The Vision of the Semantic Web

InfoTECH, London, March 14, 2007

Ivan Herman, W3C
Data(base) integration
Databases are very different in structure, in content

Lots of applications require managing *several* databases
  * after company mergers
  * combination of administrative data for e-Government
  * biochemical, genetic, pharmaceutical research
  * etc.

Most of the public data are accessible from the Web; proprietary data may not be yet
  * *this should be done for easier collaboration!*
What Is Needed?

- (Some) data should be available for machines for further processing
- Data should be possibly combined, merged on a Web scale (also referred to as “data integration”)
- Machines may also need to *reason* about that data
- *What we need is a “Web of Data”*
- Let us walk through a simple example…
### A simplified bookstore data (dataset “A”)

<table>
<thead>
<tr>
<th>ID</th>
<th>Author</th>
<th>Title</th>
<th>Publisher</th>
<th>Year</th>
</tr>
</thead>
</table>

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

<table>
<thead>
<tr>
<th>ID</th>
<th>Publisher Name</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>id_qpr</td>
<td>Harper Collins</td>
<td>London</td>
</tr>
</tbody>
</table>
1st step: export your data as a set of relations
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>ID</th>
<th>Titre</th>
<th>Auteur</th>
<th>Traducteur</th>
<th>Original</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Nom</th>
</tr>
</thead>
<tbody>
<tr>
<td>i_abc</td>
<td>Amitav Ghosh</td>
</tr>
<tr>
<td>i_qrs</td>
<td>Christiane Besse</td>
</tr>
</tbody>
</table>
2nd step: export your second set of data
3rd step: start merging your data
3rd step: start merging your data (cont.)
3rd step: merge identical resources
Start making queries…

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

- We “feel” that `a:author` and `f:auteur` should be the same
- But an automatic merge does not know that!
- Let us add some extra information to the merged data:
  - `a:author` same as `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 step revisited: use the extra knowledge
Start making richer queries!

- User of dataset “F” can now query:
  - «donnes-moi la page d’accueil de l’auteur de l’original»
  - (i.e., “give me the home page of the original’s author”)
- The data is not in dataset “F”…
- …but was made available by:
  - merging datasets “A” and datasets “F”
  - adding three simple extra statements as an extra “glue”
  - using existing terminologies as part of the “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:
  - *there is an active development to add some simple semantic “tag” to wikipedia entries*
  - *we tacitly presuppose their existence in our example*...
Merge with Wikipedia data
Is that surprising?

- Maybe but, in fact, no…
- What happened via automatic means is done all the time, every day by the users of the Web!
- The difference: a bit of extra rigor (e.g., *naming* the relationships) is necessary so that machines could do this, too
It could become even more powerful

- We could add extra knowledge to the merged datasets
  - e.g., a full classification of various type of library data
  - geographical information
  - etc.
- This is where ontologies, extra rules, etc, may come in
- Even more powerful queries can be asked as a result
What did we do?

Applications

Query, Manipulate, etc.

Data represented in abstract format

Map, Expose, etc.

Data in various formats
The abstraction pays off because...

- ... the graph representation is independent on the *exact* structures in, say, a relational database
- ... a change in local database schemas, XHTML structures, etc, do *not* affect the whole, only the “export” step
  - “schema independence”
- ... new data, new connections can be added seamlessly, regardless of the structure of other data sources
So where is the Semantic Web?

- The Semantic Web provides technologies to make such integration possible!
- It is a suite of technologies for the abstract data model, querying the data, defining ontologies, taxonomies, etc.
  - *I do not have time for the exact details here…*
The “bio” domain and the Semantic Web: ontologies

- A number of ontologies have been developed already:
  - the US Cancer Institute’s Cancer Ontology, the Gene Ontology, the BioPax Molecular Pathway Ontology, the SWAN Project for the Alzheimer Disease research community, bio-zen ontology in neuroscientific and biomedical research, BrainPharm (Pathological Mechanisms in Alzheimer's Disease) from Yale Univ., ...
  - These are available in the ontology language defined by W3C
    - huge and powerful “glues” in our example above!
A number of data sets are being exposed:

- **BrainPharm** and **SWAN** data cited above, NIST’s data on Thermodynamics of Enzyme-Catalyzed Reactions, RDF version of UniProt, …

- Work is going on to develop general methods for further data exposures; see, eg:
  - [http://esw.w3.org/topic/HCLSIG_BioRDF_Subgroup/Data](http://esw.w3.org/topic/HCLSIG_BioRDF_Subgroup/Data)
Health Care and Life Sciences Interest Group

- There has been a great interest in these technologies from a number of R&D groups
- W3C formed the “Health Care and Life Sciences Interest Group”
  - goal is to explore the feasibility of these technologies
  - build demonstrations, explore possibilities
  - group runs until end of 2007, next steps are being explored
- Participants include Merck, AstraZeneca, Pfizer, Teranode, Partners HealthCare, IBM, Oracle, Agfa, HP, Universities of Amsterdam, Manchester, Yale, …
Some work areas of the HCLS IG

- Various “task forces”:
  - *Develop techniques to convert, export, access, etc, biomedical data for data integration*
    - e.g., conversion of the Entrez Gene data into RDF (33GB of data, from XML to RDF; the size was reduced during conversion!)
  - *Evaluate, facilitate, etc, the creation of core vocabularies and ontologies in the area, possibly develop usage patterns*
  - *Look at clinical pathways, accommodate them with both patient heterogeneity and evolving clinical context*
  - *“Drug Safety and Efficacy”, ie, integrating the various steps needed in, eg, in an FDA approval process, clinical trial planning, reporting, management, etc.*

- Use cases/demonstrations are being developed on data integration; a Workshop is organized in Banff, Canada, in May 2007
Other examples

- “Online community for knowledge sharing between clinicians in oral medicine in Sweden” (by Marie Gustafsson and others)
- Application for Traditional Chinese Medicine (by Huajun Chen and others)
  - integration of over 70 databases with a shared ontology
Semantic Web Applications

- Data integration is but one paradigm of Semantic Web usage
- Some others include knowledge management, labelling (multimedia) data, content adaptation, semantically oriented search,…
- Lots of tools are at disposal; e.g., Oracle’s 10g database is prepared for Semantic Web data storage and integration
Conclusions

- The Semantic Web is there to integrate data on the Web
- The goal is the creation of a *Web of Data*
- A major new avenue for Health Care and Life Sciences
Thank you for your attention!

These slides are publicly available on:


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