



---

# Entity Linking

Kalina Bontcheva





## What is Entity Linking

---

- Entity linking is the task of identifying all mentions in text of a specific entity from a database or ontology
- Also referred to as entity disambiguation
- Researchers have used Wikipedia (e.g. TAC KBP, WikipediaMiner) or Linked Open Data (in particular DBpedia, YAGO, and Freebase)
- Typically broken down into two main phases:
  - Candidate selection (entity annotation)
  - Reference disambiguation or entity resolution



## What is EL (2)

---

- The entity linking system can either return a matching entry from the target knowledge base (e.g. DBpedia URI, Wikipedia URL) or NIL to indicate there is no matching entry in the entity database
- Much of the work on entity linking makes the closed world assumption, i.e. that there is always a target entity in the database
- This is limiting for blogs, tweets, and similar social media
- Typically focused on PER, LOC, ORG entities and English documents



## What is EL (3)

---

- Entity linking needs to handle:
  - Name variations (entities are referred to in many different ways)
  - Entity ambiguity (the same string can refer to more than one entity)
  - Missing entities – there is no target entity in the entity knowledge base/database



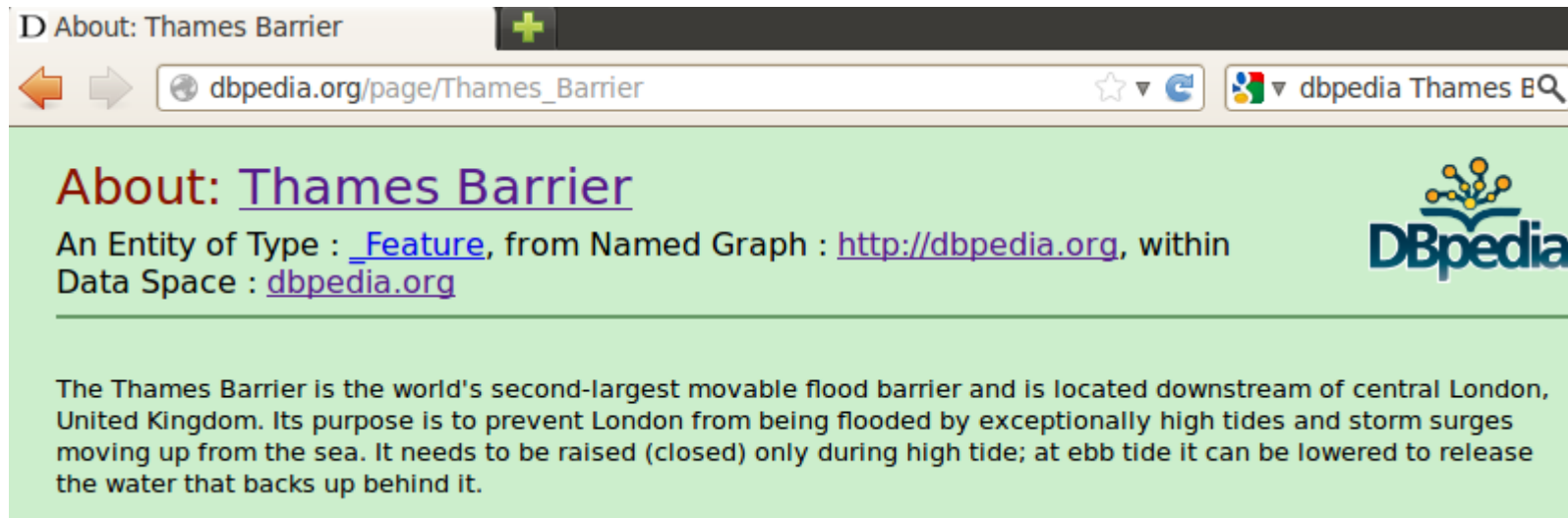
## Why entity linking?

---

- **Entity linking:** rather than just annotate the words “Berlusconi” and “Берлускони” as a Person (NER), link it to a specific ontology instance
  - Differentiate between Silvio Berlusconi, Marina Berlusconi, etc.
  - Ontologies tell us that this particular Berlusconi is a Politician, which is a type of Person. He is based in Italy, which is part of the EU. He was a prime minister, etc. This is all helpful to disambiguate and link the mention in the text to the correct entity URI in the ontology
- **Link documents across languages** and support queries for a specific entity in one language to return results in another

- 
- Machine readable knowledge on various entities and topics, including:
    - 410,000 places/locations,
    - 310,000 persons
    - 140,000 organisations
  - For each entity we have:
    - Entity name variants (e.g. IBM, Int. Business Machines)
    - a textual abstract
    - reference(s) to corresponding Wikipedia page(s)
    - entity-specific properties (e.g. latitude and longitude for places)

# Example from DBpedia



D About: Thames Barrier

dbpedia.org/page/Thames\_Barrier

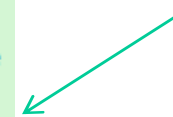
## About: Thames Barrier

An Entity of Type : [Feature](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](http://dbpedia.org)

The Thames Barrier is the world's second-largest movable flood barrier and is located downstream of central London, United Kingdom. Its purpose is to prevent London from being flooded by exceptionally high tides and storm surges moving up from the sea. It needs to be raised (closed) only during high tide; at ebb tide it can be lowered to release the water that backs up behind it.

■ ■ ■

<a href="#">owl:sameAs</a>	<ul style="list-style-type: none"><li>▪ <a href="http://cs.dbpedia.org/resource/Bariéry_na_Temži">http://cs.dbpedia.org/resource/Bariéry_na_Temži</a></li><li>▪ <a href="http://de.dbpedia.org/resource/Thames_Barrier">http://de.dbpedia.org/resource/Thames_Barrier</a></li><li>▪ <a href="http://fr.dbpedia.org/resource/Barrière_de_la_Tamise">http://fr.dbpedia.org/resource/Barrière_de_la_Tamise</a></li><li>▪ <a href="http://it.dbpedia.org/resource/Thames_Barrier">http://it.dbpedia.org/resource/Thames_Barrier</a></li><li>▪ <a href="http://sws.geonames.org/2636058/">http://sws.geonames.org/2636058/</a></li><li>▪ <a href="#">freebase:Thames Barrier</a></li></ul>
<a href="#">geo:geometry</a>	▪ POINT(0.0367 51.4977)
<a href="#">geo:lat</a>	▪ 51.497700 (xsd:float)
<a href="#">geo:long</a>	▪ 0.036700 (xsd:float)





# GeoNames

---

- 2.8 million populated places
  - 5.5 million alternate names
- Knowledge about NUTS country sub-divisions
  - use for enrichment of recognised locations with the implied higher-level country sub-divisions
- However, the sheer size of GeoNames creates a lot of ambiguity during semantic enrichment
- We use it as an additional knowledge source, but not as a primary source (DBpedia)

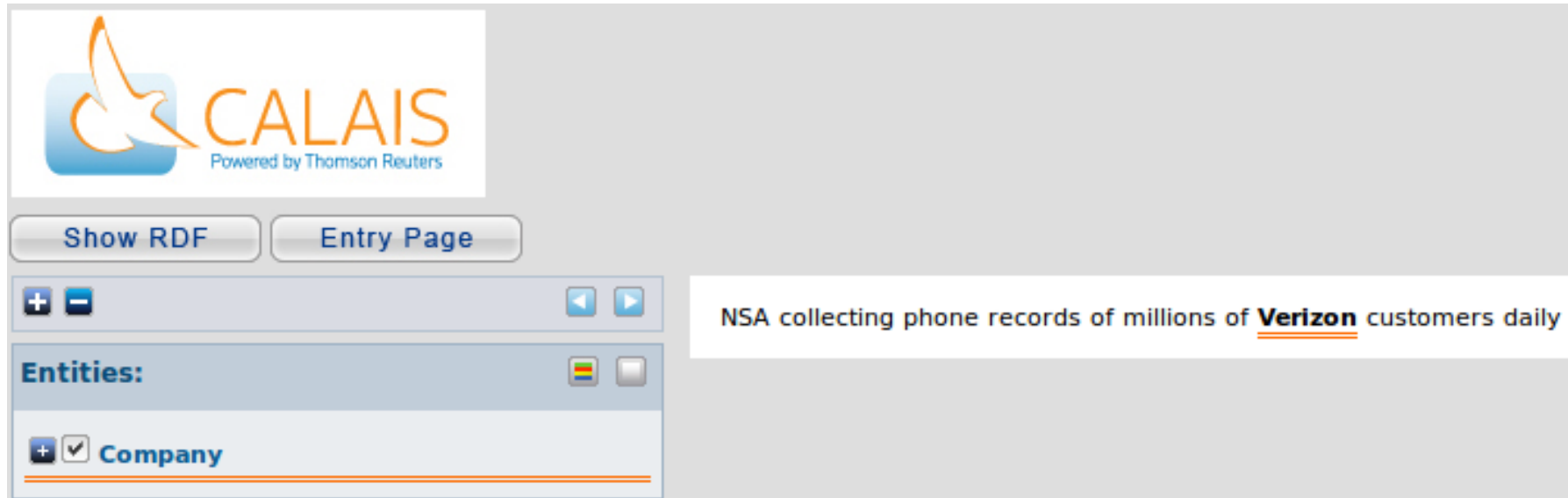






# Entity Linking Systems: OpenCalais

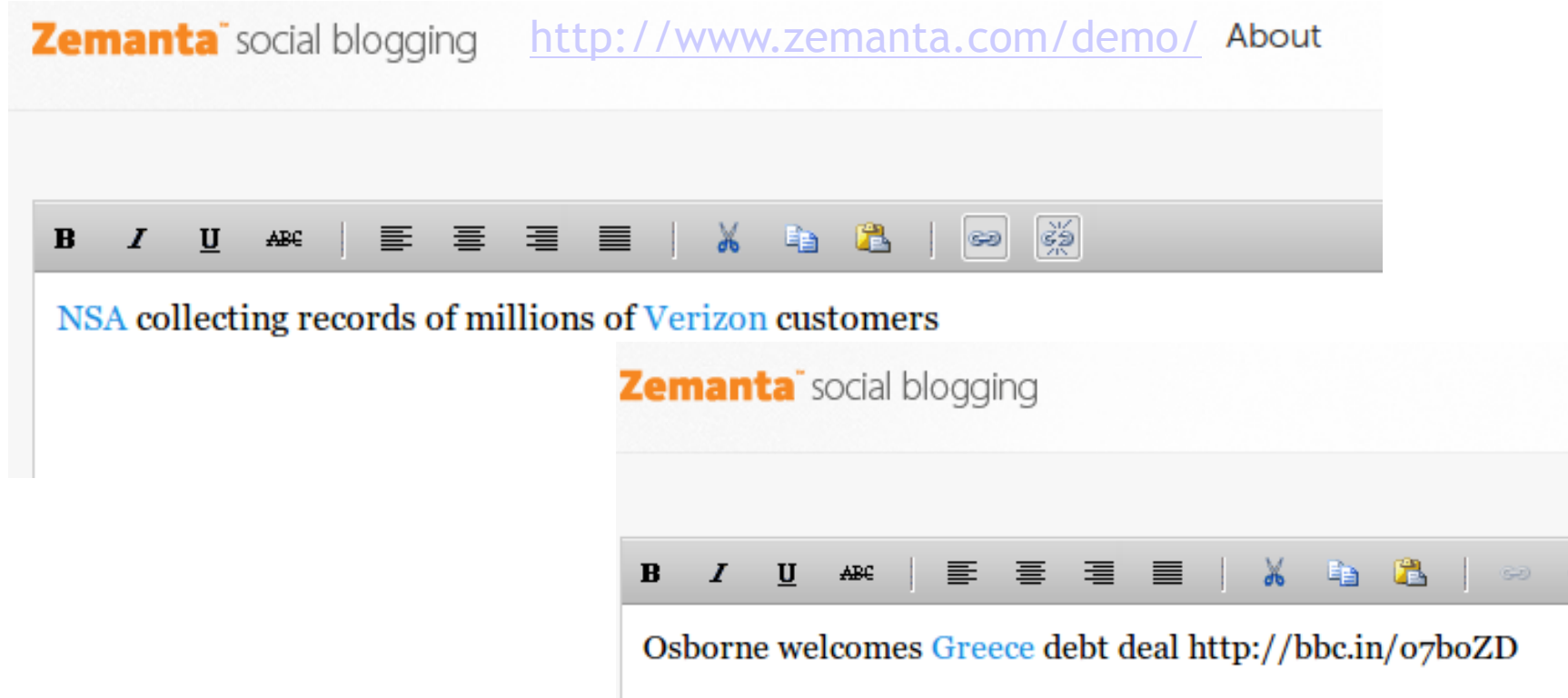
<http://viewer.opencalais.com/>

A screenshot of the OpenCalais viewer interface. The top left features the OpenCalais logo, which includes a stylized orange bird icon and the text 'CALAIS Powered by Thomson Reuters'. Below the logo are two buttons: 'Show RDF' and 'Entry Page'. A sidebar on the left contains a window with a '+' and '-' icon, and a section titled 'Entities:' with a '+', a checked checkbox, and the label 'Company'. The main content area on the right displays a text snippet: 'NSA collecting phone records of millions of Verizon customers daily', where 'Verizon' is underlined in red.

Not easily customised/extended

Domain-specific coverage varies

# EL Systems: Zemanta



The screenshot displays the Zemanta social blogging interface. At the top, it shows the text "Zemanta™ social blogging" followed by a URL <http://www.zemanta.com/demo/> and an "About" link. Below this is a rich text editor toolbar with icons for bold (B), italic (I), underline (U), text color (ABC), bulleted list, numbered list, indent, outdent, undo, redo, save, link, and unlink. The main text area contains the sentence "NSA collecting records of millions of Verizon customers". A second instance of the interface is shown below, with the text "Osborne welcomes Greece debt deal" and a link <http://bbc.in/o7boZD> inserted at the end of the sentence.

- Commercial service, inserts links in blogs to Wikipedia, news articles and similar content
- Our evaluation indicates Zemanta is better than Open Calais. On some tweets it is better than AlchemyAPI, whereas on others - Alchemy is



# EL Systems: AlchemyAPI

www.alchemyapi.com/products/demo/

By using this demo you agree to AlchemyAPI's [terms of use](#).

Osborne welcomes Greece debt deal <http://bbc.in/o7boZD>

Ready to get started with AlchemyAPI?

Free API Key

Entities

Visual

JSON

Docs

Keywords

## Concept Tagging

Concepts

Automatically tag related concepts in HTML, text, or web-based content. [Learn more](#).

Sentiment

Concept	Relevance	Linked Data
Economy of Greece	0.895888	<a href="#">DBpedia</a>

Text

Author

Back to Top



## Ontology Based Information Extraction

A simple demo showing how the [disambiguation service](#) can be used to annotate documents against [DBpedia](#).

Summary of #Greece bailout plan: 109bn aid; maturity of future EFSF loans 15-30 years, lower rates ~~expecte~~... (cont) <http://deck.ly/~A0muj>

↓ Disambiguate ↓


[Summary](#) of [#Greece](#) bailout plan: 109bn aid; maturity of future [EFSF](#) loans 15-30 years, lower rates expecte... (cont) <http://deck.ly/~A0muj>

[dbpedia.org/page/European\\_Financial\\_Stability\\_Facility](http://dbpedia.org/page/European_Financial_Stability_Facility)

About: [European Financial Stability Facility](#)

An Entity of Type : [Eurozone fiscal matters](#), from Named Graph : [http:](#)

# LODIE – English Example 2



Large-scale  
Cross-lingual Trend Mining  
Summarization

of Real-time Media Streams

## Ontology Based Information Extraction

A simple demo showing how the [disambiguation service](#) can be used to annotate documents against [DBpedia](#).

NSA collecting phone records of millions of Verizon customers daily <http://gu.com/p/3gc62/tw> via @guardian Didn't stop the #Boston bombing tho

English ▼

↓ Disambiguate ↓

[NSA](#) collecting phone records of millions of Verizon customers daily <http://gu.com/p/3gc62/tw> via @guardian Didn't stop the #Boston bombing tho

**About: [National Security Agency](#)**

An Entity of Type : [organisation](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](#)





# “South Gloucestershire” Example

Messages Lucene

Annotation Sets

Managing flood risk on the S  
January 2011  
South Gloucestershire to Hi  
Managing flood risk in the S  
We are the Environment Age  
better place \_ for you, and f  
breathe, the water you drin  
Government and society as  
healthier. The Environment  
Please click on the bookma  
brochure to specific points  
Managing flood risk in the S  
Somerset 1

Type	Set	Start
Sem_Location		57
Sem_Location		97
Sem_Location		97
Sem_Location		97
Sem_Location		149
Sem_Location		160
Sem_Location		160

211 Annotations (1 selected)

Sem\_Location

alternateName	South Gloucestershire
caption	South Gloucestershire
count	2
countryCode	GB
geonamesURI	http://sws.geonames.org/3333198/
inst	http://dbpedia.org/resource/South_Gloucestershire
latitude	51.5
longitude	-2.41667
lookupRule	fullString
matched	South Gloucestershire
name	South Gloucestershire
parentAdminURI	http://sws.geonames.org/6269131/, http://sws.geonames.org/3333198/
parentCountryInst	http://sws.geonames.org/2635167/
popularitySimilarity	1.0
randomIndexing	0.0
specificitySimilarity	0.0
string	South Gloucestershire
stringSimilarity	0.2688679
structuralSimilarity	0.0



- Organisation
  - Broadcaster
    - BroadcastNetwork
    - RadioStation
    - TelevisionStation
  - Company
    - Airline
    - LawFirm
    - Publisher
    - RecordLabel
  - EducationalInstitution
    - College
    - Library
    - School
    - University
  - GovernmentAgency
  - Non-ProfitOrganisation
  - PoliticalParty
  - TradeUnion
- Person
  - Celebrity
  - Economist
  - Journalist
  - Judge
  - MilitaryPerson
  - Monarch
  - Politician
    - Chancellor
    - Congressman
    - Deputy
    - Governor
    - Lieutenant
    - Mayor
    - MemberOfParliament
    - President
    - PrimeMinister
    - Senator
    - VicePresident
    - VicePrimeMinister
  - Presenter
    - RadioHost
    - TelevisionHost
  - Royalty
    - BritishRoyalty
  - Scientist
    - Medician
    - Professor
- Place
  - ArchitecturalStructure
  - Building
    - Library
  - PopulatedPlace

▼ Resource Information		
Organisation	Organisation	
URI	http://dbpedia.org/ontology/Organisation	
TYPE	Ontology Class	
▼ Direct Super Classes		
Agent	Agent	
▼ All Super Classes		
Agent	Agent	
▼ Direct Sub Classes		
Broadcaster	Broadcaster	
Company	Company	
EducationalInstitution	EducationalInstitution	
GovernmentAgency	GovernmentAgency	
Non-ProfitOrganisation	Non-ProfitOrganisation	
PoliticalParty	PoliticalParty	
TradeUnion	TradeUnion	
▼ All Sub Classes		
Airline	Airline	
BroadcastNetwork	BroadcastNetwork	
Broadcaster	Broadcaster	
College	College	
Company	Company	
EducationalInstitution	EducationalInstitution	
GovernmentAgency	GovernmentAgency	
LawFirm	LawFirm	
Library	Library	
Non-ProfitOrganisation	Non-ProfitOrganisation	
PoliticalParty	PoliticalParty	
Publisher	Publisher	
RadioStation	RadioStation	
RecordLabel	RecordLabel	
School	School	
TelevisionStation	TelevisionStation	
TradeUnion	TradeUnion	
University	University	
► Equivalent Classes		
▼ Property Types		
comment	[ALL RESOURCES]	
isDefinedBy	[ALL RESOURCES]	
label	[ALL RESOURCES]	
seeAlso	[ALL RESOURCES]	
versionInfo	[ALL RESOURCES]	
▼ Property Values		
label	Organisation	✗
label	organizaçao	✗
label	organizacija	✗
label	조직	✗



# Candidate ambiguity is high = tough task

---

	TAC-KBP				
	PER	LOC	ORG	UKN	TOTAL
Entities	89	361	141	274	865
Avg. number of tokens	1.91	1.20	2.12	1.87	1.78
Candidate URIs	9,427	9,553	9,502	14,649	43,131
Avg. number cand. URIs	105.02	26.46	67.39	53.46	49.86
Unambig. candidates	3	10	3	43	59



## Ontology Based Information Extraction

A simple demo showing how the [disambiguation service](#) can be used to annotate documents against [DBpedia](#).

Italiens Staatspräsident Giorgio Napolitano hat die Forderung von Ex-Premier Silvio Berlusconi nach einer umgehenden Begnadigung zurückgewiesen. Zugleich forderte er Berlusconi auf, sich im Ton zu zügeln. Bei einer Parteiveranstaltung am Samstagabend hatte Berlusconi den Präsidenten aufgefordert, ihn umgehend zu begnadigen - freilich ohne dass er ihn darum bitte. Denn das sei unter seiner Würde. Nach Berlusconis rechtskräftiger Verurteilung wegen Steuerbetrugs will der Senat am kommenden Mittwoch über seinen Ausschluss entscheiden.

German

↓ Disambiguate ↓

Italiens Staatspräsident [Giorgio Napolitano](#) hat die Forderung von Ex-Premier [Silvio Berlusconi](#) nach einer umgehenden Begnadigung zurückgewiesen. Zugleich forderte er Berlusconi auf, sich im Ton zu zügeln. Bei einer Parteiveranstaltung am Samstagabend hatte [Berlusconi](#) den Präsidenten aufgefordert, ihn umgehend zu begnadigen - freilich ohne dass er ihn darum bitte. Denn das sei unter seiner Würde. Nach Berlusconis rechtskräftiger Verurteilung wegen Steuerbetrugs will der Senat am kommenden Mittwoch über seinen Ausschluss entscheiden.

### About: [Silvio Berlusconi](#)

An Entity of Type : [agent](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](#)

Silvio Berlusconi (born 29 September 1936) is an Italian politician and media tycoon who served three times as Prime Minister of Italy from 1994 to 1995, 2001 to 2006, and 2008 to 2011. He is the controlling shareholder of Mediaset and owner of the Italian football club, A.C. Milan. He is nicknamed Il Cavaliere (The Knight) for his Order of Merit for Labour.



## Ontology Based Information Extraction

A simple demo showing how the [disambiguation service](#) can be used to annotate documents against [DBpedia](#).

След като хвърли Италия в политическа криза, Силвио Берлускони се оказа в центъра на нов скандал, предаде Ройтерс. Частната телевизия Ла7 излъчи телефонен разговор, в който експремиерът твърди, че президентът Джорджо Наполитано оказал натиск върху правосъдието в Рим по дело, свързано с медийната империя на Кавалиере.

Bulgarian

↓ Disambiguate ↓

След като хвърли [Италия](#) в политическа криза, [Силвио Берлускони](#) се оказа в центъра на нов скандал, предаде [Ройтерс](#). Частната телевизия Ла7 излъчи телефонен разговор, в който експремиерът твърди, че президентът [Джорджо Наполитано](#) оказал натиск върху правосъдието в [Рим](#) по дело, свързано с медийната империя на Кавалиере.

### About: [Silvio Berlusconi](#)

An Entity of Type : [agent](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](#)

Silvio Berlusconi (born 29 September 1936) is an Italian politician and media tycoon who served three times as Prime Minister of Italy from 1994 to 1995, 2001 to 2006, and 2008 to 2011. He is the controlling shareholder of Mediaset and owner of the Italian football club, A.C. Milan. He is nicknamed Il Cavaliere (The Knight) for his Order of Merit for Labour.

## Multilingual NEL (3)

### Ontology Based Information Extraction

A simple demo showing how the [disambiguation service](#) can be used to annotate documents against [DBpedia](#).

इटली के भू. पू. प्रधानमंत्री बरलुस्कोनी कह रहे हैं कि रिश्वतखोरी जरूरी है .... इन्हें तो भारत की नागरिकता मिल जानी चाहिये

The landscapes of Italy. Preview. Prime Brusconi saying that bribery is necessary ....  
They should be able to get the Indian citizenship

Hindi

↓ Disambiguate ↓

[इटली](#) के भू. पू. प्रधानमंत्री बरलुस्कोनी कह रहे हैं कि रिश्वतखोरी जरूरी है .... इन्हें तो [भारत](#) की नागरिकता मिल जानी चाहिये

About: [Italy](#)

An Entity of Type : [country](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](#)



## Hands On: Try the Various EL services

---

- Open a web browser, one tab per EL system
- Unpack hands-on-module6.zip
- Open examples-entity-linking.txt in an editor
- Try the 4 EL services suggested there on the provided tweets
  
- NB: The LODIE demo that you've just tried is not the latest LODIE 2 system, which will be discussed next. LODIE 2 will be available online from October
- NB: Our EL demo is not for use to process large amounts of text. If you are interested in the latter, please email Kalina and Genevieve