

Un framework basato sui Linked Data per l'arricchimento semantico dei Dati Aperti sugli eventi

Antonino Lo Bue,
Independent researcher

Alberto Machì
ICAR-CNR, Palermo, Italy

Davide Taibi
ITD-CNR, Palermo, Italy



Consiglio Nazionale delle Ricerche



Problem

- Open Data Silos
 - Islands of information not interoperable
- Semantic Gap
 - Application specific data structures
 - Data is not machine understandable
- Data is incomplete or wrong

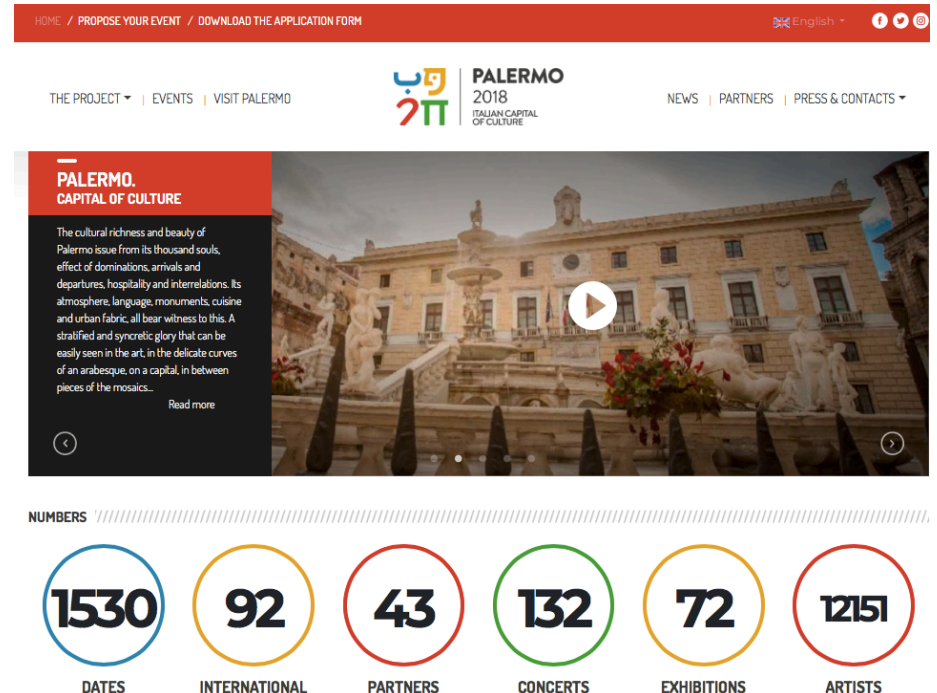
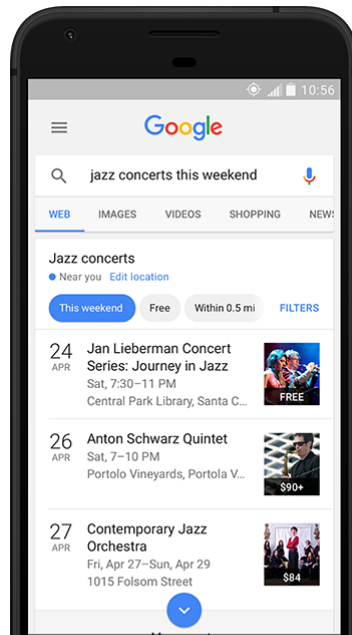
Objectives

- Applying Semantic Web principles to a real use case
- Reusing data and shared knowledge of LOD
- Information enrichments
- Know what, where, when? And what else?
- Increase awareness in local stakeholders

A case study



PALERMO
2018
ITALIAN CAPITAL
OF CULTURE



Open Data ≠ Good Data

- Data is not always good as we expect

```
"id":1,"orario":"9,30 -18,00",
"luogo":"ZAC",
"titolo":"MOSTRA SHAY FRISCH",
"ente":"Spazio Zac",
"sottocategoria":"MOSTRA",
"body":"<p style=\"margin: 0cm; margin-bottom: .0001pt;\"><span style=\"color: #222222; font-family:
Arial, sans-serif;\"><span style=\"font-size: 13.3333px;\">Visto il grande successo e l&rsquo;affluenza
di pubblico alla mostra dell&rsquo;artista israeliano <strong>Shay Frisch</strong>, in corso allo spazio
Zac dei Cantieri Culturali alla Zisa (via Paolo Gili 4), <strong>&ldquo;CAMPO 121745
B\\N&rdquo;</strong> ...</p>",
"slug":"mostra-shay-frisch",
"image":"eventi\\February2018\\shay-frish-1-1200x800.jpg",
"meta_description":"Proroga della mostra dell'artista israeliano allo Spazio Zac dei Cantieri Culturali
della Zisa",
"meta_keywords":"arte, campi elettrici, cantieri culturali della Zisa, palermo capitale cultura",
"data_inizio":"2018-01-29",
"data_fine":"2018-02-18",
"created_at":null,
"updated_at":"2018-02-06 00:13:20",
"seo_title":"Shay Frisch ai Cantieri culturali della Zisa",
"status":"PUBLISHED",
"categorie_evento":[{"id":9,"name":"mostre","slug":"mostre","created_at":"2018-01-26
23:42:11","updated_at":"2018-01-26 23:42:11","hex":"#66267B",
"pivot":{"evento_id":1,"categoria_evento_id":9,"created_at":"2018-01-29 03:12:43","updated_at":"2018-01-
29 03:12:43"}}]}
```

Open Data ≠ Good Data

- Data is not always good as we expect

Issues

Missing place information

```
"id":1,"orario":"9,30 -18,00",  
"luogo":"ZAC",  
"titolo":"MOSTRA SHAY FRISCH",  
"ente":"Spazio Zac",  
"sottocategoria":"MOSTRA",
```

Issues

Missing images

```
"image":"eventi/February2018/shay-frish-1-1200x800.jpg",  
"meta_description":"Proroga della mostra dell'  
della Zisa",  
"meta_keywords":"arte, campi elettrici, cantie  
"data_inizio":"2018-01-29",  
"data_fine":"2018-02-18",  
"created_at":null,  
"updated_at":"2018-02-06 00:13:20",
```

Issues

Missing description

```
"meta_title":"Mostre Shay Frisch e Zisa",
```

Issues

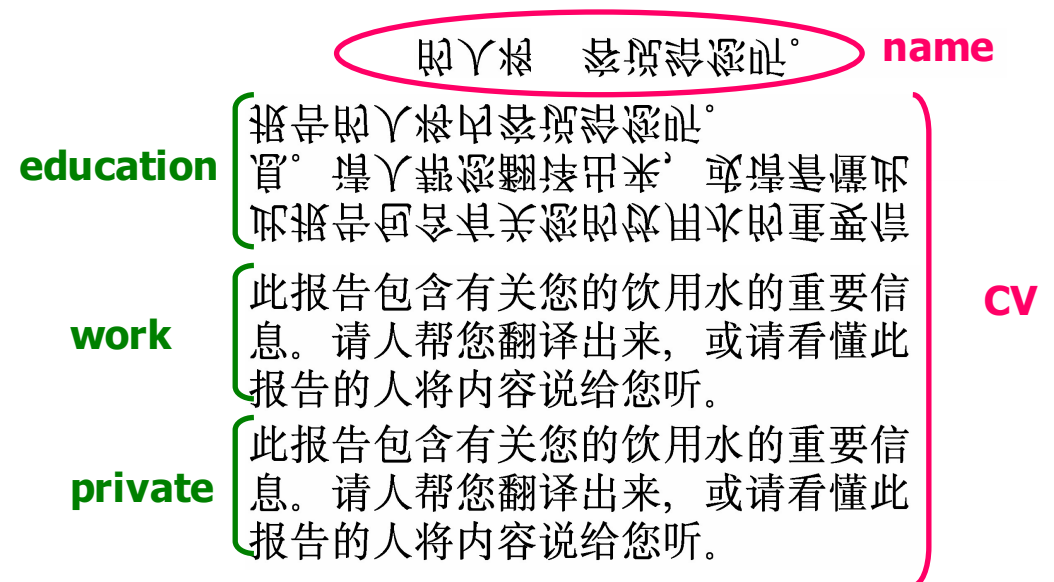
Missing geo coordinates

```
ot;\"><span style=\"color: #222222; font-family:  
px;\">>Visto il grande successo e l&rsquo;affluenza  
<strong>Shay Frisch</strong>, in corso allo spazio  
(4), <strong>&ldquo;CAMPO 121745
```

```
nostre", "created_at":"2018-01-26  
66267B",  
ed_at":"2018-01-29 03:12:43", "updated_at":"2018-01-
```

Semantic Web

- Express information in triples (subject, predicate, object)
- Describe univocally media contents
- Share concepts of a domain through ontologies
- Fill the “Semantic Gap”

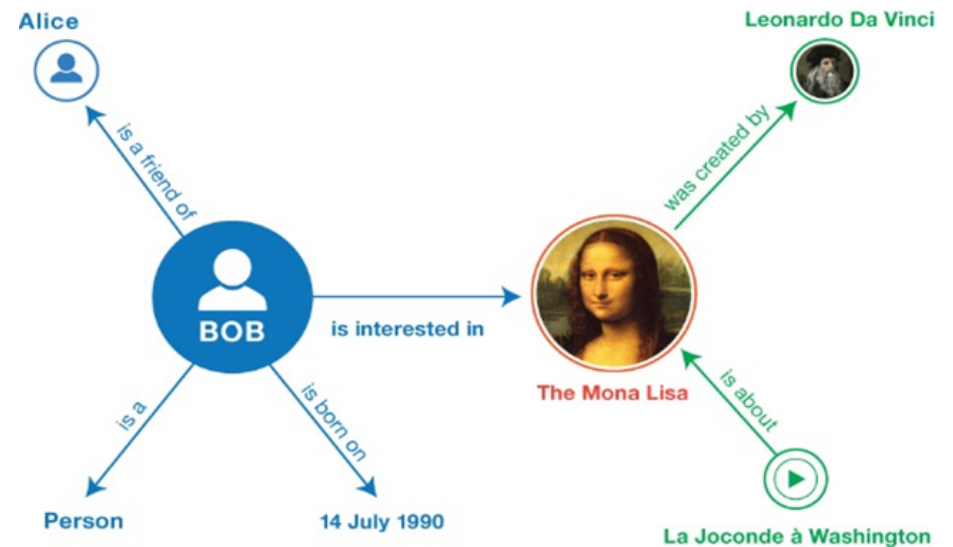
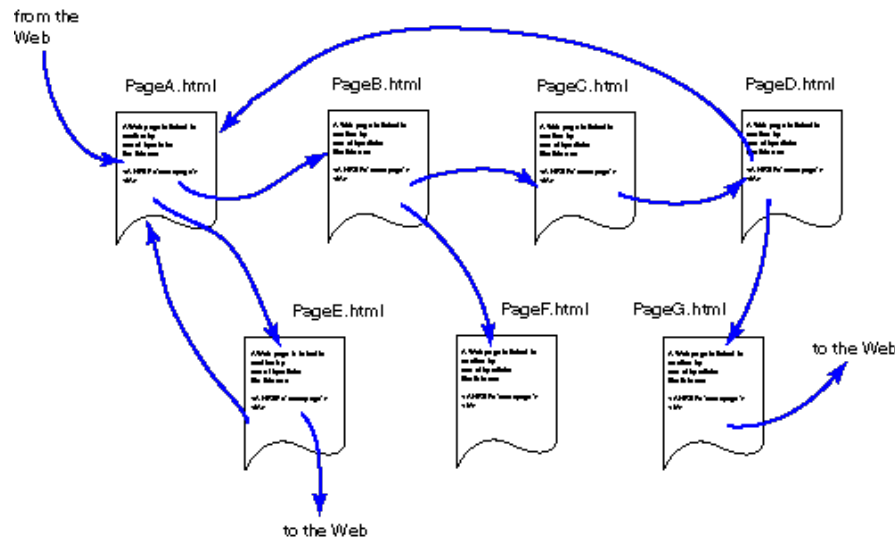


Semantic Web of Linked Data

- Hyperlinks ➡ Web

- Dataset interlinking ➡ Semantic Web

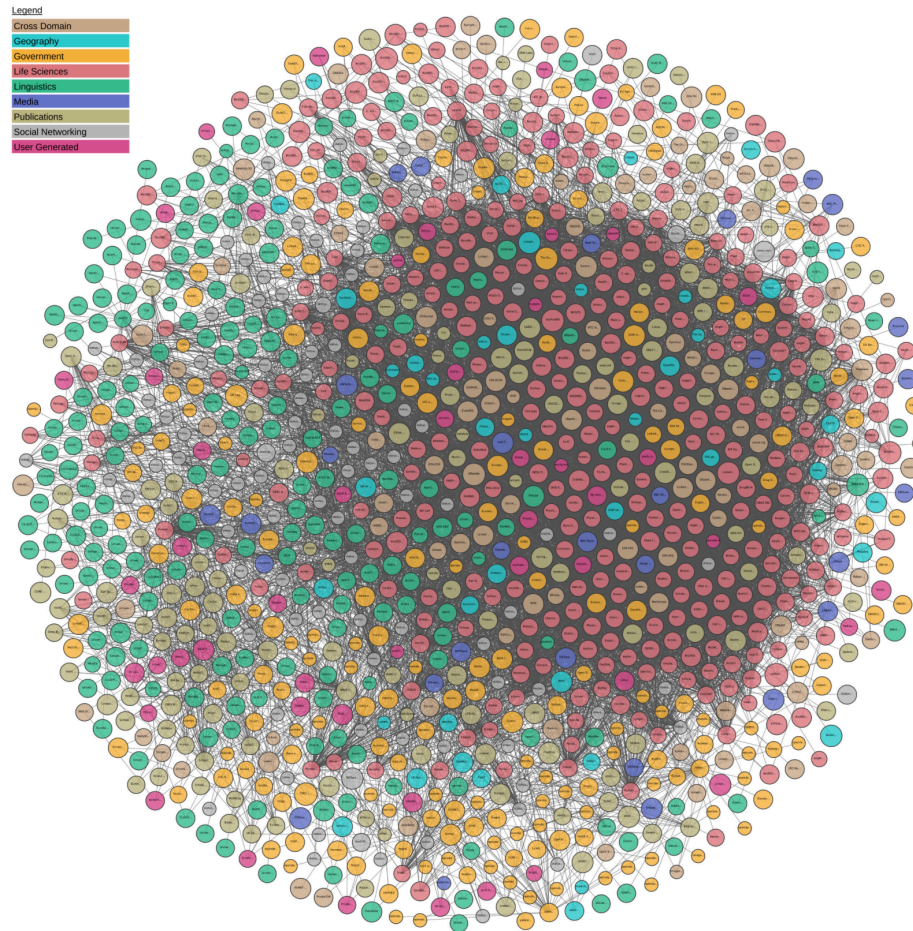
- “The **Semantic Web** *isn't* just about **putting data** on the web. It is about **making links**, so that a **person** or **machine** can **explore the web of data**. ” Tim Berners-Lee



Linked Open Data cloud

- How much knowledge is already shared and linked?

<http://lod-cloud.net>



How to publish good Linked Data?

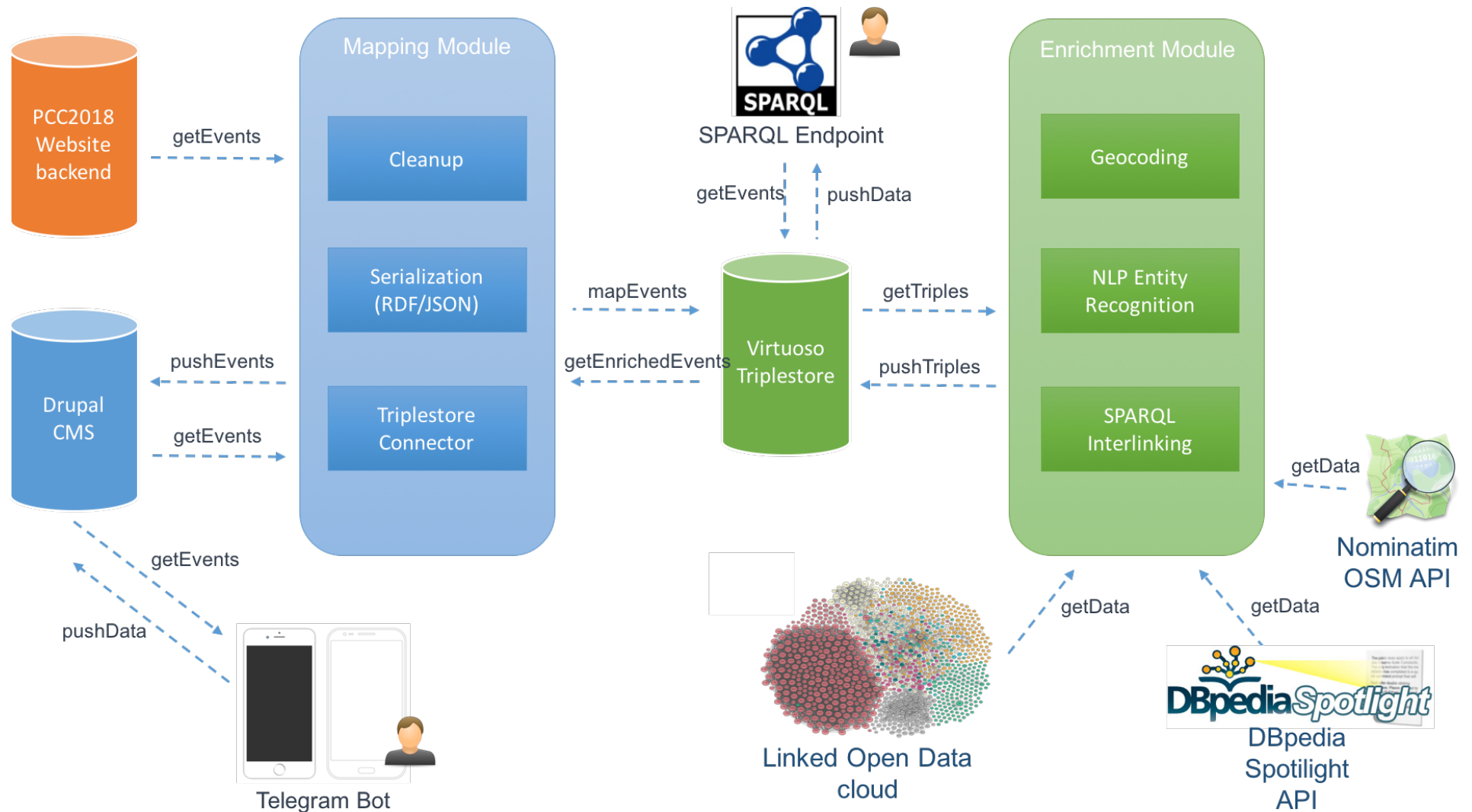
AgID guidelines to produce Linked (Open) Data

- *Dataset selection*
- *Cleaning*
- *Model creation*
- *Enrichment*
- *Interlinking*
- *Validation*
- *Publication*

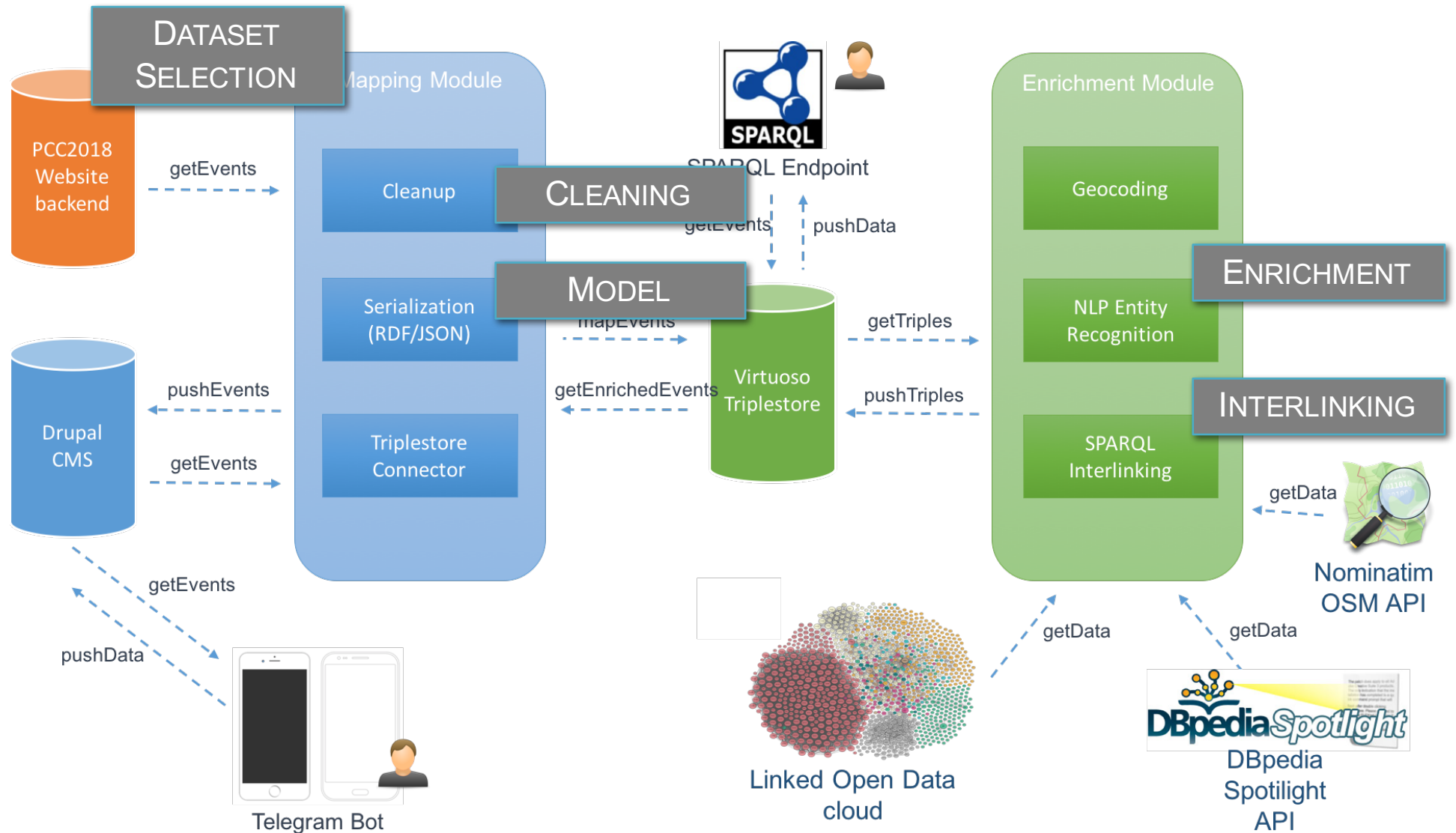


Source: Agenzia per l'Italia Digitale - Digital Italy Agency

Overall Architecture



Overall Architecture



Data Model: schema.org

- Types and properties defined in the *schema.org* vocabulary, are used by search engines to extract structured data
- Better ranking (SEO)
- Structured information for users

Barack Obama

44th U.S. President



 barackobama.com

Barack Hussein Obama II is an American politician who served as the 44th President of the United States from January 20, 2009, to January 20, 2017. A member of the Democratic Party, he was the first African American to be elected to the presidency and previously served as a United States Senator from Illinois. [Wikipedia](#)

Party: [Democratic Party](#) Trending

Born: August 4, 1961 (age 57 years), [Kapi'olani Medical Center for Women and Children, Honolulu, Hawaii, United States](#)

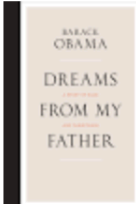
Height: 1.85 m

Vice president: [Joe Biden](#) (2009–2017)

Education: [Harvard Law School](#) (1988–1991), [MORE](#)

Books

View 35+ more



Dreams From My Father
1995



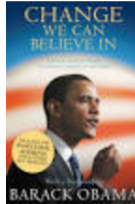
The Audacity of Hope: Th...
2006



Of Thee I Sing
2010

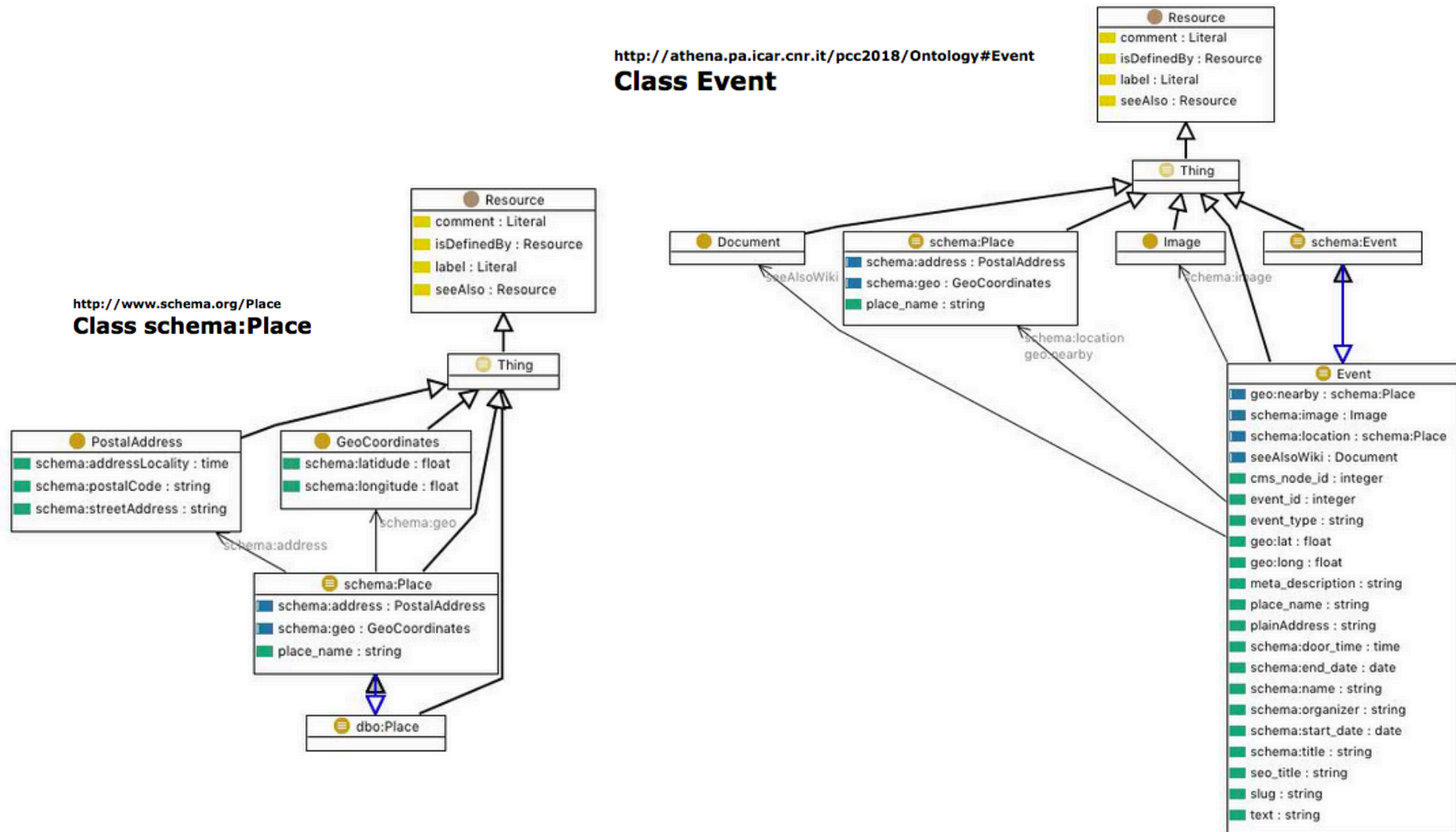


Barack Obama
2011



Change We Can Believe In
2008

Data Model: ontology

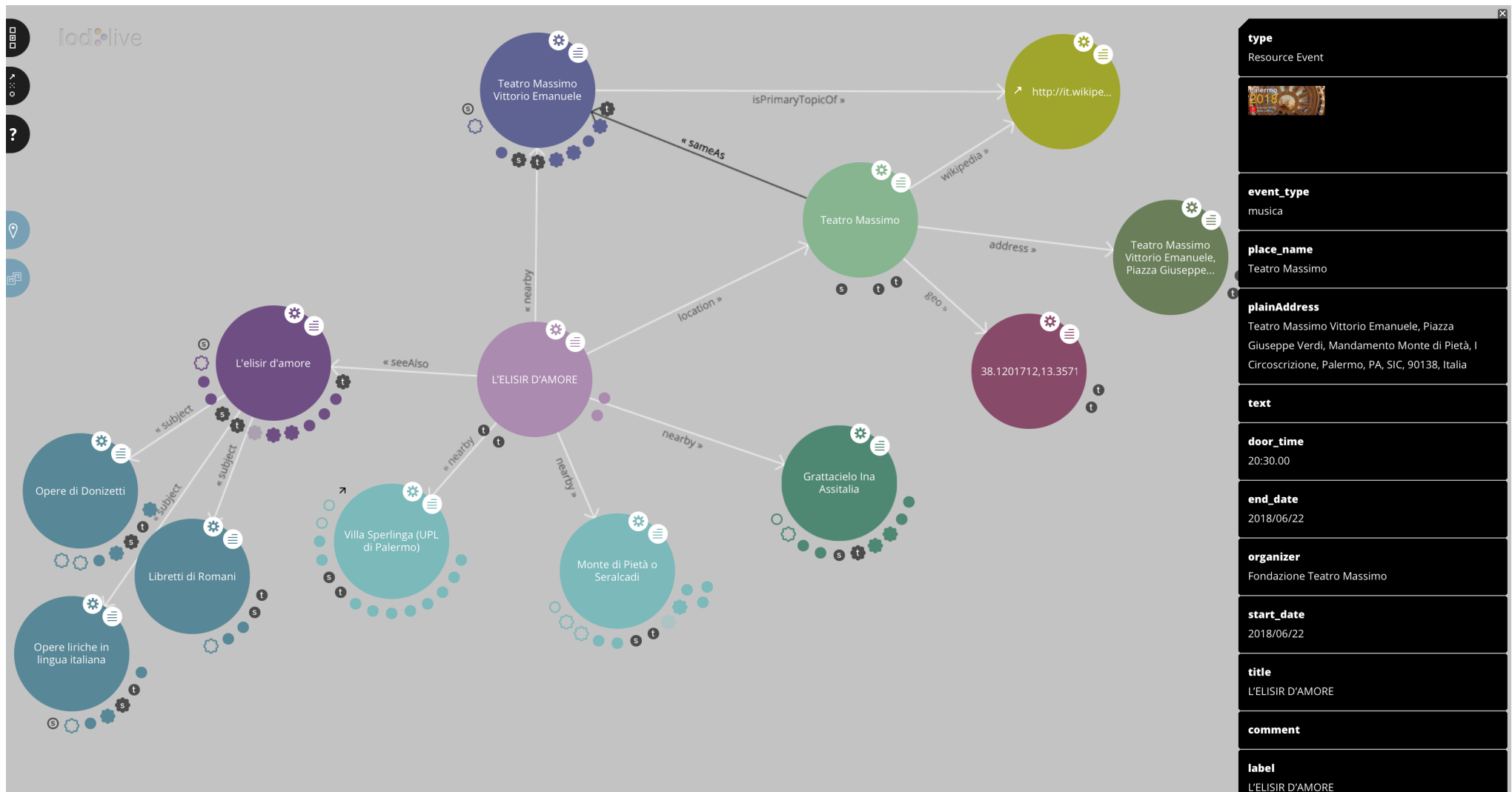


Semantic Enrichment and Interlinking

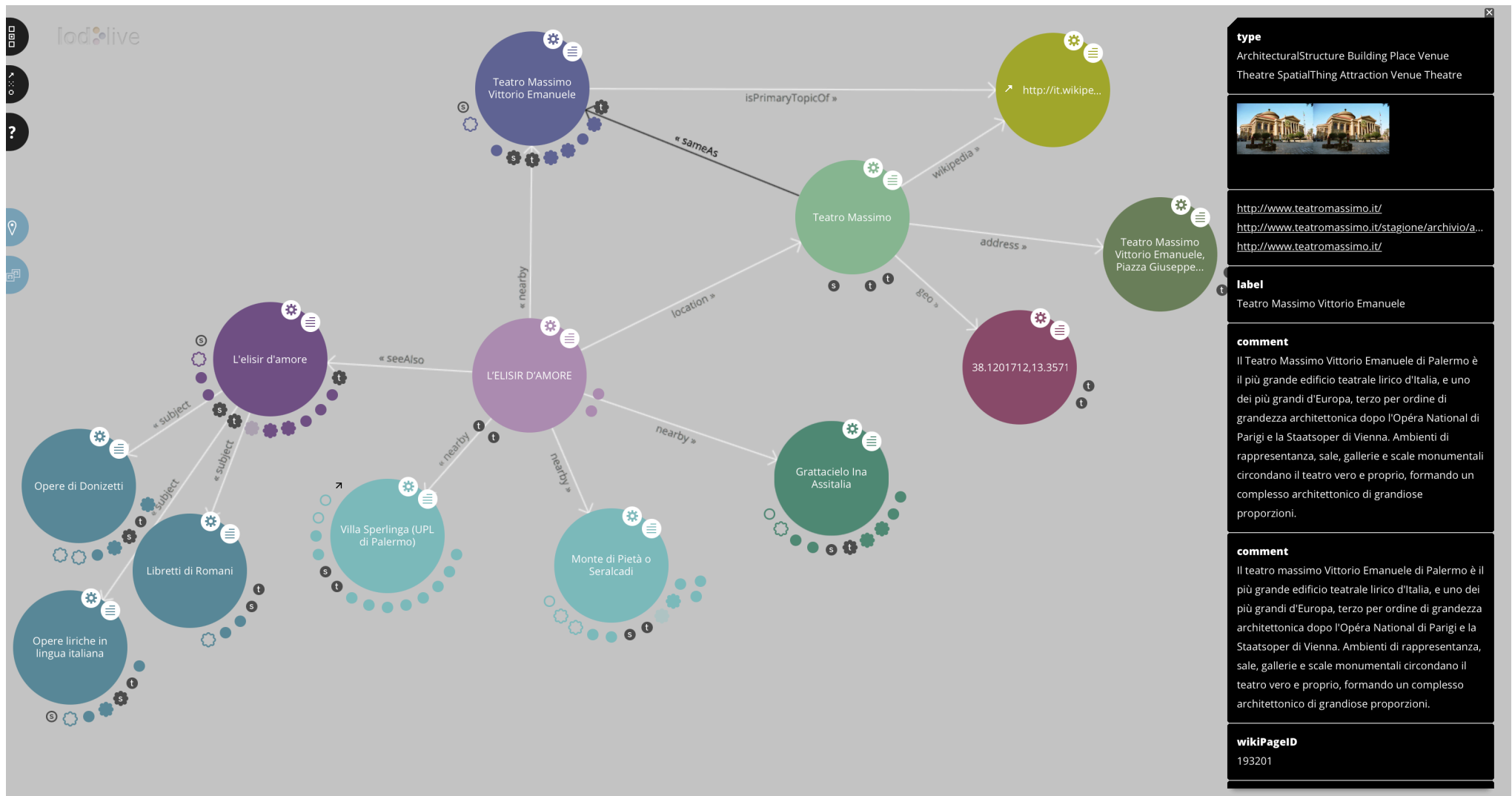
- Semantic named-entity recognition
 - Using DBpedia Linked Data as vocabulary for the named entities
 - Infer concept and knowledge related to the Event description
- Geocoding enrichment
 - Using Nominatim Open Street Map API,
 - Add the geolocation data to the Event plus Wikidata and Wikipedia references about the place (when present)
- Spatial interlinking
 - Using a federated query on the pcc2018 dataset and the Italian DBpedia dataset
 - Infer via spatial distance (Haversine) point of Interests near the Event location
- Information expansion
 - Using federated query add Place description and images

Semantic Enrichment and Interlinking

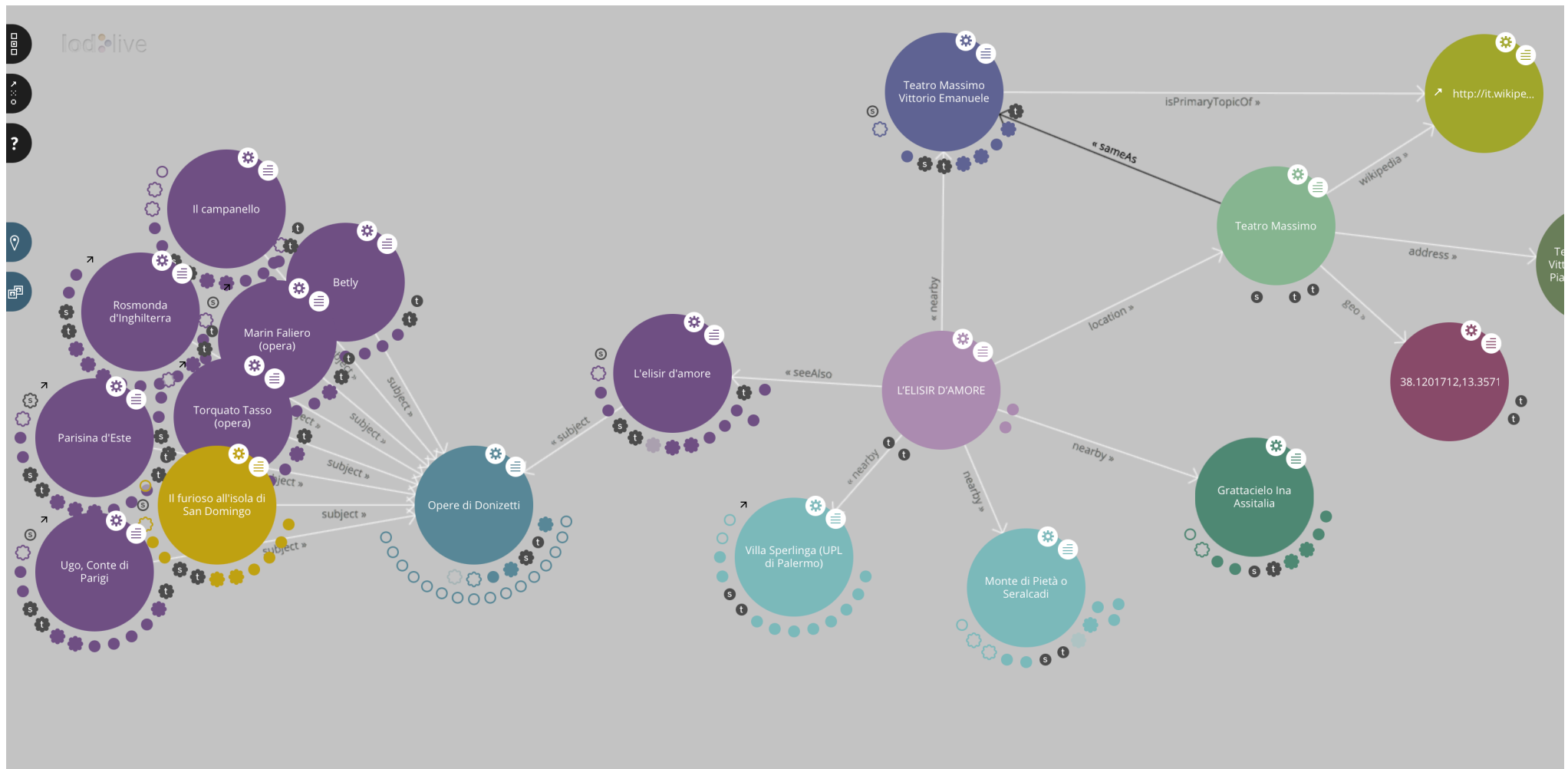
<http://athena.pa.icar.cnr.it:8080/lodview/Event/471.html>



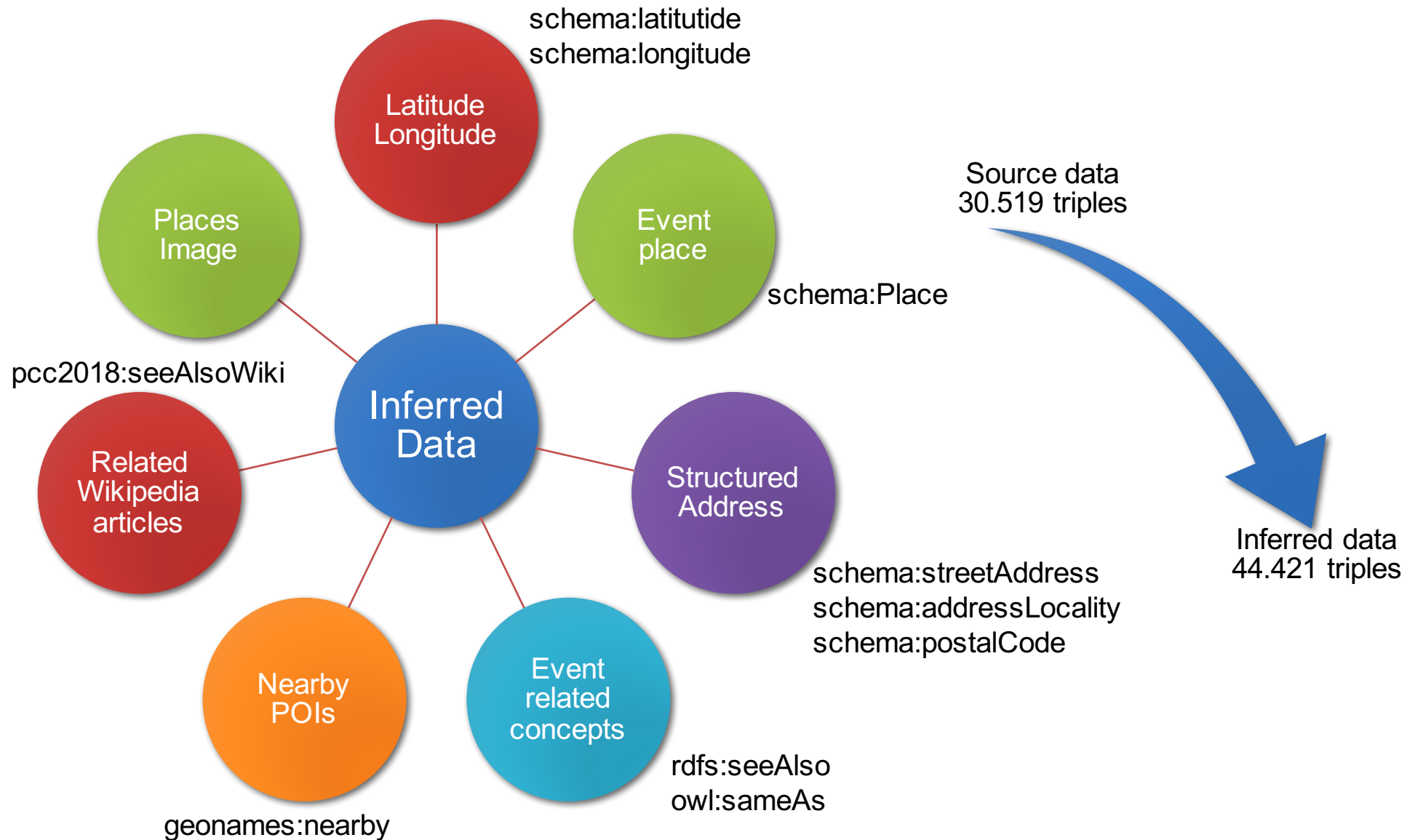
Semantic Enrichment and Interlinking



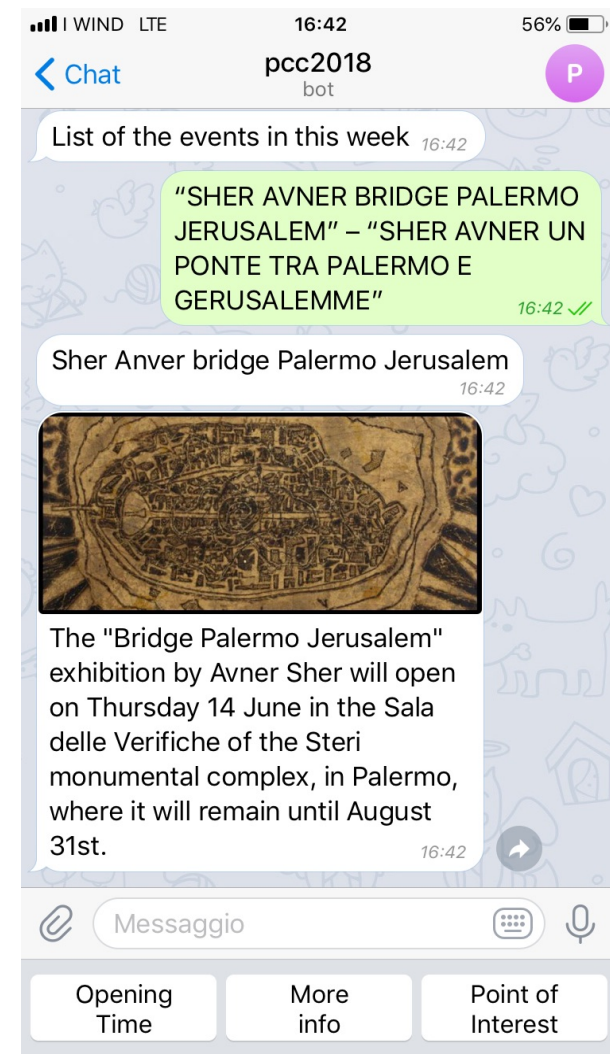
Semantic Enrichment and Interlinking



Semantic Enrichment and Interlinking



pcc2018 Bot Telegram



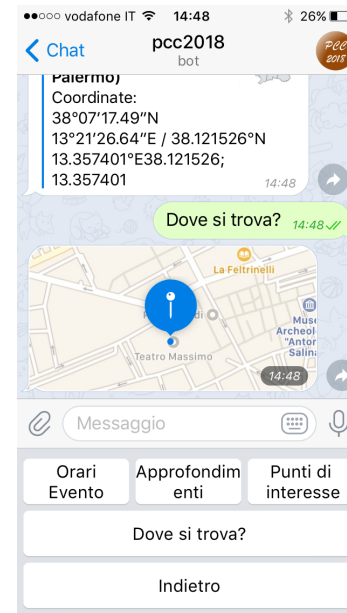
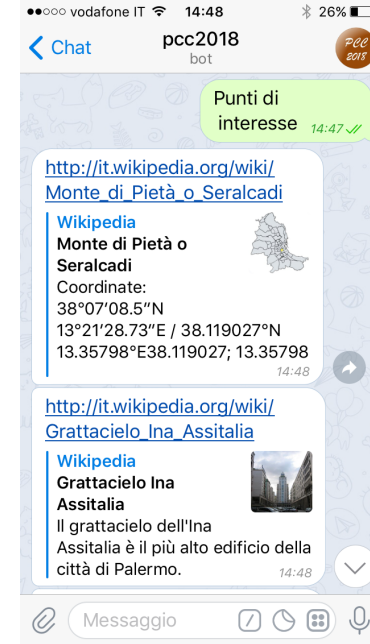
pcc2018 Bot Telegram



see also

nearby places

place information



Issues

- Geocoding limitations
 - Nominatim 719 over 1792
 - Google 1792 over 1792
- Data validation
 - Wrong geolocation
 - Named Entity Recognition issues
- Source data limitations
 - Char encoding
 - All capitalised data
 - Wrong fields used
 - Duplicates
 - ...
- Licenses?

MOSTRA SHAY FRISCH

<http://athena.pa.icar.cnr.it/pcc2018/Event/1>

rdfs:comment

Visto il grande successo e l'affluenza di pubblico alla mostra dell'artista Cantieri Culturali alla Zisa (via Paolo Gili 4), "CAMPO 121745 B/N" e' stata consiste nella modellazione e la manipolazione dell'energia attraverso l' da un assemblaggio di adattatori elettrici di uso comune, attraverso i qu imbrigliata all'interno di ripetizioni modulari e trasformata dall'elettricit prevede una sequenza di installazioni a pavimento e sulla grande parete martedì' a domenica 9,30-18,30. Ingresso libero

Visto il grande successo e l'affluenza di pubblico alla mostra dell'artista Cantieri Culturali alla Zisa (via Paolo Gili 4), "CAMPO 121745 B/N" e' stata consiste nella modellazione e la manipolazione dell'energia attraverso l' da un assemblaggio di adattatori elettrici di uso comune, attraverso i qu imbrigliata all'interno di ripetizioni modulari e trasformata dall'elettricit prevede una sequenza di installazioni a pavimento e sulla grande parete martedì' a domenica 9,30-18,30. Ingresso libero



rdfs:label

MOSTRA SHAY FRISCH

MOSTRA SHAY FRISCH

Conclusions

- Plain metadata (and Open Data) can be powered by Linked Data and expressed through semantic vocabularies to become **rich-machine understandable data** used by automatic bots to improve user experiences
- A case study with more than **1500 events** in the framework of the initiative *“Palermo Italian Capital of Culture 2018”* enriched
- A compact semantic search GUI was provided through a **Telegram bot** to allow users to easily search and share information

Future work

- Enhancement of the enrichment inference engine
- Content validation rules
 - Shapes Constraint Language (SHACL)
- Serialization of the inferred data as microdata/JSON-LD for website enhancement and SEO
- Upgrade of the bot to a conversational bot (chat-bot)

Grazie

*A framework to enrich events Open Data
producing Linked Open Data*

Ontology Documentation

<http://athena.pa.icar.cnr.it/htdocs/PCC2018/ontology>

SPARQL Endpoint

<http://kossyra.pa.icar.cnr.it:8890/sparql>

Telegram BOT

pcc2018

Antonino Lo Bue

@ninokeys

linkedin.com/in/antoninolobue



Example

Source data	Inferred data
<i>Pausa Sismica, Cinquantanni dal terremoto del Belice</i> (schema:title)	<http://it.dbpedia.org/resource/Gibellina> <http://it.dbpedia.org/resource/Ludovico_Corrao> <http://it.dbpedia.org/resource/Terremoto_del_Belice_del_1968> (rdfs:seeAlso)
<i>Nell'anniversario delle scosse, la mostra Pausa Sismica1968/2018 Cinquant'anni dal terremoto del Belice. Vicende e visioni alla Fondazione Sant Elia, (...) rifondata sul sogno del suo sindaco Ludovico Corrao</i> (schema:description)	
<i>Piazza San Domenico, 1 - Palermo</i> (fce:place_name)	<i>Palermo</i> (schema:addressLocality) 90133 (schema:postalCode) <i>Piazza San Domenico, 1</i> (schema:streetAddress) lat: 38.1187555 (schema:latitude) long:13.3634137 (schema:longitude)
lat: 38.1187555 (schema:latitude) long:13.3634137 (schema:longitude)	<http://it.dbpedia.org/resource/Genio_del_Garraffo> <http://it.dbpedia.org/resource/Castellammare_o_Loggia> <http://it.dbpedia.org/resource/Fontana_del_Garraffello> (geonames:nearby)

DBpedia Spotlight

- DBpedia is a collection of entity descriptions extracted from Wikipedia & shared as Linked Data
- DBpedia Spotlight uses data from DBpedia and text from associated Wikipedia pages
- Learns how to recognize that a DBpedia resource was mentioned
- Given plain text as input, generates annotated text