

# Structured Data within the Web of Data

**Cristian LAI**

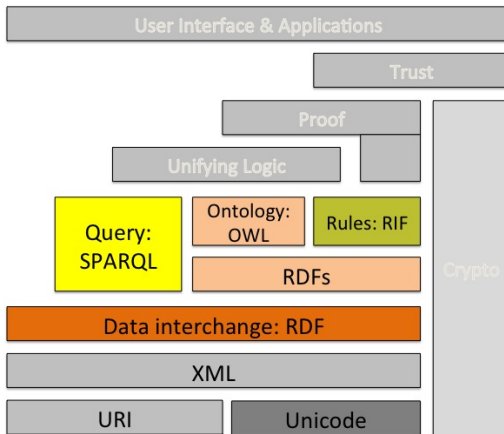
ISOC, NIT

# Outline

- Motivation
- UnStructured Data
- Structured Data
- Applications

# Context

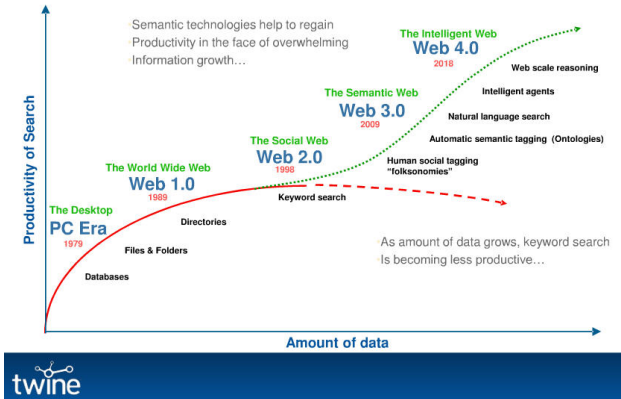
## Semantic Web



# Motivation

## Search on the Web

### The Future of Search



## Outline

- Motivation
- UnStructured Data
- Structured Data
- Applications

## Wikipedia

- Started in 2001.
- 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

**WIKIPEDIA**  
The free encyclopedia

Article Talk Read Edit View history Search

**Cagliari**

*"Cagliari" redirects here. For each genus, see *Carex* (disambiguation). For other uses, see *Cagliari* (disambiguation).*

**Cagliari** (Italian: [kadjali] as seen:  listen; Sardinian: Càrrali; Latin: Carralis) is the capital of the island of Sardinia, a region of Italy. Cagliari's Sardinian name *Càrralis* literally means castle. It has about 158,000 inhabitants, or about 400,000 including the outlying townships (metropolitan area). Ethnically, Sardinians, Capoterra, Serrargius, Senis, Monreale, Quartucciu, Quartu Serteddu.

An ancient city with a long history, Cagliari has seen the rule of several civilisations. It was the capital of the Kingdom of Sardinia which in the 1861 became the Kingdom of Italy from 1924 to 1930 and from 1946 to 1962. Seat of the important University of Cagliari and the Primate Roman Catholic archdiocese of Sardinia,<sup>[a]</sup> the city is an important regional cultural, educational, political and artistic centre. Known for its diverse art, historical, architectural and several monuments,<sup>[b]</sup> it is also Sardinia's economic and industrial hub, being one of the biggest ports in the Mediterranean sea, an international airport, and the 28th-highest income rate in Italy, comparable to several Northern cities, such as Turin, Florence and Genoa.<sup>[c]</sup>

**Contents** (hide)

- History
  - Early history
  - Division of Cagliari
  - 11th to 15th century
  - 14th to 17th centuries
  - 18th century
  - Modern age
  - Projects for the future
- Geography
  - Climate
  - Urban plans
- Economy
- Transport
- Flags
- Culture
- Religion
- Consulates
- International relations
- Twin towns – Sister cities
- Image gallery
- References
- Notes

**Location of Cagliari in Italy**  
**Coordinates: 39°14′47″N 09°03′27″E﻿ / ﻿39.24639°N 9.05750°E﻿ / 39.24639; 9.05750**

<b>Country</b>	<span><span><span></span></span><span> </span></span> Italy
<b>Region</b>	Sardinia
<b>Province</b>	Cagliari (CA)
<b>President</b>	Italo
<b>Government</b>	
<span> </span> - <b>Mayor</b>	Massimo Zedda (SDI)
<b>Area</b>	
<span> </span> - <b>Total</b>	85.45 <span> </span> km <sup>2</sup> (32.98 <span> </span> sq <span> </span> mi)
<b>Elevation</b>	4 <span> </span> m (13 <span> </span> ft)
<b>Population</b> (30 November 2010)	

# Structured Information in Wikipedia

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

The screenshot shows the Wikipedia article for 'Cagliari'. A red box highlights the 'Languages' section on the right, which lists various language versions of the article. Another red box highlights the 'Contents' table in the middle, which provides a structured overview of the article's sections. A third red box highlights the 'Location of Cagliari' section at the bottom, which includes the coordinates 39°14'47''.

**Languages**

- Afrikaans
- العربية
- aragonés
- Armãneashtce
- беларуская
- български
- bosanski
- brezhoneg
- catàla
- češky
- corsu
- Cymraeg
- dansk
- Deutsch
- deutch
- Ελληνικά
- español
- Espéranto
- euskara
- فارسی
- français
- Gaeilge
- galego
- 한국어
- hrvatski
- Ido
- Bahasa Indonesia
- interlingua
- Ирон
- italiano
- עברית
- ქართული
- Kiswahili
- Ladino
- Latina

**Contents (hide)**

- 1 History
  - 1.1 Early history
  - 1.2 Division of Cagliari
  - 1.3 11th to 15th century
  - 1.4 16th to 17th centuries
  - 1.5 18th century
  - 1.6 Modern age
  - 1.7 Projects for the future
- 2 Geography
  - 2.1 Climate
  - 2.2 Urban plans
- 3 Location of Cagliari
- Coordinates: 39°14′47″ ﻿ / ﻿﻿ /
- 4 Economy
- 5 Transport
- 6 Sport
- 7 Culture
- 8 Nightlife
- 10 Constitutes
- 11 International relations
- 11.1 Twin towns – Sister cities
- 12 Image gallery
- 13 References

**Location of Cagliari**  
Coordinates: 39°14′47″ ﻿ / ﻿﻿ /

# Structured Information in Wikipedia

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

The image displays two versions of the Wikipedia article for Cagliari, Italy, illustrating the structured information in the infobox template.

**Left Screenshot (Full Article):** Shows the top of the article. The infobox is highlighted with a red box. It contains:

- Title: Cagliari
- Image gallery: Four images showing the city and its landmarks.
- Text: "A heritage showing different parts and features of the city of Cagliari"
- Flag and Coat of arms
- Map of Italy with Cagliari highlighted.
- Coordinates: 39°14′47″N 09°10′27″E﻿ / ﻿39.24639°N 9.17417°E﻿ / 39.24639; 9.17417
- Country: Italy
- Region: Sardinia
- Province: Cagliari (CA)
- Mayor: Massimo Zedda (DLC)
- Area: Total 80.45 km<sup>2</sup> (30.99 sq mi)
- Population (30 November 2010):
  - Total: 150,340
  - Density: 1,869/km<sup>2</sup> (4,869/sq mi)
- Time zone: CET (UTC +1)
- Summer (DST): CEST (UTC +2)
- Postal code: 09100
- Calling code: 070
- Patron saint: St. Saturninus
- Saint day: October 30
- Website: Official website

**Right Screenshot (Zoomed Infobox):** Shows a zoomed-in view of the infobox template. It contains:

- Title: Cagliari
- Image gallery: Four images showing the city and its landmarks.
- Text: "A heritage showing different parts and features of the city of Cagliari"
- Flag and Coat of arms
- Map of Italy with Cagliari highlighted.
- Coordinates: 39°14′47″N 09°10′27″E﻿ / ﻿39.24639°N 9.17417°E﻿ / 39.24639; 9.17417
- Country: Italy
- Region: Sardinia
- Province: Cagliari (CA)
- Mayor: Massimo Zedda (DLC)
- Area:
  - Total: 80.45 km<sup>2</sup> (30.99 sq mi)
- Population (30 November 2010):
  - Total: 150,340
  - Density: 1,869/km<sup>2</sup> (4,869/sq mi)
- Time zone: CET (UTC +1)
- Summer (DST): CEST (UTC +2)
- Postal code: 09100
- Calling code: 070
- Patron saint: St. Saturninus
- Saint day: October 30
- Website: Official website

## Outline

- Motivation
- UnStructured Data
- Structured Data
- Applications

## RDF representation Knowledge Base

```
dbp:Cagliari rdf:type dbp:City  
dbp:Cagliari dbp>Title "Cagliari"  
dbp:Cagliari dbp:Country dbp:Italy  
dbp:Cagliari dbp:postalCode 09100  
dbp:Cagliari geo:lat "39.246387"xsd:float  
dbp:Cagliari geo:long "9.057500"xsd:float  
dbp:Cagliari rdf:type yago:MediterraneanPortCitiesAndTownsInItaly
```

...

- An environment for collecting and structuring data.
- Well defined structure of classification.

## RDF

- Triples: (*subject, predicate, object*)
- *Subject and object*
  - are both URIs that each identify a resource, or a URI and a string literal respectively.
  -
- Predicate
  - specifies how the subject and object are related, and is also represented by a URI.
- For example:
  - *A knows B*
  - *C 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

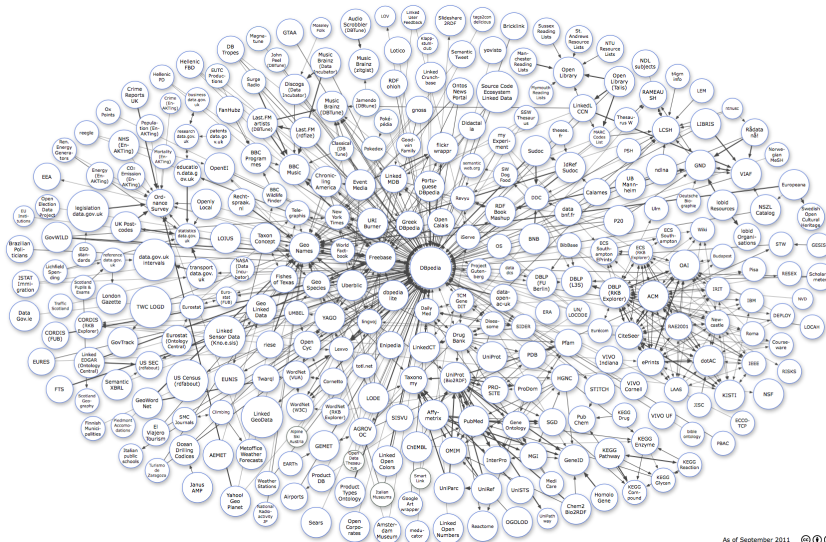
- Started in 2007.
- Is the result of a community effort to extract structured information from Wikipedia.
- Makes Wikipedia data available as RDF.
- Results: **The DBpedia Data Set**
  - describes 3.64 million "*things*" with over half a billion "*facts*" (July 2011), 364k persons, 462k places, 99k 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>



## 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:
  - ...
  - *Abstracts of movies starring Tom Cruise, released before 1999.*
  - *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

SPARQL

```
SELECT ?subject ?label ?released ?abstract WHERE {
  ?subject rdf:type <http://dbpedia.org/ontology/Film>.
  ?subject dbpedia2:starring <http://dbpedia.org/resource/Tom_Cruise>.
  ?subject rdfs:comment ?abstract.
  ?subject rdfs:label ?label.

  FILTER(lang(?abstract) = "en" && lang(?label) = "en").

  ?subject <http://dbpedia.org/ontology/releaseDate> ?released.

  FILTER(xsd:date(?released) < "2000-01-01"^^xsd:date).
} ORDER BY ?released
```

subject	label	released	abstract
<a href="#">:Losin%27_It</a>	"Losin' It"@en	"1983-04-08"^^xsd:date	"Losin' It is a 1983 comedy film starring Tom Cruise, Shelley Long, Jackie Earle Haley, and John Stockwell. The film is directed by Curtis Hanson. It was filmed largely in Calexico, California."@en
<a href="#">:Risky_Business</a>	"Risky Business"@en	"1983-08-05"^^xsd:date	"Risky Business is a 1983 American teen comedy-drama film written by Paul Brickman in his directorial debut. It stars Tom Cruise and Rebecca De Mornay. The hit film launched Cruise to stardom."@en
<a href="#">:All_the_Right_Moves</a>	"All the Right Moves"@en	"1983-10-21"^^xsd:date	"For the OneRepublic song, see All the Right Moves (song) All the Right Moves is a 1983 drama film directed by Michael Chapman and starring Tom Cruise, Craig T. Nelson, Lea Thompson, Chris Penn, and Gary Graham. It was filmed on location during WPIAL football season in Johnstown, Pennsylvania, and Pittsburgh."@en
<a href="#">:Legend_%2Bfilm%29</a>	"Legend (film)"@en	"1985-12-13"^^xsd:date	"Legend is a 1985 fantasy film released by Universal Pictures, directed by Ridley Scott and starring Tom Cruise, Mia Sara, and Tim Curry. Though not a very notable success when first released, it received an Academy Award nomination (for Best Makeup) and has since developed a cult following."@en
<a href="#">:The_Color_of_Money</a>	"The Color of Money"@en	"1986-10-17"^^xsd:date	"The Color of Money is a 1986 film based on a 1984 novel of the same name by American writer Walter Tevis. The film continues the story of pool hustler and stakehorse Edward "Fast Eddie" Felson from Tevis' first novel, The Hustler (1959), with Paul Newman reprising his role from its film adaptation (1961)."@en

## Outline

- Motivation
- UnStructured Data
- Structured Data
- Applications

## Linked Data Search Engines and Indexes

- 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.

Tom Heath and Christian Bizer (2011) *Linked Data: Evolving the Web into a Global Data Space (1st edition)*. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1:1, 1-136. Morgan & Claypool.

- 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](http://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?...](http://www.starwoodhotels.com/westin/property/overview/index.html?...)

Experience a Legend. Since 1906, 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/](http://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-d203080-Reviews-Th...](http://www.tripadvisor.com/Hotel_Review-g187791-d203080-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** ...

## Westin Excelsior, Rome

[www.westin.com/excelsiorrome](http://www.westin.com/excelsiorrome)

Zagat: **25** / 30 - 16 Google reviews - \$379 ▼



Via V  
06 47

Check in:

07/26

31

Check out:

07/27

31

## The Wes

[www.starw](http://www.starw)

Experienc

statesmen

+ Show m

↳ Rooms

## Hotel Th

[excelsior](http://excelsior)

Hotel The

refurbished

## The Wes

[www.tripad](http://www.tripad)

★★★★☆

The Westin **Excelsior Rome, Rome:** See 678

Ads

per night

Travelocity

\$410 incl taxes & fees

Priceline

\$419 incl taxes & fees

getaroom

\$380

Booking.com

\$425 incl taxes & fees

Currency disclaimer

[www.westin.com](http://www.westin.com)

Owner site



# Twitter, #annotations

Semantic annotations for Twitter

localhost:8080/SemanticTwitterConsumer/main.jsp

Google

Most Visited | Getting Started | Overview (Java ... | Latest Headlines | Apple | Yahoo! | Google Maps | YouTube | Wikipedia | News | Popular | Bookmarks


Reports | Navigation | Text Equivalents | Scripting | Style | Validators | Tools | Keyboard | Options

## Semantic #annotations for Twitter

[Logout](#)

crlai

Cristian Lai



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

Type your tweet with semantic annotations

Keyword

Message

44 Characters  
 96 Characters Left  
 7 Words

### Twitter Feed

[TEST] Usain Bolt [#dbpedia:Usain\\_Bolt](#) vs Tayson Gay [#dbpedia:Olympic\\_Games](#)  
16 hours ago

---

[TEST] Usain Bolt [#dbpedia:Usain\\_Bolt](#) vs Tayson Gay  
16 hours ago

---

[TEST] Google [#dbpedia:Google](#)  
16 hours ago

---

[#dbpedia:United\\_States](#)  
19 days ago

---

RT @DART2012ws: DART2012 <http://t.co/mY6gvZt3>. Deadline EXTENDED, June 29.  
21 days ago

## Q & A