-IntroSW-IH/
http://www.w3.org/2009/Talks/0615-SanJose-tutorial-IH/
Let’s organize a trip to Budapest from Amsterdam using the Web!
You try to find a proper flight with ...
... a big, reputable airline, or ...
... the airline of the target country, or ...

**SELECT DELIVERY METHOD**

Please select ticket delivery method below!

<table>
<thead>
<tr>
<th>Delivery method</th>
<th>Service fee</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-ticket</td>
<td>EUR 10</td>
<td>with e-invoice</td>
</tr>
<tr>
<td>Airport</td>
<td>EUR 30</td>
<td>KLM ticket office, departure hall 2.</td>
</tr>
<tr>
<td>Office</td>
<td>EUR 30</td>
<td>KLM ticket office, departure hall 2.</td>
</tr>
<tr>
<td>Courier</td>
<td></td>
<td>Courier or mail delivery is available only within the country of departure! You will be asked to provide a delivery address on the next page. Please note that in case of courier delivery we will be unable to deliver your ticket(s) to a PO box.</td>
</tr>
<tr>
<td>Mail</td>
<td>EUR 30</td>
<td></td>
</tr>
</tbody>
</table>

**FLIGHT SUMMARY**

Outbound flight
From: Amsterdam, Schiphol (AMS), Netherlands
Budapest, Ferihev 2A (BUD).

Return flight
From: Budapest, Ferihev 2A (BUD), Hungary
Amsterdam, Schiphol (AMS), Netherlands.
... or a low cost one
You have to find a hotel, so you look for...
... a really cheap accommodation, or ...

Hostel Traveler - Instant online booking for youth hostels, budget hotels and hostelling the globe.

Find Hostels and Lodging at your destinations.

Look for for instant online booking.

3 accommodations have been found matching your criteria. Select your accommodations and click for rates, availability, and reservations.

Tip: Click for rates and instant secure confirmations.

Sort by: Price (Lo-Hi) Price (Hi-Lo) Traveler Rating Hostel Class Hostel Name

Balaton, Hungary

Unity Hostel Balaton
Rakoczi Ut 268
Hostel 8 Units
Write a Review

From €12
... or a really luxurious one, or ...
... an intermediate one ...
oops, that is no good, the page is in Hungarian that almost nobody understands, but...
...this one could work...
Of course, you could decide to trust a specialized site...
... like this one, or...
... or this one
You may want to know something about Budapest; look for some photographs...
... on flickr ...
... on Google ...
... or you can look at mine😊
...or at a (social) travel site

Budapest Travel Guide

This capital city - made up of two parts, Buda and Pest - sits on one of the most beautiful areas of the Danube River and it shows. Often dubbed the “Paris of Eastern Europe”, it is a combination of Old World grandeur and a thriving cultural scene. Budapest has a vibrancy and vitality that never slows and the numerous sights can occupy travelers for weeks. With so much history and culture to explore, no traveler leaves unsatisfied.

more Budapest photos ▶
Destinations near Budapest ▶

Travel Guide Information From Our Partners
What happened here?

- You had to consult a large number of sites, all different in style, purpose, possibly language…
- You had to mentally *integrate* all those information to achieve your goals
- We all know that, sometimes, this is a long and tedious process!
• All those pages are only tips of respective icebergs:
  • the real *data* is hidden somewhere in databases, XML files, Excel sheets, …
  • you have only access to what the Web page designers allow you to see
• Specialized sites (Expedia, TripAdvisor) do a bit more:
  • they gather and combine data from other sources (usually with the approval of the data owners)
  • but they still control how you see those sources
• But sometimes you want to personalize: access the original data and combine it yourself!
• The value is in the *combination* of the data
Here is another example...
Another example: social sites. I have a list of “friends” by...
... Dopplr,
… LinkedIn,
... and, of course, Facebook
• I had to type in and connect with friends again and again for each site independently 😞
• This is even worse then before: I feed the icebergs, but I still do not have an easy access to data…
What would we like to have?

- Use the data on the Web the same way as we do with documents:
  - be able to link to data (independently of their presentation)
  - use that data the way I want (present it, mine it, etc)
  - agents, programs, scripts, etc, should be able to interpret part of that data
Put it another way…

- We would like to extend the current Web to a “Web of data”:
  - allow for applications to exploit the data directly
But wait! Isn’t what mashup sites are already doing?
A “mashup” example:
• In some ways, yes, and that shows the huge power of what such Web of data provides
• But mashup sites are forced to do very ad-hoc jobs
  • various data sources expose their data via Web Services
  • each with a different API, a different logic, different structure
  • these sites are forced to reinvent the wheel many times because there is no standard way of doing things
Put it another way (again)…

• We would like to extend the current Web to a *standard* way for a “Web of data”
But what does this mean?

• What makes the current (document) Web work?
  • people create different documents
  • they give an address to it (ie, a URI) and make it accessible to others on the Web
Steven’s site on Amsterdam (done for some visiting friends)
Then some magic happens…

- Others discover the site and they link to it
- The more they link to it, the more important and well known the page becomes
  - remember, this is what, eg, Google exploits!
- This is the “Network effect”: some pages become important, and others begin to rely on it even if the author did not expect it…
This could be expected...
but this one, from the other side of the Globe, was not…
What would that mean for a Web of Data?

• Lessons learned: we should be able to:
  • “publish” the data to make it known on the Web
    • standard ways should be used instead of ad-hoc approaches
    • the analogous approach to documents: give URI-s to the data
  • make it possible to “link” to that URI from other sources of data (not only Web pages)
    • ie, applications should not be forced to make targeted developments to access the data
    • generic, standard approaches should suffice
  • and let the network effect work its way…
Example: combine data from experiments

- A drug company has huge amount of old experimental data on its Intranet
- Data in different formats (XML, databases, …)
- To reuse them:
  - make the important facts available on the Web via standards
  - use off-the-shelf tool to integrate, display, search

 Courtesy of Nigel Wilkinson, Lee Harland, Pfizer Ltd, Melliyal Annamalai, Oracle (SWEO Case Study)
But it is a little bit more complicated 😞

- On the traditional Web, humans are implicitly taken into account
- A Web link has a “context” that a person may use
Eg: address field on my page:

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: 5a2f9241f0812f2a2d7736a203e167095f45c20b8)

Postal address:
C/o Centre for Mathematics and Computer Sciences (CWI)
Krijn de Graft 413, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands.

Phone numbers:
phone: +31-20-5924163
mobile phone: +31-541044153
fax: +31-20-5924312

PGP/GPG:
My GnuPG key and signature is available on-line.

FOAF:
You can either extract a short FOAF information from this page, or consult my more complete, public FOAF file.

Misc:
... leading to this page
• A human understands that this is an institution’s home page
• He/she knows what it means (realizes that it is a research institute in the Netherlands)
• On a Web of Data, something is missing; machines can’t make sense of the link alone
• New lesson learned:
  • extra information ("label") must be added to a link: “this links to an institution, which is a research institute”
  • this information should be machine readable
• This is a characterization (or “classification”) of both the link and its target
  • in some cases, the classification should allow for some limited “reasoning”
    • eg, if an address refers to Amsterdam, then this means it is also in the Netherlands
Let us put it together

- What we need for a Web of Data:
  - use URI-s to publish data (not only full documents)
  - allow the data to link to other data
  - characterize/classify the data and the links (the “terms”) to convey some extra meaning
  - and use standards for all these!
So What is the Semantic Web?
It is a collection of standard technologies to realize a Web of Data
• It is that simple…
• Of course, the devil is in the details
  • a common model has to be provided for machines to describe, query, etc, the data and their connections
  • technologies should be around to “export” the data
  • the “classification” of the terms can become very complex for specific knowledge areas: this is where ontologies, thesauri, etc, enter the game…
  • but these details are fleshed out by experts as we speak!

[Icons for RDF, SPARQL, OWL, SKOS]
Example: find the right experts at NASA

- NASA has nearly 70,000 civil servants over the whole of the US
- Their expertise is described in 6-7 databases, geographically distributed, with different data formats, access types...
- Task: find the right expert for a specific task within NASA!
Example: find the right experts at NASA

- Approach: integrate all the data with standard means, and describe the data and links using generic (and simple) vocabularies.

Michael Grove, Clark & Parsia, LLC, and Andrew Schain, NASA, (SWEO Case Study)
Wait! Does it mean that I have to convert all my data in some way?
• Not necessarily; this would not always be feasible
• There are technologies to make your data accessible to standard means without converting it
  • run-time “bridges” (eg, rewriting queries on the fly)
  • annotate existing data (eg, XHTML pages)
  • extract data from XHTML/XML files
  • etc
• Some of these techniques are still being developed
Example: “Linking Open Data Project”

- Goal: “expose” open datasets for integration
- *Set links among the data items* from different datasets
- Set up query endpoints
- Altogether billions of relationships, millions of links…
Example data source: DBpedia

- DBpedia is a community effort to
  - extract structured (“infobox”) information from Wikipedia
  - provide a query endpoint to the dataset
  - interlink the DBpedia dataset with other datasets on the Web
The LOD “cloud”, March 2008
The LOD “cloud”, September 2008

As of September 2008
The LOD “cloud”, March 2009
All this sounds nice, but isn’t that just a dream?
The 2007 Gartner predictions

During the next 10 years, Web-based technologies will improve the ability to embed semantic structures [... it] will occur in multiple evolutionary steps...

By 2017, we expect the vision of the Semantic Web [...] to coalesce [...] and the majority of Web pages are decorated with some form of semantic hypertext.

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

(note: “semantic hypertext” refers to pages “prepared” for integration)
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, GE, Northrop Gruman, Altova, Microsoft, Dow Jones, …

• Others are using it (or consider using it) as part of their own operations: Novartis, Pfizer, Telefónica, …

• 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, Pfizer, Sun, Eli Lilly, …
Lots of Tools (not an exhaustive list!)

- Categories:
  - Triple Stores
  - Inference engines
  - Converters
  - Search engines
  - Middleware
  - CMS
  - Semantic Web browsers
  - Development environments
  - Semantic Wikis
  - …

- Some names:
  - Jena, AllegroGraph, Mulgara, Sesame, flickurl, …
  - TopBraid Suite, Virtuoso environment, Falcon, Drupal 7, Redland, Pellet, …
  - Disco, Oracle 11g, RacerPro, IODT, Ontobroker, OWLIM, Tallis Platform, …
  - RDF Gateway, RDFLib, Open Anzo, DartGrid, Zitgist, Ontotext, Protégé, …
  - Thetus publisher, SemanticWorks, SWI-Prolog, RDFStore…
  - …
Some deployment communities

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

• Health care and life science sector is now very active
  • also at W3C, in the form of an Interest Group
Application specific portions of the cloud

- Eg, “bio” related datasets
  - done, partially, by the “Linking Open Drug Data” task force of the HCLS IG at W3C
Help in choosing the right drug regimen

- Help in finding the best drug regimen for a specific case, per patient
- Integrate data from various sources (patients, physicians, Pharma, researchers, ontologies, etc)
- Data (e.g., regulation, drugs) change often, but the tool is much more resistant against change

_Courtesy of Erick Von Schweber, PharmaSURVEYOR Inc., (SWEO Use Case)
Yahoo’s SearchMonkey

- Search based results may be customized via small applications
- Metadata embedded in pages are reused
- Publishers can export extra data via other formats

Courtesy of Peter Mika, Yahoo! Research, (SWEO Case Study)
What is the Semantic Web? (In 15 minutes...)

ISOC Nieuwjaarsreceptie 2009

2009-01-15, Amsterdam, The Netherlands

Ivan Herman, W3C
Information in Web Pages: SlideShare
**Improved Search (GoPubMed)**

- Search results are re-ranked using ontologies
- Related terms are highlighted, usable for further search
**Improved Search (Go3R)**

- Same dataset, different ontology
  - (ontology is on non-animal experimentation)
New type of Web 2.0 applications

• New Web 2.0 applications come every day
• Some begin to look at Semantic Web as possible technology to improve their operation
  • more structured tagging, making use of external services
  • providing extra information to users
  • etc.
• Some examples: Twine, Revyu, Faviki, …
Integration of “social” software data

- Internal usage of wikis, blogs, RSS, etc, at EDF
  - goal is to manage the flow of information better
- Items are integrated via
  - Semantic Web based unifying format
  - simple, public vocabularies
  - internal data is combined with linked open data like Geonames
  - Semantic Web queries are used for internally
- Details are hidden from end users (via plugins, extra layers, etc)

Courtesy of A. Passant, EDF R&D and LaLIC, Université Paris-Sorbonne, (SWEO Case Study)
Integration of “social” software data

| Inline macro | [onto|members] |
|--------------|---------------|
| Domaines:    | [onto|domaines]|
| Localisation:|               |
| Ville, Pays:  | Paris, France |
| Répéter ce champ |               |
| Rattachement:|               |
| Affiliation principale: |     |
| Lien:         |               |
| Répéter ce champ |               |
| Domaine et Métier: |       |
| Domaine:      |               |
| Activités Associatives | |
| Métier:       |               |
| Associatif    |               |
| Zone Géographique: | France  |

Courtesy of A. Passant, EDF R&D and LaLIC, Université Paris-Sorbonne, (SWEO Case Study)
Integration of “social” software data

Courtesy of A. Passant, EDF R&D and LaLIC, Université Paris-Sorbonne, (SWEO Case Study)
Conclusions…

- More and more data should be “published” on the Web
  - this can lead to the “network effect” on data
- New breeds of applications come to the fore
  - “mashups on steroids”
  - better representation and usage of community knowledge
  - new customization possibilities
  - …
The rough structure of data integration

1. Map the various data onto an abstract data representation
   - make the data independent of its internal representation…
2. Merge the resulting representations
3. Start making queries on the whole!
   - queries that could not have been done on the individual data sets
A simplified bookstore data (dataset “A”)

<table>
<thead>
<tr>
<th>ID</th>
<th>ISBN0-00-651409-X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>id_xyz</td>
</tr>
<tr>
<td>Title</td>
<td>The Glass Palace</td>
</tr>
<tr>
<td>Publisher</td>
<td>id_qpr</td>
</tr>
<tr>
<td>Year</td>
<td>2000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Home Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>id_xyz</td>
<td>Ghosh, Amitav</td>
<td><a href="http://www.amitavghosh.com">http://www.amitavghosh.com</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Publ. Name</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>id_qpr</td>
<td>Harper Collins</td>
<td>London</td>
</tr>
</tbody>
</table>
1st: export your data as a set of relations

- The Glass Palace
  - a:title
- 2000
  - a:year
- London
  - a:city
- Harper Collins
  - a:p_name
- Ghosh, Amitav
  - a:name
- http://www.amitavghosh.com
  - a:homepage
- http://.../isbn/000651409X
  - a:author
  - a:publisher
Some notes on the exporting the data

- Relations form a graph
  - the nodes refer to the “real” data or contain some literal
  - how the graph is represented in machine is immaterial for now
- Data export does not necessarily mean physical conversion of the data
  - relations can be generated on-the-fly at query time
    - via SQL “bridges”
    - scraping HTML pages
    - extracting data from Excel sheets
    - etc.
- One can export part of the data
Another bookstore data (dataset “F”)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID</td>
<td>Titre</td>
<td>Traducteur</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6</th>
<th>ID</th>
<th>Auteur</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>ISBN-0-00-651409-X</td>
<td>A12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11</th>
<th>Nom</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Ghosh, Amitav</td>
</tr>
<tr>
<td>13</td>
<td>Besse, Christianne</td>
</tr>
</tbody>
</table>
2\textsuperscript{nd}: export your second set of data
3rd: start merging your data
3rd: start merging your data (cont.)
3rd: merge identical resources
Start making queries…

- User of data “F” can now ask queries like:
  - “give me the title of the original”
    - well, … « donnez-moi le titre de l’original »
  - This information is not in the dataset “F”…
  - …but can be retrieved by merging with dataset “A”!
However, more can be achieved...

- We “feel” that \texttt{a:author} and \texttt{f:auteur} should be the same
- But an automatic merge does not know that!
- Let us add some extra information to the merged data:
  - \texttt{a:author} same as \texttt{f:auteur}
  - both identify a “Person”
  - a term that a community may have already defined:
    - a “Person” is uniquely identified by his/her name and, say, homepage
    - it can be used as a “category” for certain type of resources
3rd revisited: use the extra knowledge

The Glass Palace
- a:title
- a:year
- a:city
- a:p_name
- foaf:name
- foaf:homepage

2000
- a:year

London
- a:city

Harper Collins
- a:p_name

Le palais des mirroirs
- f:titre
- f:trajecteur

http://.../isbn/000651409X
- a:author
- f:auteur
- f:original

http://.../isbn/2020386682
- f:original

Ghosh, Amitav
- foaf:name

http://www.amitavghosh.com

Besse, Christiane
- foaf:name
Start making richer queries!

- User of dataset “F” can now query:
  - “donnes-moi la page d’accueil de l’auteur de l’originale”
  - well… “give me the home page of the original’s ‘auteur’”

- The information is not in datasets “F” or “A”…

- …but was made available by:
  - merging datasets “A” and datasets “F”
  - adding three simple extra statements as an extra “glue”
Combine with different datasets

- Using, e.g., the “Person”, the dataset can be combined with other sources
- For example, data in Wikipedia can be extracted using dedicated tools
  - e.g., the “dbpedia” project can extract the “infobox” information from Wikipedia already…
Merge with Wikipedia data
Merge with Wikipedia data
**Is that surprising?**

- It may look like it but, in fact, it should not be…
- What happened via automatic means is done every day by Web users!
- The difference: a bit of extra rigour so that machines could do this, too
What did we do?

• We combined different datasets that
  • are somewhere on the web
  • are of different formats (mysql, excel sheet, XHTML, etc)
  • have different names for relations
• We could combine the data because some URI-s were identical (the ISBN-s in this case)
• We could add some simple additional information (the “glue”), possibly using common terminologies that a community has produced
• As a result, new relations could be found and retrieved
It could become even more powerful

- We could add extra knowledge to the merged datasets
  - e.g., a full classification of various types of library data
  - geographical information
  - etc.
- This is where **ontologies**, extra **rules**, etc, come in
  - ontologies/rule sets can be relatively simple and small, or huge, or anything in between…
- Even more powerful queries can be asked as a result
What did we do? (cont)