

# NLP & Linked Data: OpeNER and NewsReader

**German Rigau**

<http://adimen.si.ehu.eus/~rigau>



NAZIOARTEKO  
BIKAINTASUN  
CAMPUSA  
CAMPUS DE  
EXCELENCIA  
INTERNACIONAL

# IXA NLP group

- **MCR** (Multilingual Central Repository)
- Academia de la lengua vasca
  - **Observatorio del léxico** (datos enlazados)
- FP7 EU research projects:
  - PATHS: Personalised Access To cultural Heritage Spaces
  - LoCloud: Local content in a Europeana cloud
  - **OpeNER**: Open Polarity Enhanced Named Entity Recognition
  - **NewsReader**: Building structured event Indexes of large volumes of financial and economic Data for Decision Making
  - Readers: Evaluation And DEvelopment of Reading System
  - QTLeap: Quality Translation by Deep Language Engineering Approaches

# MCR

- **MCR** (Multilingual Central Repository)

The screenshot shows a Mozilla Firefox browser window displaying the Multilingual Central Repository website. The browser's address bar shows the URL `adimen.si.ehu.es/web/MCR`. The website's navigation menu on the left includes 'adimen', 'Demos' (with 'MCR' selected), 'Resources', 'Links', and 'User Login'. The main content area is titled 'Multilingual Central Repository' and contains the following text:

Home > Demos

## Multilingual Central Repository

The current version of the **Multilingual Central Repository** (MCR) (Atserias et al. 04, Gonzalez-Agirre et al. 12) is a result of the 5th Framework MEANING project (IST-2001-34460) and Spanish government KNOW (TIN2006-15049-C03), KNOW2 (TIN2009-14715-C04-01) projects and the ongoing SKaTer (TIN2012-38584-C06) projects..

The MCR integrates in the same EuroWordNet framework wordnets from five different languages: English, Spanish, Catalan, Basque and Galician. The Inter-Lingual-Index (ILI) allows the connection from words in one language to equivalent translations in any of the other languages thanks to the automatically generated mappings among WordNet versions. The current ILI version corresponds to WordNet 3.0. Furthermore, the MCR is enriched with the semantically tagged glosses.

The MCR also integrates WordNet Domains, new versions of the Base Concepts and the Top Ontology, and the AdimenSUMO ontology.

In that way, the MCR constitutes a natural multilingual large-scale semantic resource for a number of semantic processes that need large amount of multilingual knowledge to be effective tools.

### News

The Basque WordNet is now distributed under CreativeCommons Attribution 3.0 Unported (CC BY 3.0) license.

This script transforms the Multilingual Central Repository (MCR) 3.0 database so that it can be loaded using the NLTK WordNet reader.

The MCR is integrated into the [Open Multilingual WordNet](#) initiative, BabelNet, and used by Google.

### MCR using WordNet 3.0 as ILI

The current version of the MCR (using **WordNet 3.0** as ILI) can be consulted using the Web EuroWordNet Interface (consult mode).

The current version of the MCR (using **WordNet 3.0** as ILI) can be also edited also edited using the Web EuroWordNet Interface (edit mode).

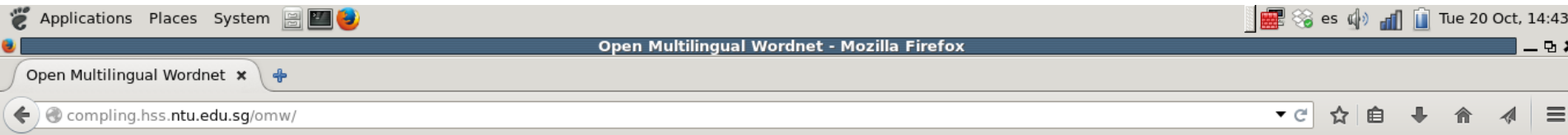
**Download the MCR 3.0** and install it as an SQL database (for both MySQL and PostgreSQL): [zip (37M)] [tgz (37M)]

The MCR 3.0 is distributed under 3 different licenses:

The taskbar at the bottom shows several open windows, including '2015-10-Workshop-Lider', 'Multilingual Central Re...', and '2015-10-GermanRigau...'. The system tray on the right shows the date and time as 'Tue 20 Oct, 14:47'.

# MCR

- **MCR** (Multilingual Central Repository)



## Open Multilingual Wordnet

This page provides access to wordnets in a variety of languages, all linked to the [Princeton Wordnet of English](#) (PWN). The goal is to make it easy to use wordnets in multiple languages. The individual wordnets have been made by many different projects and vary greatly in size and accuracy. This page has (i) extracted and normalized the data, (ii) linked to it Princeton WordNet 3.0 and (iii) put it in one place. This page only includes those with a license that allows redistribution. There is a fuller list at the Global Wordnet Association's [Wordnets in the World page](#).

If you use these wordnets, please cite the original projects who created them (linked in Table 1), if you got value from this aggregation, please cite [Bond and Paik \(2012\)](#).

You can access the wordnets through the (python) [Natural Language Tool-Kit wordnet interface \(NLTK\)](#).

We have an [extended version](#) with automatically extracted data for over a 150 languages from [Wiktionary](#) and the [Unicode Common Locale Data Repository \(Bond and Foster, 2013\)](#).

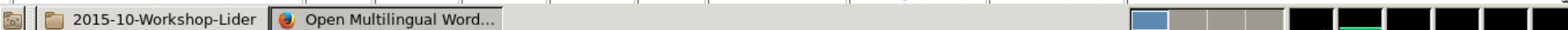
[Documentation](#), [News and Updates](#)

### Search

We have a [simple search interface](#) (search [the extended wordnet](#)). It uses the SQL database originally developed by the Japanese Wordnet.

28 Available Wordnets

Wordnet	Lang	Synsets	Words	Senses	Core	Licence	Data	Citation
<a href="#">Albanet</a>	<a href="#">als</a>	4,675	5,988	9,599	31%	<a href="#">CC BY 3.0</a>	<a href="#">als.zip (+xml)</a>	<a href="#">cite:als; (.bib)</a>
<a href="#">Arabic WordNet (AWN v2)</a>	<a href="#">arb</a>	9,916	17,785	37,335	47%	<a href="#">CC BY SA 3.0</a>	<a href="#">arb.zip (+xml)</a>	<a href="#">cite:arb; (.bib)</a>



# Observatorio del Léxico

- **Euskaltzaindia** (Academia de la Lengua Vasca)
- Corpus que se viene desarrollando desde 2007
- Euskera actual (desde 2000):
  - medios de comunicación
  - Analizado automáticamente: lematización
- Objetivo principal: monitorización del uso real del léxico
- Elhuyar, Uzei, Grupo IXA.

# Observatorio del Léxico

- Datos Enlazados (Corpus - Diccionario Unificado)
- Soporte para la actualización del Diccionario Unificado de la Academia (DU, *Hiztegi Batua*)
  - observar el **uso real** del léxico
  - monitorizar el cumplimiento (o no) de las normas de la Academia
- La normativa léxica dictada por la Academia ha sido integrada en la base de datos léxica
- La BD contiene también formas no estándares
- El corpus lematizado automáticamente **se enlaza al DU**
- 1 millón (2007) – 53 millones de palabras (2015)

# OpeNER

- **OpeNER**: Open Polarity Enhanced Named Entity Recognition
- 2012/07/01 - 2014/06/30
- Vicomtech, VUA, CNR, Synthema, Olery, UPV/EHU.
- Main goal:
  - Easy to use NLP tools in 6 languages for big data text analytics
- Application domain:
  - Tourism
- Focusing on:
  - Opinion Mining and Sentiment Analysis of reviews
- Example:
  - Monitoring Hotel customer opinions

# OpeNER

- **OpeNER**: Open Polarity Enhanced Named Entity Recognition



Consortium:



Home Pre-processed content Live analysis Demo KAF file visualization

Choose language:

English

Select news from inside a cluster:

Egypt judges call for nationwide strike

Egypt's Morsi says new powers temporary

Egypt judges furious over Morsi's decree

Egypt presidency: Morsi decrees 'temporary'

Clashes continue into the night in Egypt

Judiciary slams Morsi's 'unprecedented attack'

Cairo Shares Plunge 9% as Political Crisis Deepens

Pro- and Anti-Morsi Journalists Clash during Crisis Talks

Morsi says new powers 'temporary'

Egypt Clashes Continue as Islamists Call Mass Show of Support for Morsi

Egypt's Morsi faces judicial revolt over decree

Morsi faces judicial rebellion

Morsi's Judicial Decree Draws High-Level Dissent

Egypt stocks plunge after Morsi power grab ? Saudi hits 10-month low as Egypt's crisis weighs

Egypt's Morsi says new powers temporary, urges dialogue

Judges slam all powerful president

NER Legend



Sentiment Legend



- Polarity +

Text NERC **Sentiment** KAF Images Map Opinion Coreference

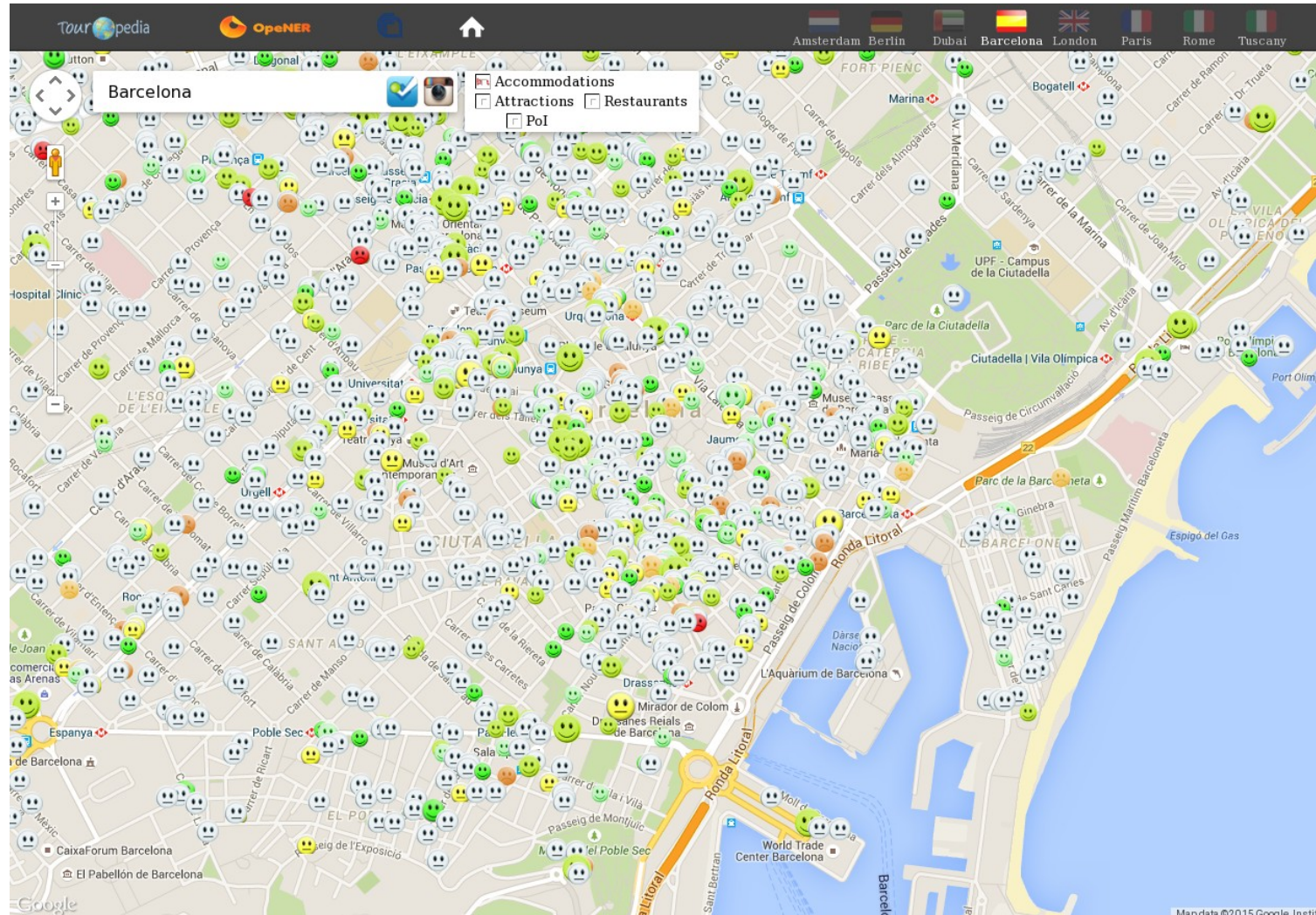
## Egypt's Morsi says new powers temporary

Egypt Islamists call mass show of **support** Egyptian President Mohamed Morsi says the sweeping new powers he has assumed are only temporary and has called for dialogue, as clashes in the Nile Delta saw a member of his party **killed**, medics said. Morsi's constitutional declaration on Thursday, allowing him to issue decisions and **laws** unchallenged, triggered a wave of **protest** and has set him on course for a showdown with Egypt's judges, whom he is due to meet on Monday in a bid to defuse the crisis. Clashes between **supporters** and opponents of the president, outside a Muslim Brotherhood headquarters in Damanhour, saw one Islamist **killed** and 10 people wounded, a doctor at the hospital in the Nile Delta town told AFP. Witnesses said the clashes, in which protesters hurled petrol **bombs** and stones, followed three days of unrest there, with Morsi's opponents trying to storm the Brotherhood office. Several offices belonging to the Muslim Brotherhood's Freedom and Justice Party have been torched since Thursday's announcement. "The death of this young Islamist and the fires targeting the party's offices show that certain people are trying to lead the country towards **chaos**," the party's president Saad Al Katatni said on his Facebook page. Last week's constitutional declaration states that Morsi can issue "any decision or measure to protect the revolution," which are final and **not** subject to appeal. The announcement has sparked **charges** that he is taking on dictatorial powers. In a move to assuage his **critics**, Morsi is to meet the Supreme Judicial Council on Monday, after his **justice** minister Ahmed Mekki held preliminary talks with the council, the president's spokesman Yasser Ali said. The president on Sunday emphasised the "temporary nature" of the new measures, which would apply only until a new constitution is adopted and elections held, and which "are **not** meant to concentrate power," but devolve it to a democratically-elected parliament. The measures were also "deemed necessary in order to hold accountable those **responsible** for the corruption as **well** as the other crimes during the previous regime and during the transitional period." Tahrir Square, one of the **capital**'s crossroads, remained closed to traffic on Sunday as Morsi opponents **pressed** a sit-in, while **nearby** clashes between police and protesters, which entered their second week, occasionally **spilled** into the square. Separately, hundreds of Morsi **supporters** demonstrated in front of mosques in Cairo and across the country in **protests** called for by the Muslim Brotherhood. Anti-riot police began erected a concrete **barrier** to keep the Tahrir protesters away from **nearby** government buildings, witnesses said.



# OpeNER

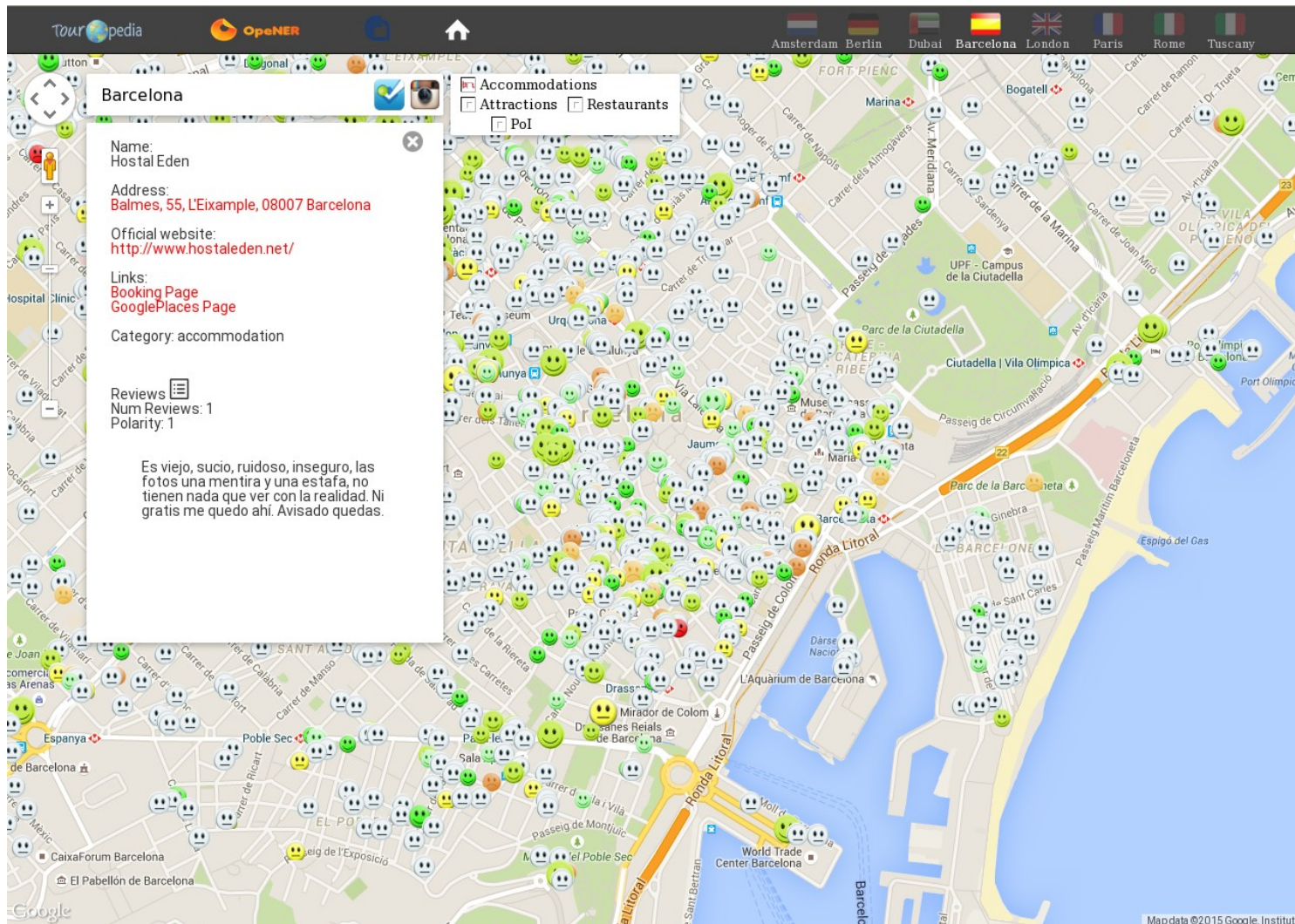
- **OpeNER**: Open Polarity Enhanced Named Entity Recognition





# OpeNER

- **OpeNER**: Open Polarity Enhanced Named Entity Recognition



# NewsReader

- **NewsReader**: Building structured event Indexes of large volumes of financial and economic Data for Decision Making
- 2013/01/01 - 2015/12/31
- VUA, FBK, LexisNexis, ScraperWiki, UPV/EHU.
- Main goal:
  - StoryLines in 4 languages at large scale
  - From NLP to Semantic Web
- Application domain:
  - Economy & Finance
- Focusing on:
  - Cross-lingual Event Mining (*who did what when and where*)
- Example:
  - Monitoring car industry, wikinews, etc.

# NewsReader

- **NewsReader**: Building structured event Indexes of large volumes of financial and economic Data for Decision Making

KnowledgeStore UI Lookup SPARQL query Reports

ID  Lookup example URI 1 resource found

**Resource text** Download Select resource metadata Select entity (102) Select mention (190)

In this week's interview on Fox News Sunday, former U.S. President Bill Clinton described a question by the host Chris Wallace as a "conservative hitjob." According to the introduction given by Wallace, a pre-interview agreement to split the interview evenly between questions on Clinton's Global Initiative and any topic of the news. After a few opening questions, Wallace raised the issue of Clinton's efforts to deal with Osama bin Laden, "Why didn't you do more to put bin Laden and al-Qaeda out of business record, highlighting his efforts to kill bin Laden and comparing them with the record of the current administration before September 11, 2001. They had no meetings on bin Laden. Clinton's determination to rebut the implication that he had not done enough to kill bin Laden took the majority of the first two sections of the three-part interview. As well as context for describing his efforts to kill bin Laden as an "obsession", Clinton asked Wallace why the same questions were not put to the current administration about their pre-September 11 Clinton pointed out that the political-right attacked his efforts to kill bin Laden as an attempt to "wag the dog." He also said that he could not get the FBI or CIA to certify that bin Laden was attacking the Taliban in Afghanistan were not put into action.

In the much reduced amount of time devoted to Clinton's Global Initiative it was revealed that compared to last year's 2.5 billion dollars, this year commitments totaling more than \$1 billion pledged to give all the profits from his rail operations.

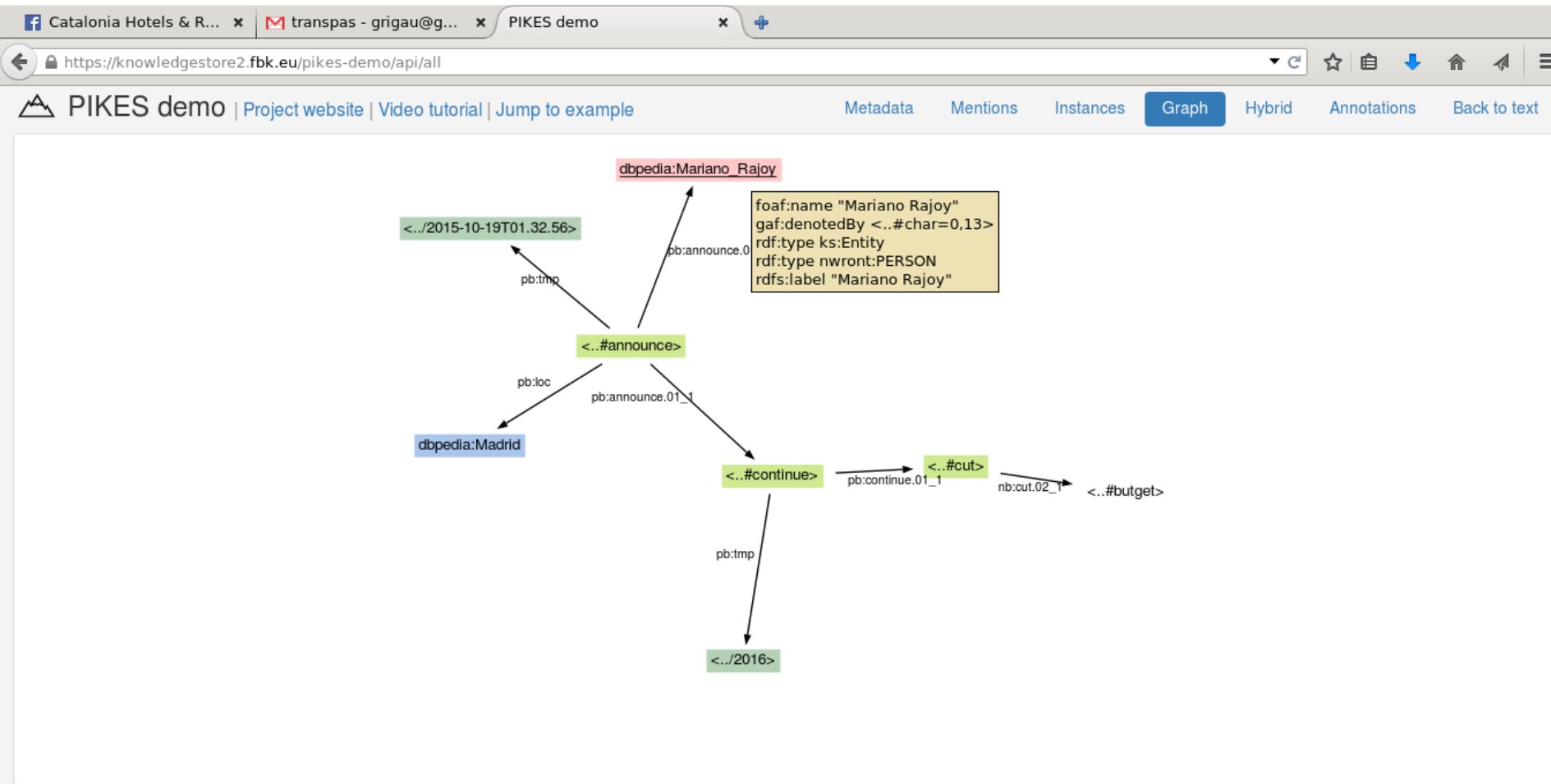
**Entity <../Laden> (437 triples out of 437)**

subject	predicate	object
<../Laden>	rdf:type	<../LOCATION>
<../Laden>	rdf:type	<../MISC>
<../Laden>	rdf:type	<../ORGANIZATION>
<../Laden>	rdf:type	<../PERSON>
<../Laden>	rdfs:label	Bin Laden
<../Laden>	rdfs:label	Osama bin Laden
<../Laden>	rdfs:label	bin Laden
<../Laden>	rdfs:label	his
<../Laden>	rdfs:label	whose
<../Laden>	rdfs:label	Laden
<../Laden>	rdfs:label	Osama bin Laden 's
<../Laden>	rdfs:label	bin Laden 's
<../Laden>	rdfs:label	bin laden
<../Laden>	rdfs:label	i ]
<../Laden>	rdfs:label	laden
<../Laden>	rdfs:label	their leader Osama bin Laden
<../Laden>	rdfs:label	a trophy
<../Laden>	gaf:denotedBy	<.#char=1435,1440>

dbpedia:Alfred\_Russel\_Wallace  
dbpedia:Bill\_Clinton  
dbpedia:Central\_Intelligence\_Agency  
dbpedia:Chris\_Wallace  
dbpedia:Federal\_Bureau\_of\_Investigation  
dbpedia:Fox\_News\_Channel  
dbpedia:Hillary\_Rodham\_Clinton  
dbpedia:Osama\_bin\_Laden  
dbpedia:Richard\_Branson  
dbpedia:Taliban  
dbpedia:USS\_Cole\_(DDG-67)  
dbpedia:United\_States\_Armed\_Forces  
dbpedia:War\_in\_Afghanistan\_(2001-2021)  
dbpedia:War\_in\_Afghanistan\_(2001-present)  
<.#ev15>  
<.#ev16>  
<.#ev17>  
<.#ev18\_1>  
<.#ev18\_2>  
<.#ev18\_3>  
<.#ev18\_4>  
<.#ev18\_5>  
<.#ev18\_6>  
<.#ev18\_7>  
<.#ev18\_8>  
<.#ev18\_9>  
<.#ev19>  
<.#ev20>  
<.#ev21>  
<.#ev22>

# NewsReader

- **NewsReader**: Building structured event Indexes of large volumes of financial and economic Data for Decision Making



# NLP & Linked Data: OpeNER and NewsReader

**German Rigau**

<http://adimen.si.ehu.eus/~rigau>



NAZIOARTEKO  
BIKAINTASUN  
CAMPUSA  
CAMPUS DE  
EXCELENCIA  
INTERNACIONAL