Structured Data within the Web of Data

Cristian LAI

ISOC, NIT
Outline

- Motivation
- Unstructured Data
- Structured Data
- Applications
Context
Semantic Web

http://www.w3.org/2006/Talks/1023-sb-W3CTechSemWeb/
Motivation
Search on the Web

The Future of Search

- Semantic technologies help to regain
- Productivity in the face of overwhelming
- Information growth...

The Semantic Web
Web 3.0
2009
- Natural language search
- Automatic semantic tagging (Ontologies)

The Social Web
Web 2.0
2009
- Keyword search
- Human social tagging “folksonomies”

The World Wide Web
Web 1.0
1999
- Directories
- Files & Folders
- Databases

PC Era
1979

The Desktop

The Intelligent Web
Web 4.0
2018
- Web scale reasoning
- Intelligent agents

As amount of data grows, keyword search is becoming less productive...

http://www.slideshare.net/novaspivack/web-evolution-nova-spivack-twine
Outline

- Motivation
- UnStructured Data
- Structured Data
- Applications
Wikipedia

- Is a multilingual, web-based, free-content encyclopedia project based on an openly editable model.
- Is the 5th site on the web and serves 454 million unique visitors monthly as of March 2011.
- Has fewer than 100 employees.
- Wikipedia holds an annual fundraiser instead of accepting advertising. You may have seen "A personal appeal from Wikipedia founder Jimmy Wales" if you’ve used the online encyclopedia during the last weeks of 2011. Google co-founder Sergey Brin and his wife, Anne Wojcicki, has given a 500,000 dollars grant to help Wikipedia fund its 28.3 million dollars annual budget.
Wikipedia

● Pros:
  ○ Is a highly-efficient not-for-profit organization.
  ○ Is the finest example of truly collaborative created content: >19M articles;
    >270 languages, >82k active contributors.
  ○ Covers many topics and domains, articles are a result of a community
    consensus.

● Cons:
  ○ Contains many inconsistencies.
    ● Disclaimer: Wikipedia cannot guarantee the validity of the information found here.
  ○ Is not very well integrated with other data sources.
  ○ Queries and search are not facilitated due to the lacks of structured
    representation.
Issues

- UnStructured data, keywords based search.
- Simple questions are hard to answer.
  - People who were born in Rome before 1900.
  - Italian musicians with English and French descriptions.
  - The official websites of companies with more than 500 employees.
- The information required to answer these is contained in Wikipedia.
- Transforming Wikipedia into a knowledge base.
  - To reveal the structure and semantics of Wikipedia content
  - The DBpedia project.
Structure in Wikipedia

- Wikipedia articles consist mostly of free text, but also contain different types of structured information, such as infobox templates, categorisation information, images, geo-coordinates, and links to external Web pages.
- Title
- Abstract
- Infobox Template
- Geo-coordinates
- Categories
- Images
- Links
  - other language version
  - other Wikipedia pages
  - redirects
  - disambiguation
Structured Information in Wikipedia

- Title
- Abstract
- Infobox Template
- Geo-coordinates
- Categories
- Images
- Links
  - other language version
  - other Wikipedia pages
  - redirects
  - disambiguation

![Location of Cagliari in Italy](image)

Coordinates: 39°14′47″N 09°03′27″E

July 18, 2012
Structured Information in Wikipedia

- Title
- Abstract
- Infobox Template
- Geo-coordinates
- Categories
- Images
- Links
  - other language version
  - other Wikipedia pages
  - redirects
  - disambiguation
Structured Information in Wikipedia

- Title
- Abstract
- Infobox Template
- Geo-coordinates
- Categories
- Images
- Links
  - other language versions
  - other Wikipedia pages
  - redirects
  - disambiguation
Outline

- Motivation
- UnStructured Data
- Structured Data
- Applications
An environment for collecting and structuring data.

Well defined structure of classification.
RDF

- Triples: \((subject, predicate, object)\)
- \textit{Subject} and \textit{object}
  - are both URIs that each identify a resource, or a URI and a string literal respectively.
- \textit{Predicate}
  - specifies how the subject and object are related, and is also represented by a URI.
- For example:
  - A \textit{knows} B
  - C \textit{isAuthorOf} D
  - Two resources linked in this fashion can be drawn from different data sets on the Web, allowing data in one data source to be linked to that in another, thereby creating a Web of Data.
DBpedia

- Started in 2007.
- Is the result of a community effort to extract structured information from Wikipedia.
- Makes Wikipedia data available as RDF.
- Results: **The DBpedia Data Set**
  - describes 3.64 million "things" with over half a billion "facts" (July 2011), 364k persons, 462k places, 99k music albums, 54k films, 148k organisations;
  - extraction in 97 different languages;
  - 672M RDF triples
- It is maintained by: Universität Leipzig, Freie Universität Berlin, OpenLink Software, Inc.
- See [http://wiki.dbpedia.org/Team](http://wiki.dbpedia.org/Team)
Nucleus of the Web of Data

- Within the W3C Linking Open Data (LOD) community effort.
- Tim Berners-Lee’s Linked Data principles.
  - URI
  - HTTP
  - RDF, SPARQL
  - Interlinking among data providers
- An increasing number of data providers have started to publish and interlink data on the Web.
- Several billion RDF triples and covers domains such as geographic information, people, companies, online communities, films, music, books and scientific publications.
LOD Datasets
The DBpedia SPARQL endpoint

- All data sets are available for queries via the DBpedia SPARQL endpoint (http://dbpedia.org/sparql).
- Querying the data set:
  - ... 
  - The official websites of companies with more than 50000 employees.
  - Cities with more than 2 million habitants.
  - ...
Abstracts of movies starring Tom Cruise, released before 1999

```
    ?subject rdf:type <http://dbpedia.org/ontology/Film>.
    FILTER(lang(?abstract) = "en" && lang(?label) = "en").
    FILTER(xsd:date(?released) < "2000-01-01"^^xsd:date).
} ORDER BY ?released
```

<table>
<thead>
<tr>
<th>subject</th>
<th>label</th>
<th>released</th>
<th>abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td>:Losin%27_It</td>
<td>&quot;Losin' It&quot;@en</td>
<td>&quot;1983-04-08&quot;</td>
<td>&quot;Losin' It is a 1983 comedy film starring Tom Cruise, Shelley Long, Jackie Earle Haley, and John</td>
</tr>
<tr>
<td></td>
<td></td>
<td>xsd:dateTime</td>
<td>Stockwell. The film is directed by Curtis Hanson. It was filmed largely in Calexico, California.&quot;@en</td>
</tr>
<tr>
<td>:Risky_Business</td>
<td>&quot;Risky Business&quot;@en</td>
<td>&quot;1983-08-05&quot;</td>
<td>&quot;Risky Business is a 1983 American teen comedy-drama film written by Paul Brickman in his</td>
</tr>
<tr>
<td></td>
<td></td>
<td>xsd:dateTime</td>
<td>directorial debut. It stars Tom Cruise and Rebecca De Mornay. The hit film launched Cruise to</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>stardom. @en</td>
</tr>
<tr>
<td>:All_the_Right_Moves</td>
<td>&quot;All the Right Moves&quot;@en</td>
<td>&quot;1983-10-21&quot;</td>
<td>&quot;For the OneRepublic song, see All the Right Moves (song) All the Right Moves is a 1983 drama</td>
</tr>
<tr>
<td></td>
<td></td>
<td>xsd:dateTime</td>
<td>film directed by Michael Chapman and starring Tom Cruise, Craig T. Nelson, Lea Thompson, Chris</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Penn, and Gary Graham. It was filmed on location during WPFL football season in Johnstown,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pennsylvania, and Pittsburgh,&quot;@en</td>
</tr>
<tr>
<td>:Legend%28film%29</td>
<td>&quot;Legend (film)&quot;@en</td>
<td>&quot;1985-12-13&quot;</td>
<td>&quot;Legend is a 1985 fantasy film released by Universal Pictures, directed by Ridley Scott and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>xsd:dateTime</td>
<td>starring Tom Cruise, Mia Sara, and Tim Curry. Though not a very notable success when first</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>released, it received an Academy Award nomination (for Best Makeup) and has since developed a</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cult following,&quot;@en</td>
</tr>
<tr>
<td></td>
<td></td>
<td>xsd:dateTime</td>
<td>Walter Tevis. The film continues the story of pool hustler and stakehorse Edward &quot;Fast Eddie&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Felson from Tevis' first novel, The Hustler (1959), with Paul Newman reprising his role from its</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>film adaptation (1961).&quot;@en</td>
</tr>
</tbody>
</table>
Outline

- Motivation
- UnStructured Data
- Structured Data
- Applications
A number of search engines have been developed that crawl Linked Data from the Web by following RDF links, and provide query capabilities over aggregated data.


- Google, Bing and Yahoo! agree to create and support a common vocabulary for structured data markup on web pages.
- Facebook has started to support RDF and Linked Data URIs and now provides access to parts of its user data via a Linked Data API.
Google rich snippets

Westin Excelsior, Rome
www.westin.com/excelsiorrome
Zagat: 25 / 30 - 16 Google reviews - $379
Via Veneto, 125 00187 Province of Rome, Italy
06 47081

The Westin Excelsior, Rome - Starwood Hotels & Resorts Worldwid... www.starwoodhotels.com/westin/property/overview/index.html?
Experience a Legend. Since 1908, The Westin Excelsior, Rome has hosted celebrities, statesmen and artists visiting the Eternal City. The hotel, which was ...
+ Show map of Via Veneto, 125, 00187 Rome, Italy
Rooms - Contact Us - Photos - Local Area

Hotel The Westin Excelsior Rome, Rome Hotels, Italy Hotels excelsior.hotelinroma.com/
Hotel The Westin Excelsior Rome, Rome Hotels: Located in Via Veneto and recently refurbished, this hotel is a monument to turn-of-the-century style.

The Westin Excelsior, Rome (Rome, Italy) - Hotel Reviews ... www.tripadvisor.com/Hotel_Review-g187791-d203060-Reviews-Th...
★★★★★ 678 reviews - Price range: $$$
The Westin Excelsior, Rome, Rome: See 678 traveler reviews, 228 candid photos, and great deals for The Westin Excelsior, Rome, ranked #229 of 1278 hotels ...
Twitter, #annotations

Semantic #annotations for Twitter

Cristian Lai
Sardegna, Italia.
Timezone: Rome
Followers: 25
Friends: 49
Statuses: 80

Twitter Feed
[TEST] Usain Bolt #dbpedia:Usain_Bolt vs Taysom Gay #dbpedia:Olympic_Games 36 hours ago

[TEST] Usain Bolt #dbpedia:Usain_Bolt vs Taysom Gay 36 hours ago

[TEST] Google #dbpedia:Google 36 hours ago

#dbpedia:United_States

RT @DART2012ws: DART2012 http://t.co/mY6gvZt3. Deadline EXTENDED, June 29. 21 days ago
Q & A