



Querying Wikipedia like a database

Jinesh Shaji George

Wikipedia

- 5th most popular website (according to Alexa.com)
- Maybe the finest example of truly collaboratively created content (>20M articles, >280 languages, >90,000 contributors)
- Lacks structured representation of content which facilitate querying and search
- Search capabilities are limited to keyword matching

Wikipedia

- Simple questions, hard to answer
 - Which are the companies owned by Sony and located in India?
 - Which universities in India, located in Mumbai, were established before 1975?
 - Who is Shakespeare's child's spouse?

Sony Pictures Imageworks	
	SONY PICTURES imageworks
Type	Subsidiary of Sony Pictures Entertainment
Industry	CGI visual effects Motion pictures
Founded	1992
Headquarters	Culver City, California, USA
Number of locations	Albuquerque, New Mexico, US Novato, California, US Chennai, India Vancouver, British Columbia, Canada
Products	visual effects
Owner(s)	Sony
Parent	Sony Pictures Entertainment
Website	imageworks.com

DBpedia

- Tool which facilitates querying Wikipedia
- An effort to extract structured information from Wikipedia and to make this information available on the Web under an open license
- Project undertaken by the **Free University of Berlin** (Germany) and the **University of Leipzig** (Germany), in collaboration with OpenLink Software
- A uniform dataset which can be queried or linked to other data sets on the web

Structure in Wikipedia

Title

Other language versions

Categories

Geo-Coordinates

Infoboxes

The screenshot shows the Wikipedia article for Mumbai. Annotations include:

- Title:** Points to the main title "Mumbai".
- Other language versions:** Points to the "Languages" section in the left sidebar.
- Geo-Coordinates:** Points to the coordinate display "Coordinates: 18°58′30″N 72°49′33″E".
- Infoboxes:** Points to the information box on the right side of the article.
- Categories:** Points to the category list at the bottom of the article.

Categories: Mumbai | Cities and towns in Maharashtra | Populated coastal places in India | Former Portuguese colonies | Indian capital cities | Metropolitan cities in India | Port cities in India

Former name	Bombay
Country	India
State	Maharashtra
District(s)	Mumbai City Mumbai Suburban
Municipal commissioner	Subodh Kumar
Mayor	Shraddha Jadhav (SS)
Population	12,478,447 ⁽¹⁾⁽²⁾ (1st) (2011)
• Density	• 20,694 /km ² (53,597 /sq mi)
• Metro	• 18,414,288 ⁽³⁾ (1st) (2011)
Time zone	IST (UTC+05:30)
Area	603 km ² (233 sq mi)
• Elevation	• 14 metres (46 ft)
Codes	[show]
Website	www.mcgm.gov.in

DBpedia Extraction Framework

- Identifies structured information in Wikipedia and converts it into to RDF



```
{{Infobox Indian jurisdiction
|former_name = Bombay
|other_name = Bombay
|type = Metropolitan City
|type_2 = Finance Capital
|latd=18 |latm=58 |lats=30
|longd=72 |longm=49 |longs=33
|locator_position = right
|state_name = Maharashtra
|district = [[Mumbai City district|Mumbai City]]
<br />[[Mumbai Suburban District|Mumbai Suburban]]
|language=Marathi language
|leader_title_2 = [[Mayor of Mumbai|Mayor]]
|leader_name_2 = [[Shraddha Jadhav]] ([[Shiv Sena|SS]])
|leader_title = [[Municipal commissioner]]
|leader_name = Subodh Kumar
|altitude = 14
|population_total = 12,478,447
```



dbpedia-owl:areaCode	▪ 9122-XXXX XXXX
dbpedia-owl:areaTotal	▪ 1561762830.532608 (xsd:double)
dbpedia-owl:country	▪ dbpedia:India
dbpedia-owl:district	▪ dbpedia:Mumbai_Suburban_District ▪ dbpedia:Mumbai_City_district
dbpedia-owl:elevation	▪ 14.000000 (xsd:double)
dbpedia-owl:leaderTitle	▪ Municipal commissioner
dbpedia-owl:populationDensity	▪ 7989.998069 (xsd:double)
dbpedia-owl:populationMetro	▪ 21900967 (xsd:integer)
dbpedia-owl:populationTotal	▪ 12478447 (xsd:integer)
dbpedia-owl:postalCode	▪ 400 xxx
dbpedia-owl:state	▪ dbpedia:Maharashtra
dbpedia-owl:synonym	▪ Bombay






What is RDF?

- **R**esource **D**escription **F**ramework
- Knowledge representation language for describing resources
- DBpedia uses RDF for representing the extracted information
- A typical format of an RDF statement comprises of 3 components
 - **Subject** - Any entity like place, person, book etc.
 - **Predicate** – A characteristic of the subject
 - **Object**– A value for the subject

RDF – Example

- “Mumbai is the birth place of Rajiv Gandhi”
- The corresponding RDF statement would be :

<http://dbpedia.org/resource/Rajiv_Gandhi>		Subject
<http://dbpedia.org/property/birthPlace>		Predicate
<http://dbpedia.org/resource/Mumbai>.		Object
- Subject and Predicate are identified by URIs while the Object can be an URI or a value
- The URIs in the RDF statement can be shortened by using a “prefix”

RDF – Prefixes

- PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX dbp: <http://dbpedia.org/property/>

The previous RDF statement can be now represented as

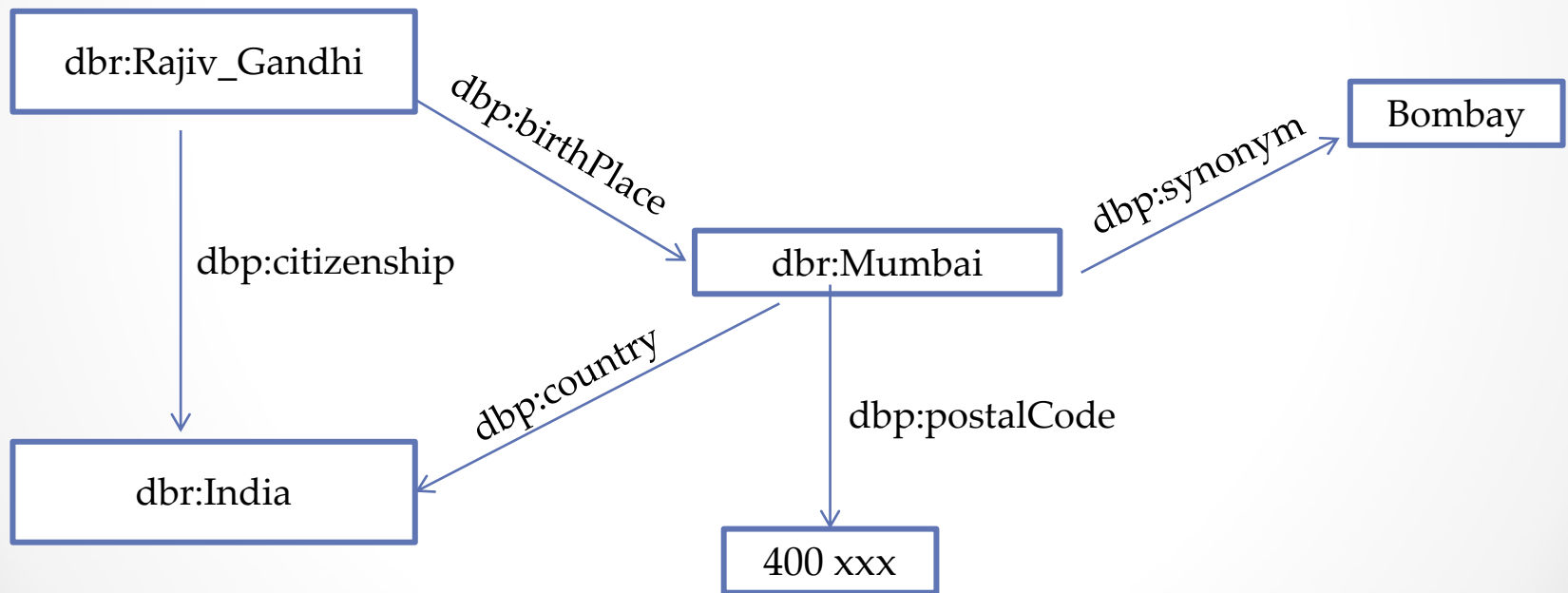
dbr:Rajiv_Gandhi dbp:birthPlace dbr:Mumbai.

The diagram illustrates the structure of the RDF statement "dbr:Rajiv_Gandhi dbp:birthPlace dbr:Mumbai.". Three blue curly braces are positioned below the text to group the components. The first brace is under "dbr:Rajiv_Gandhi", the second is under "dbp:birthPlace", and the third is under "dbr:Mumbai.". Below each brace is a label: "Subject" under the first, "Predicate" under the second, and "Object" under the third.

Subject Predicate Object

Why RDF?



- Structured knowledge base
- Relations between Resources can be formed by combining multiple RDF statements resulting in a “graph”



How to query this RDF graph?

- **Simple Protocol And RDF Query Language (SPARQL)**

- **Format**

Select <variable(s)>  Denoted by ? mark
Where <condition(s) . >  RDF statement (s)

- **Example**

```
PREFIX dbr: <http://dbpedia.org/resource/>  
PREFIX dbp: <http://dbpedia.org/property/>  
SELECT ?capital  
WHERE {  
  dbr:India    dbp:capital    ?capital .  
}
```

How to query DBpedia?

- The DBpedia dataset consists of over 1 billion pieces of RDF triples from Wikipedia
- The dataset can be accessed via
 1. SPARQL query endpoint [-http://dbpedia.org/snorql/](http://dbpedia.org/snorql/)
 2. SPARQL Linked Data Interface
Example - <http://dbpedia.org/page/India>
 3. Download the DBpedia dataset

Example

- **Universities in India, located in Mumbai, which were established before 1975:**

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

```
PREFIX dbr: <http://dbpedia.org/resource/>
```

```
PREFIX dbp: <http://dbpedia.org/property/>
```

```
PREFIX dbo: <http://dbpedia.org/ontology/>
```

```
SELECT ?institutes
```

```
WHERE {
```

```
    ?institutes dbp:country dbr:India .
```

```
    ?institutes dbp:city dbr:Mumbai .
```

```
    ?institutes rdf:type dbo:University .
```

```
    ?institutes dbp:established ?date.
```

```
    FILTER(?date < 1975)
```

```
}
```

Example

- **List of Hindi movies having run time greater than 3 hours and starring Anil Kapoor**

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

```
PREFIX dbr: <http://dbpedia.org/resource/>
```

```
PREFIX dbp: <http://dbpedia.org/property/>
```

```
PREFIX dbo: <http://dbpedia.org/ontology/>
```

```
SELECT ?movie, ?runtime
```

```
WHERE {
```

```
    ?movie rdf:type dbo:Film.
```

```
    ?movie dbo:runtime ?runtime.
```

```
    ?movie dbo:language dbr:Standard_Hindi.
```

```
    ?movie dbo:starring dbr:Anil_Kapoor.
```

```
    FILTER (?runtime > 10800)
```

```
}
```

Example

- **Information about Germany in German language**

```
PREFIX dbr: <http://dbpedia.org/resource/>
```

```
PREFIX dbp: <http://dbpedia.org/property/>
```

```
PREFIX dbo: <http://dbpedia.org/ontology/>
```

```
SELECT ?abstract
```

```
WHERE {
```

```
  dbr:Germany dbo:abstract ?abstract .
```

```
  FILTER langMatches( lang(?abstract), 'de')
```

```
}
```

DBpedia – Use Cases

1. Improving Wikipedia Search
2. DBpedia data for your web pages
3. Mobile and Geographic Applications
 - Example - [DBpedia Mobile](#)
4. Document Enhancement
 - [DBpedia Spotlight](#) (Annotation)
 - [Zemanta](#) (Blogging)
5. Support Wikipedia Authors
6. Other Applications
 - [RelFinder](#)

DBpedia - Improvements

- Synchronization with Wikipedia data
 - <http://live.dbpedia.org/LiveStats/>

Entity Statistics

Description	Total
Number of instances updated within last minute	83
Number of instances updated within last 5 minutes	76
Number of instances updated within last hour	4131
Number of instances updated within last day	132826
Number of instances since the start of the database	11762481

20 Most Recently Updated Entities

#	Time	Title	DBpedia	Wikipedia
1	19 seconds ago	The Judas Kiss	DBpedia	Wikipedia
2	20 seconds ago	Rebirth of the Temple	DBpedia	Wikipedia
3	20 seconds ago	'Abd al-Ilah	DBpedia	Wikipedia
4	21 seconds ago	Muddy Waters	DBpedia	Wikipedia
5	22 seconds ago	Jean-Claude Van Damme	DBpedia	Wikipedia
6	23 seconds ago	Sufism in Sindh	DBpedia	Wikipedia
7	23 seconds ago	Dashboard/Requested moves	DBpedia	Wikipedia
8	25 seconds ago	List of Diary of a Wimpy Kid characters	DBpedia	Wikipedia
9	25 seconds ago	Alfonso XIII of Spain	DBpedia	Wikipedia
10	26 seconds ago	Donald Bren School of Information and Computer Sciences	DBpedia	Wikipedia
11	27 seconds ago	Atticus Finch	DBpedia	Wikipedia
12	28 seconds ago	Duffy Conroy	DBpedia	Wikipedia
13	28 seconds ago	Hadise	DBpedia	Wikipedia

- Data Cleansing required
 - Properties like PlaceOfBirth and BirthPlace

Thank you!