Introduction to Semantic Web

Ivan Herman
The Music site of the BBC

Eric Clapton
Born 30 March 1945.

Biography

Eric Patrick Clapton, CBE (born 30 March 1945) is an English blues-rock guitarist, singer, songwriter and composer. Clapton has been inducted into the Rock and Roll Hall of Fame as a solo performer, as a member of rock bands: the Yardbirds and Cream. Clapton is the only person ever to be inducted three times. Often viewed by critics and fans alike as one of the most important and influential guitarists of all time, Clapton was ranked fourth in Rolling Stone magazine's list of the "100 Greatest Guitarists of All Time."
The Music site of the BBC
How to build such a site 1.

- Site editors roam the Web for new facts
  - may discover further links while roaming
- They update the site manually
- And the site gets soon out-of-date 😞
How to build such a site 2.

- Editors roam the Web for new data published on Web sites
- “Scrape” the sites with a program to extract the information
  - Ie, write some code to incorporate the new data
- Easily get out of date again… 😞
How to build such a site 3.

- Editors roam the Web for new data via API-s
- Understand those…
  - input, output arguments, datatypes used, etc
- Write some code to incorporate the new data
- Easily get out of date again… 😞
The choice of the BBC

- Use external, public datasets
  - Wikipedia, MusicBrainz, …
- They are available *as data*
  - not API-s or hidden on a Web site
  - data can be extracted using, eg, HTTP requests or standard queries
In short...

- Use the Web of Data as a Content Management System
- Use the community at large as content editors
And this is no secret...
Data on the Web

- There are more and more data on the Web
  - government data, health related data, general knowledge, company information, flight information, restaurants,…
- More and more applications rely on the availability of that data
But… data are often in isolation, “silos”
Imagine...

- A “Web” where
  - documents are available for download on the Internet
  - but there would be no hyperlinks among them
And the problem is real...
**Data on the Web is not enough...**

- We need a proper infrastructure for a real **Web of Data**
  - data is available on the Web
    - accessible via standard Web technologies
  - data are interlinked over the Web
  - ie, data can be *integrated* over the Web
- This is where Semantic Web technologies come in
i.e., ..., connect the silos

Photo credit "kxlly", Flickr
In what follows…

- We will use a simplistic example to introduce the main Semantic Web concepts
The rough structure of data integration

- Map the various data onto an abstract data representation
  - make the data independent of its internal representation...
- Merge the resulting representations
- Start making queries on the whole!
  - queries not possible on the individual data sets
We start with a book...

Amitav Ghosh

THE GLASS PALACE

The magnificent, poignant, fascinating novel of three generations that starts in Mandalay...
## A simplified bookstore data (dataset “A”)

<table>
<thead>
<tr>
<th>ISBN</th>
<th>Author</th>
<th>Title</th>
<th>Publisher</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>0006511409X</td>
<td>id_xyz</td>
<td>The Glass Palace</td>
<td>id_qpr</td>
<td>2000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Homepage</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>Publisher’s 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
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
Same book in French…
Another bookstore data (dataset “F”)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID</td>
<td>Titre</td>
<td>Traducteur</td>
<td>Original</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>ID</td>
<td>Auteur</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ghosh, Amitav
Besse, Christianne
2\textsuperscript{nd}: export your second set of data

http://…isbn/000651409X

http://…isbn/2020386682

Le palais des miroirs

Ghosh, Amitav

Besse, Christianne

f:nom

f:original

f:titre

f:traducteur

f:nom
3rd: start merging your data

The Glass Palace
2000
London
Harper Collins
Ghosh, Amitav

http://www.amitavghosh.com
http://...isbn/000651409X
http://...isbn/2020386682

f:nom
f:auteur
f:traducteur
f:titre
f:nom

Ghosh, Amitav
Le palais des miroirs
Besse, Christianne
3rd: start merging your data (cont)

The Glass Palace

2000

London

Harper Collins

Ghosh, Amitav

http://www.amitavghosh.com

http://...isbn/000651409X

http://...isbn/000651409X

Le palais des miroirs

Besse, Christianne

http://...isbn/2020386682

Same URI!
3rd: start merging your data

The Glass Palace

2000

London

Harper Collins

Ghosh, Amitav

http://www.amitavghosh.com

http://...isbn/000651409X

Le palais des miroirs

http://...isbn/2020386682

Besse, Christianne

Ghosh, Amitav

http://www.amitavghosh.com
User of data “F” can now ask queries like:
  - “give me the title of the original”
    - well, … « donnes-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 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 revisited: use the extra knowledge

- The Glass Palace
- 2000
- London
- Harper Collins
- Ghosh, Amitav

http://www.amitavghosh.com

- Le palais des miroirs
- Ghosh, Amitav

http://www.amitavghosh.com

- Besse, Christianne

http://www.amitavghosh.com
Start making richer queries!

- User of dataset “F” can now query:
  - “donnes-moi la page d’accueil de l’auteur de l’original”
    - 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

The Glass Palace
2000
London
Harper Collins
Ghosh, Amitav
http://www.amitavghosh.com
http://dbpedia.org/..//Amitav_Ghosh

http://...isbn/000651409X
f:original
Le palais des miroirs
http://...isbn/2020386682
f:titre
Besse, Christianne

f:name
f:nom
a:name
a:nom

f:author
a:author
r:type
http://...foaf/Person

f:traducteur
r:type

f:auteur
r:type

f:publishe

a:city

a:country

f:year

a:year

f:publishe

a:title

a:author

f:author

w:reference

foaf:name

a:homepage

http://www.amitavghosh.com
Merge with Wikipedia data

The Glass Palace
2000
London
Harper Collins
Ghosh, Amitav
http://www.amitavghosh.com
http://dbpedia.org/..../Amitav_Ghosh
http://dbpedia.org/..../The_Hungry_Tide
http://dbpedia.org/..../The_Calcutta_Chromosome

Le palais des miroirs
f:titre
f:traducteur

Besse, Christianne

http://dbpedia.org/..../Kolkata

http://dbpedia.org/..../The_Glass_Palace

http://dbpedia.org/..../The_Hungry_Tide

http://dbpedia.org/..../The_Calcutta_Chromosome

http://dbpedia.org/..../Kolkata
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
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)

Applications

Data represented in abstract format

Data in various formats

Manipulate
Query

Map, Expose,
...
The abstraction pays off because…

- … the graph representation is independent of the exact structures
- … a change in local database schema’s, XHTML structures, etc, do not affect the whole
  - “schema independence”
- … new data, new connections can be added seamlessly
The network effect

- Through URI-s we can link any data to any data
- The “network effect” is extended to the (Web) data
- “Mashup on steroids” become possible
So where is the Semantic Web?

- The Semantic Web provides technologies to make such integration possible!