

Cross-Linguality as Global Social Sensor

<http://eventregistry.org/>

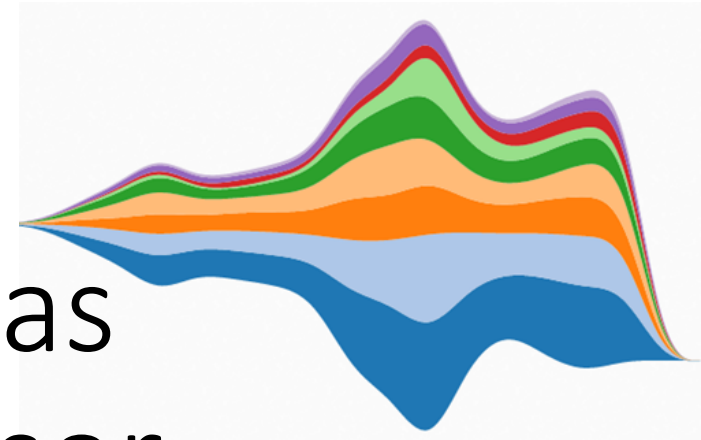
Marko Grobelnik

Marko.Grobelnik@ijs.si

Artificial Intelligence Lab, Jozef Stefan Institute
Ljubljana, Slovenia

Contributions from Gregor Leban, Blaz Fortuna, Janez Brank, Jan Rupnik,
Andrej Muhic, Aljaz Kosmerlj, Evgenia Belyaeva, Pat Moore

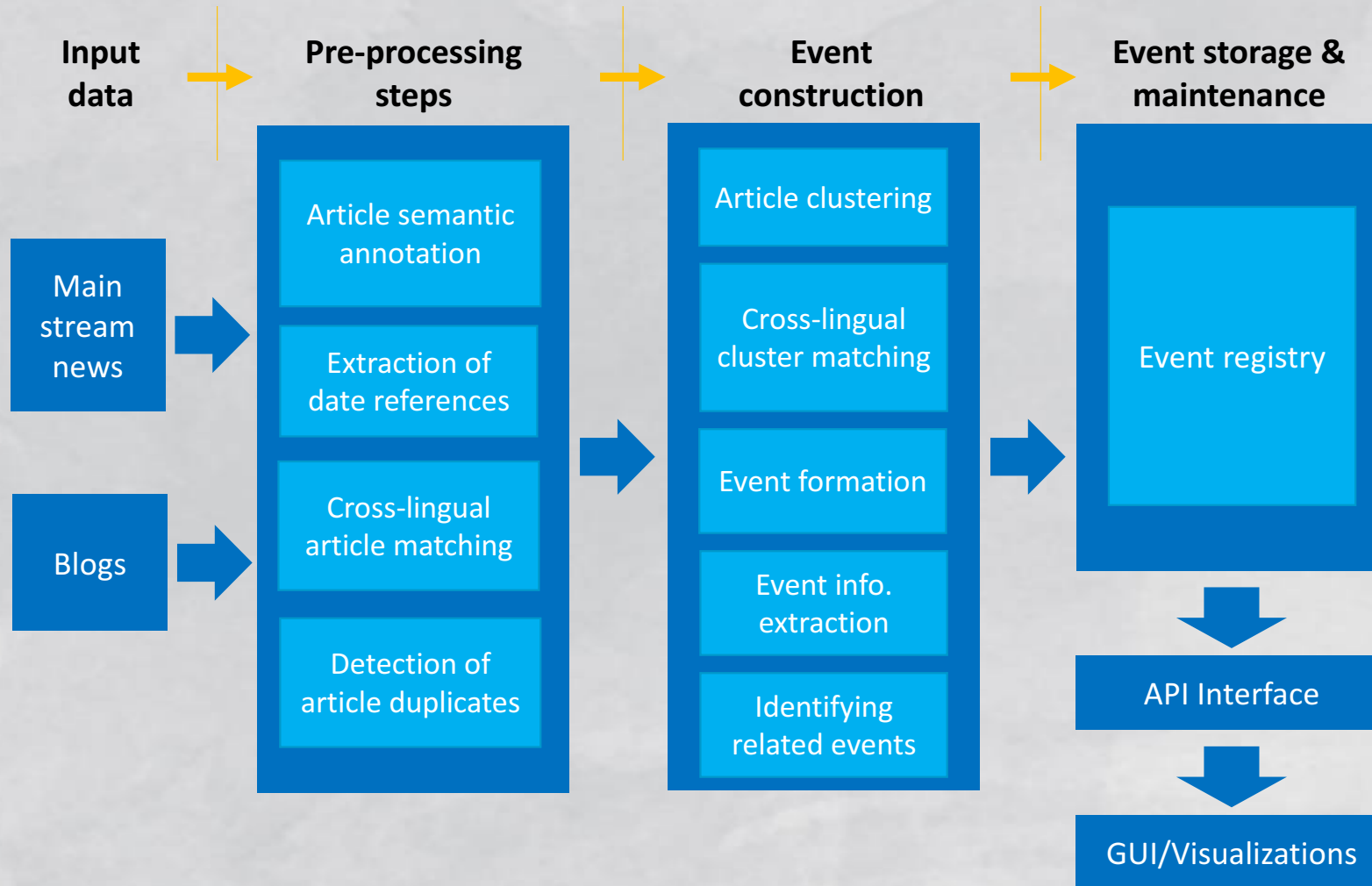
MetaForum, 2017, Nov 13th, Brussels



What questions are we addressing?

- Where to get global media data?
 - ...in real-time, from any language, as complete as possible
- What is extractable from media documents?
 - ...**knowledge extraction**
- How to connect information across languages?
 - ...**cross-linguality**
- How to structure news data in a form of events?
 - ...**modelling social dynamics**
- How to approach diversity in news reporting?
- How to query & visualize global event dynamics?

Global Media Monitoring pipeline



<http://EventRegistry.org>

Collecting Media Data

<http://newsfeed.ijs.si/>

Collecting global media data

- Data collection service News-Feed
 - <http://newsfeed.ijs.si/>
 - ...crawling global main-stream and social media
- Monitoring
 - ~150k main-stream publishers (RSS feeds+special feeds)
 - ~250k most influential blogs (RSS feeds)
 - free Twitter feed
- Data volume: ~350k articles & blogs per day (+5M tweets)
- Languages: eng (50%), ger (10%), spa (8%), fra (5%), ...



Example: Brussels Media

- **Brussels News**

- Actua TV
- De Morgen
- De Standaard
- Het Laatste Nieuws
- Het Nieuwsblad
- Journal du Mardi Magazine
- Knack
- La Capitale
- La Dernière Heure
- La Libre Belgique
- La Première
- La Province
- Le Soir
- Meervoud
- Radio 1 Brussels
- Tele Bruxelles

- **Brussels Business**

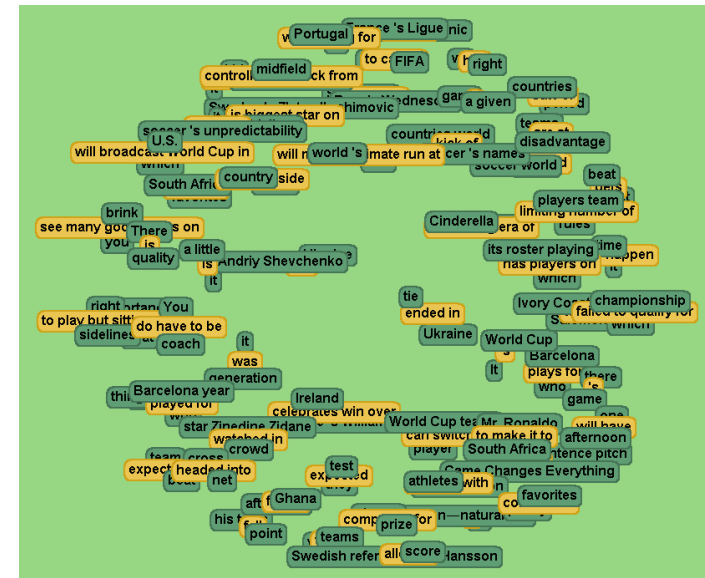
- Canal Z
- L'Echo Brussels
- Brussels Entertainment
- AB3 TV
- Be 1 TV
- Bruzz
- Canvas
- Een TV
- Humo
- La Deux
- La Trois TV Belgium
- La Une TV
- Radio 2 Brussels
- VivaCite
- VT4 TV
- VTM TV

- **Brussels Sports**

- Sport Wereld
- Voetbal

- **Brussels Society**

- European Voice
- Moniteur Belge



Document Enrichment

<http://wikifier.org/>

<http://enrycher.ijs.si/>

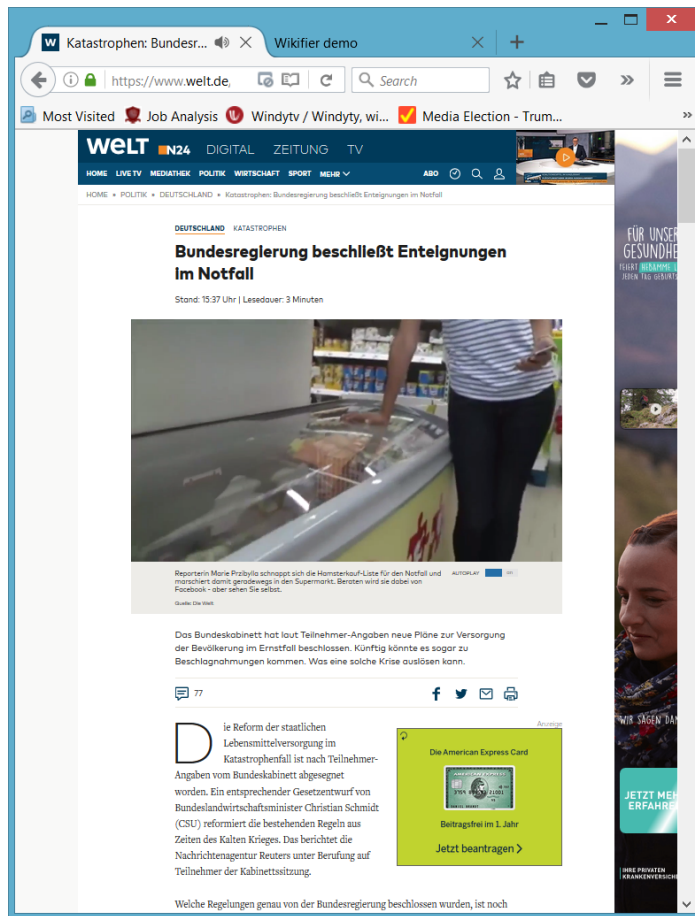
What can be extracted from a document?

- Lexical level
 - **Tokenization** – extracting tokens from a document (words, separators, ...)
 - **Sentence splitting** – set of sentences to be further processed
- Linguistic level
 - **Part-of-Speech** – assigning word types (nouns, verbs, adjectives, ...)
 - **Deep Parsing** – constructing parse trees from sentences
 - **Triple extraction** – subject-predicate-object triple extraction
 - **Name entity extraction** – identifying names of people, places, organizations
- Semantic level
 - **Co-reference resolution** – replacing pronouns with corresponding names; merging different surface forms of names into single entity
 - **Semantic labeling** – assigning semantic identifiers to names (e.g. LOD/DBpedia/Freebase) including disambiguation
 - **Topic classification** – assigning topic categories to a document (e.g. DMoz)
 - **Summarization** – assigning importance to parts of a document
 - **Fact extraction** – extracting relevant facts from a document

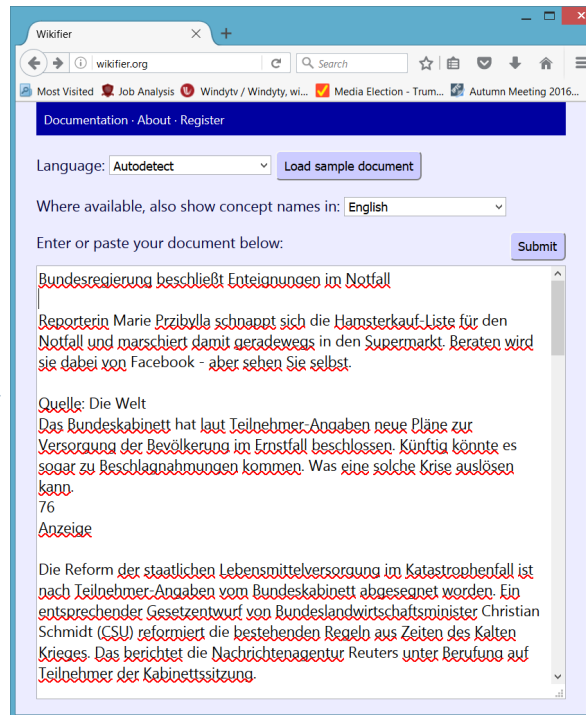
Semantic annotation with [Wikifier.Org](https://www.wikifier.org/)

(operates for 100 languages)

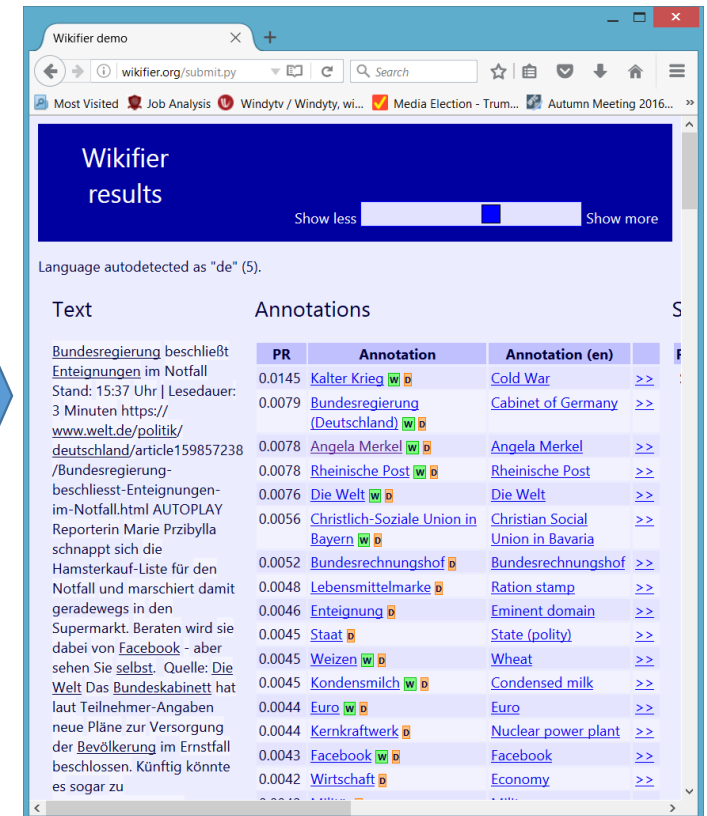
Original text



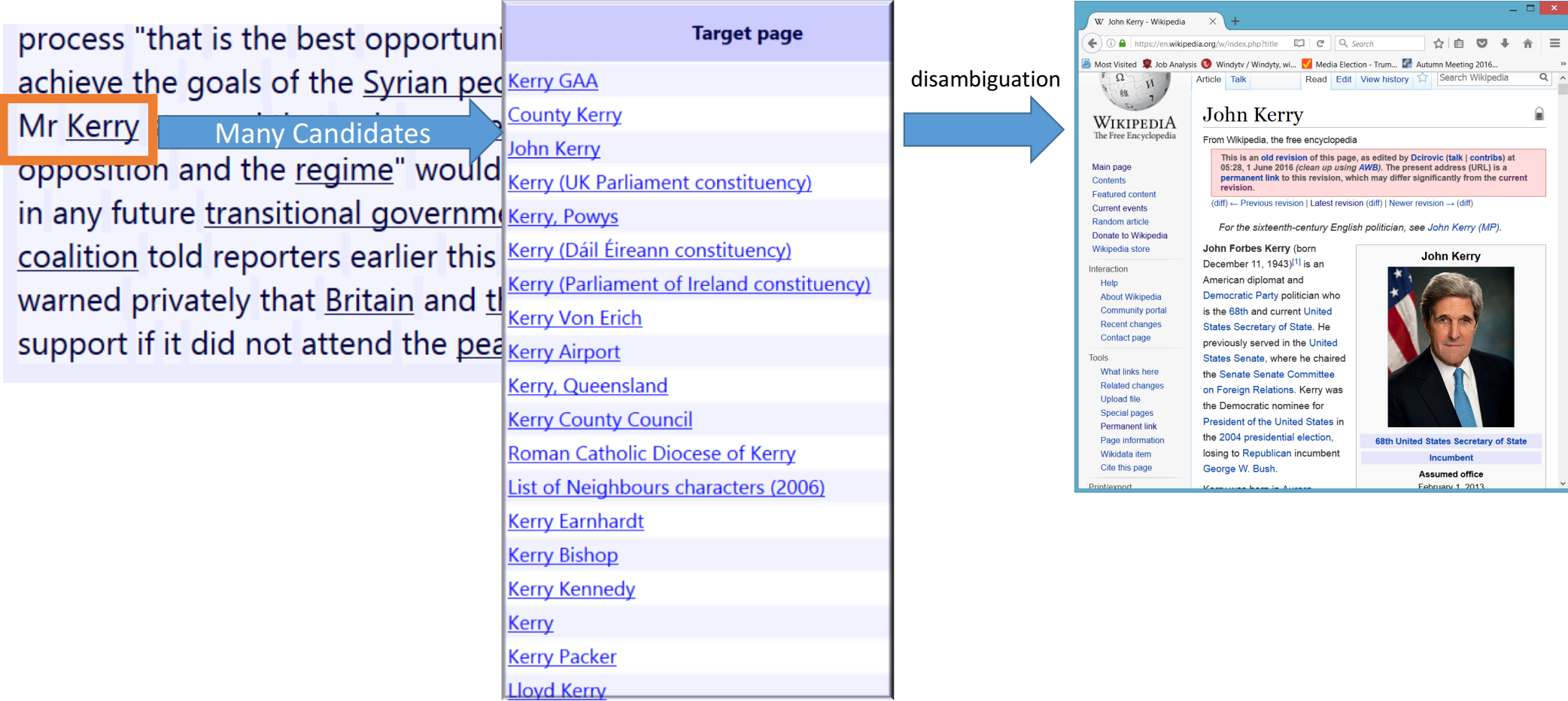
Plain clean text



Annotated text in original language and English



Key problem in semantic annotation (wikification) is concept disambiguation



Enrycher (<http://enrycher.ijs.si/>)

Plain text

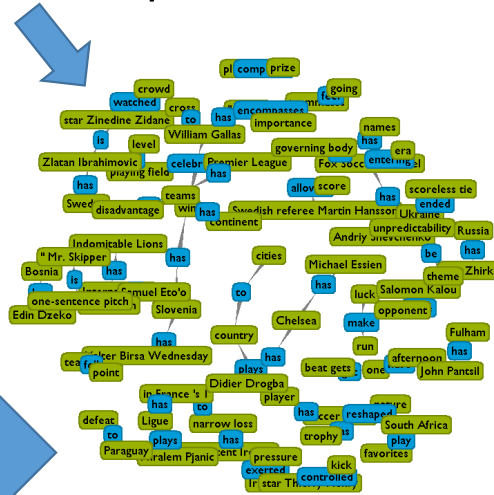


Slovenia's dramatic win over Russia Wednesday, and to a lesser extent Ireland's narrow loss to France, capped off a grueling two-year qualifying period that saw some of the smallest countries in the world kick some of soccer's biggest names in the teeth. After a century of near domination from the likes of Brazil, Italy and Germany, international soccer is entering the era of the Cinderella. It may not happen this time around, but given the increasing flow of talent, training and information across borders, it's almost certain that a small upstart nation blessed with good athletes and better luck will make a legitimate run at the world's most coveted trophy.

Russia's Yuri Zhirkov, right, fights for the ball with Slovenia's Edin Dzeko Wednesday.

Text
Enrichment

Extracted graph of triples from text



entities

- [Brazil](#)
- [Italy](#)
- [Germany](#)
- [Cinderella](#)
- [Paris](#)
- [John O'Shea](#)
- [Manchester United](#)
- [Robbie Keane](#)
- [Shay Given](#)
- [Greece](#)
- [Portugal](#)
- [Bosnia-Herzegovina](#)
- [Cristiano Ronaldo](#)
- [Uruguay](#)

keywords

Sports, Soccer, CONCACAF, Competitions, United States, Sports and Hobbies, Kids and Teens, World Cup, Women,

categories

- [Top/Kids_and Teens /Sports_and Hobbies /Sports/Soccer](#)
- [Top/Sports/Soccer /Competitions](#)
- [Top/Sports/Soccer /Competitions/World_Cup](#)
- [Top/Sports/Soccer /CONCACAF](#)

“Enrycher” is available as a web-service generating Semantic Graph, LOD links, Entities, Keywords, Categories, Text Summarization, Sentiment

Diego Maradona Semantics:

owl:sameAs: http://dbpedia.org/resource/Diego_Maradona

owl:sameAs: <http://sw.opencyc.org/concept/Mx4rvofERZwpEbGdrcN5Y29ycA>

rdf:type: <http://dbpedia.org/class/yago/ArgentinianInternationalFootballers>

rdf:type: <http://dbpedia.org/class/yago/ArgentineExpatriatesInItaly>

rdf:type: <http://dbpedia.org/class/yago/ArgentineFootballManagers>

rdf:type: <http://dbpedia.org/class/yago/ArgentineFootballers>

Robbie Keane Semantics:

owl:sameAs: http://dbpedia.org/resource/Robbie_Keane

rdf:type: <http://dbpedia.org/class/yago/CoventryCityF.C.Players>

rdf:type: <http://dbpedia.org/class/yago/ExpatriateFootballPlayersInItaly>

rdf:type: <http://dbpedia.org/class/yago/F.C.InternazionaleMilanoPlayers>

Cross-linguality

<http://xling.ijs.si/>

Cross-linguality

How to operate in many languages?

- Cross-linguality is a set of functions on how to transfer information across the languages
 - ...having this, we can track information independent of the language borders
 - Machine Translation is expensive and slow, so the goal is to avoid machine translation to gain speed and scale
- The key building block is the function for comparing and categorization of documents in different languages
 - <http://XLing.ijs.si> is an open web service to bridge information across 100 languages

Languages covered by XLing (top 100 Wikipedia languages)

Afrikaans	Chinese	Hungarian	Malagasy	Simple English
Albanian	Chuvash	Icelandic	Malay	Slovak
Arabic	Croatian	Ido	Malayalam	Slovenian
Aragonese	Czech	Indonesian	Marathi	Spanish
Armenian	Danish	Irish	Nepali	Sundanese
Asturian	Dutch	Italian	Newar / Nepal Bhasa	Swahili
Azerbaijani	English	Japanese	Norwegian (Bokm?)	Swedish
Bashkir	Esperanto	Javanese	Norwegian (Nynorsk)	Tagalog
Basque	Estonian	Kazakh	Occitan	Tamil
Belarusian	Finnish	Kirghiz	Persian	Tatar
Belarusian (Taraškievica)	French	Korean	Piedmontese	Telugu
Bengali	Galician	Kurdish	Polish	Thai
Bishnupriya Manipuri	Georgian	Latin	Portuguese	Turkish
Bosnian	German	Latvian	Quechua	Ukrainian
Breton	Greek	Lithuanian	Romanian	Urdu
Bulgarian	Gujarati	Lombard	Russian	Uzbek
Burmese	Haitian	Low Saxon	Serbian	Vietnamese
Cantonese	Hebrew	Luxembourgish	Serbo-Croatian	West Frisian
Catalan	Hindi	Macedonian	Sicilian	Western Panjabi
Cebuano				Yoruba

XLing (<http://XLing.ijs.si>)

service for comparing and categorization of documents across 100 languages

The screenshot displays the XLing interface for comparing two documents. At the top, a similarity score of 0.669834 is shown between two documents. Below this, there are two columns of text, one in English and one in Chinese. The English text is on the left, and the Chinese text is on the right. The interface includes several annotations:

- Automatically Extracted Keywords:** Two boxes point to the keyword lists for each document. The English keywords are: china, economi, econom, growth, bank, market, chines, rate, that, global. The Chinese keywords are: 经济, 中国, 银行, 增长, 全球, 企业, 在, 改革, 政府, 资本.
- Similarity Between Two Documents:** A box points to the similarity score of 0.669834.
- Automatically Extracted Keywords:** A box points to the Chinese keyword list.
- English Text:** A box points to the English document content.
- Chinese Text:** A box points to the Chinese document content.
- Selection Of 100 Languages:** A box points to the language selection dropdown menu.

The English document content includes: "World Bank cuts China and Thailand's growth forecasts", "World Bank trims East Asia growth forecasts", "Good morning, and welcome to our rolling coverage of events across the financial markets, the global economy, the eurozone and business.", "The World Bank has kicked off the week by cutting its growth forecasts for the Asian economy, including a sharp downgrade to Thailand following months of political unrest.", "It warned that 'there was a bumpy start to 2014, notably in China and the United States', as it lowered its forecast for growth across the East Asia and Pacific region.", "It now expects GDP to rise by 7.1% in 2014 and 2015 across the area, down from the 7.2% previously forecast for both years.", "And it admitted that this could be too optimistic, given the wider

The Chinese document content includes: "世界银行:中国经济'硬着陆'风险仍存", "我们认为,中国经济的挑战还是存在的,经济放缓依然可能失去控制.....'硬着陆'仍是需要考虑的一种情形。"世界银行发展展望部全球宏观趋势项目主管伯恩斯(Andrew Burns)周二(6月10日)在伦敦接受BBC中文网记者专访时表示。", "相关内容", "中国5月份出口猛涨贸易顺差飙升", "IMF下调中国增长预期 呼吁北京继续改革", "中国经济放缓 冲击全球民航利润", "更多相关的故事", "相关新闻话题", "中国, 亚洲,"

Example: Cross-lingual News Recommendation

- What local media (e.g. German) is writing about the topic we are reading in English?
- Usual fear of publishers: are users we are sending away coming back?
 - ...evaluation shows they all come back

The screenshot shows the Bloomberg website interface. At the top, there's a navigation bar with 'Bloomberg.com', 'Businessweek.com', 'Bloomberg TV', and 'Premium'. Below this is a 'MARKET SNAPSHOT' section with a table of market indices:

	U.S.	EUROPE	ASIA
DJIA	12,938.10	-158.20	-1.21%
S&P 500	1,402.43	-15.67	-1.10%
NASDAQ	2,960.31	-25.59	-0.86%

Below the market snapshot is a green banner for 'SEIZE THE WEEK WITH THE WEALTH WATCH NEWSLETTER' with a 'SIGN UP NOW >>>' button. The main navigation bar includes 'HOME', 'QUICK', 'NEWS', 'OPINION', 'MARKET DATA', 'PERSONAL FINANCE', 'TECH', 'POLITICS', 'SUSTAINABILITY', 'TV', 'VIDEO', and 'RADIO'. The 'POLITICS' tab is selected.

The main article is titled 'Taxes to Rise for Workers as Budget Deal Still Elusive' by Kathleen Hunter & Roxana Tiron, dated Dec 31, 2012. The article text reads: 'With taxes set to increase for almost every U.S. worker at midnight, Congress hasn't reached a budget deal that Democrats and Republicans say is necessary to prevent a blow to the U.S. economy.'

Below the article is a photo of Harry Reid and Mitch McConnell with the caption: 'Private talks between Senate Majority Leader Harry Reid and Minority Leader Mitch McConnell that began Dec. 28 stalled yesterday because of disputes over income tax rates, the estate tax and other issues.'

On the right side, there's a sidebar with a 'GET THE POLITICAL CAPITAL NEWSLETTER' sign-up button. Below that is a 'HEADLINES' section with a 'DEUTSCH NACHRICHTEN' tab selected. The headlines are:

- US-Kongress kommt am Silvestertag zu Sitzung über Budget zusammen Cash Magazine
- US-Haushaltsstreit steuert auf dramatisches Finale zu (2. AF) Cash Magazine
- Ägyptisches Pfund stürzt ab Cash Magazine
- Lästige Konkurrenz: Deutsche Bank neidisch auf Datenkrake Google Die Welt
- Die Frauen, der Tod und ein hilfloser Staat Frankfurter Allgemeine Zeitung
- Erdogan spricht syrischen Flüchtlingen mit zu Cash Magazine

A blue circle highlights the German news recommendations in the sidebar.

News Reporting Bias

<http://aidemo.ijs.si/diversinews/>

News Reporting Bias example

UK SOLDIERS CLEARED IN IRAQI DEATH – SEVEN BRITISH SOLDIERS WERE ACQUITTED ON THURSDAY OF CHARGES OF BEATING AN INNOCENT IRAQI TEENAGER TO DEATH WITH RIFLE BUTTS. A JUDGE AT A SPECIALLY CONVENED MILITARY COURT IN EASTERN ENGLAND ORDERED THE ADJUDICATING PANEL TO RETURN 'NOT GUILTY' VERDICTS AGAINST THE SEVEN BECAUSE HE DID NOT BELIEVE THERE WAS SUFFICIENT EVIDENCE AGAINST THEM, THE MINISTRY OF DEFENCE SAID. . . .

BRITISH MURDERERS IN IRAQ ACQUITTED – THE JUDGE AT A COURT-MARTIAL ON THURSDAY DISMISSED MURDER CHARGES AGAINST SEVEN SOLDIERS, FROM THE 3RD BATTALION, THE PARACHUTE REGIMENT, WHO'RE ACCUSED OF MURDERING IRAQI TEENAGER; CLAIMING THERE'S INSUFFICIENT EVIDENCE TO SECURE A CONVICTION, THE ASSOCIATED PRESS REPORTED THURSDAY. . . .

Detecting News Reporting Bias

- The task:
 - Given a news story, are we able to say from which news source it came?
- We compared **CNN** and **Aljazeera** reports about the same events from the war in Iraq
 - ...300 aligned articles describing the same story from both sources
- The same topics are expressed in both sources with the following keywords:
 - CNN with:
 - **Insurgents**, Troops, Baghdad, Iran, **Militant**, Police, **Suicide**, **Terrorist**, United, National, Hussein, **Alleged**, Israeli, Syria, Terrorism...
 - Aljazeera with:
 - Attacks, Claims, **Rebels**, Withdrawing, Report, **Fighters**, President, **Resistance**, Occupation, Injured, Army, Demanded, Hit, Muslim, ...

Event Representation

<http://eventregistry.org/>

Feature vector event representation

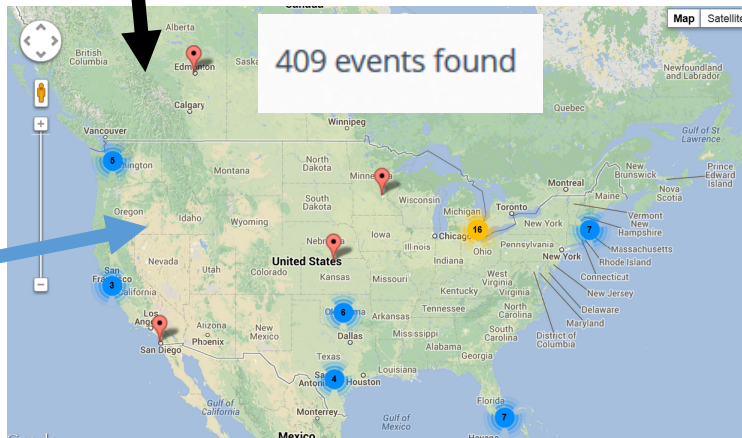
- Feature vectors easily extractable from news documents:
 - **Topical dimension** – what is being talked about? (keywords)
 - **Social dimension** – which entities are mentioned? (named entities)
 - **Temporal aspect** – what is the time of an event? (temporal distribution)
 - **Geographical aspect** – where an event is taking place? (location)
 - **Publisher aspect** – who is reporting? (publisher identifiers)
 - **Sentiment/bias aspect** – emotional signals (numeric estimates)
- Scalable Machine Learning techniques can easily deal with such representation
 - ...in “Event Registry” system we use this representation to describe events

Example of “feature vector” event representation: Event Registry “Chicago” related events

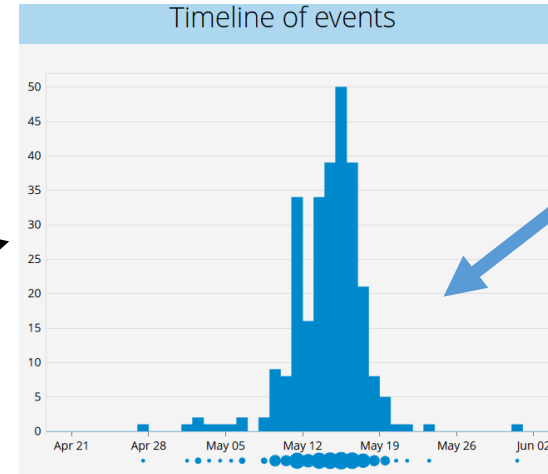
Query: “Chicago”

409 events found

Emanuel says he'll seek 2nd term as Chicago mayor
 WHEN: May 8, 2013
 ARTICLES: 9 (eng: 9, deu: 0, spa: 0, zho: 0, slv: 0)
 ENTITIES: Rahm Emanuel, Chicago, African American, Mayor, Richard M. Daley, President of the United States, Chicago Tribune, Homicide, Chicago Sun-Times
 KEYWORDS: School, Voting, Percentage, Question, Richard Wilkins (Buffy the Vampire Slayer), Race and ethnicity in the United States Census, Justice, Mayor of Chicago, City, United States presidential approval rating



Where? (geography)



When? (temporal distribution)

Event is about...

ENTITIES

Rahm Emanuel	100
Chicago	92
African American	69
Mayor	66
Richard M. Daley	66
Barack Obama	40
President of the United States	34
Chicago Tribune	34
Homicide	32
Chicago Sun-Times	27

KEYWORDS

School	32
Voting	32
Percentage	25
Question	24
Richard Wilkins (Buffy the Vampire Slayer)	23
Race and ethnicity in the United States Census	23
Justice	23
Mayor of Chicago	22
City	21
United States presidential approval rating	21

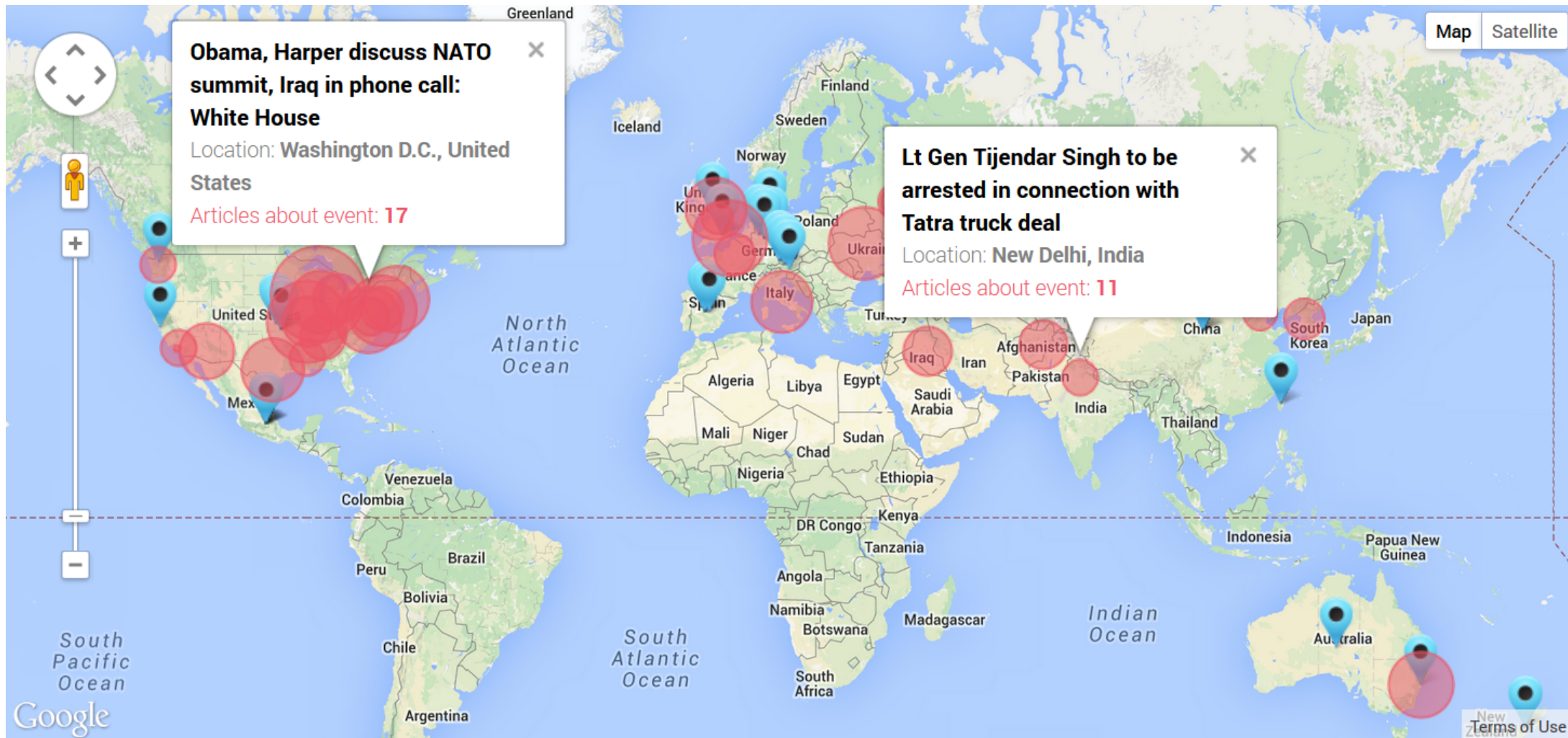
Who? (named entities)

What? (keyword/topics)

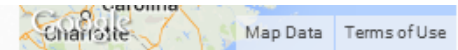
Event Visualization

<http://eventregistry.org/>

Live Event tracking with <http://EventRegistry.org/>



Representatives from tech companies are meeting with white House staff on Friday.



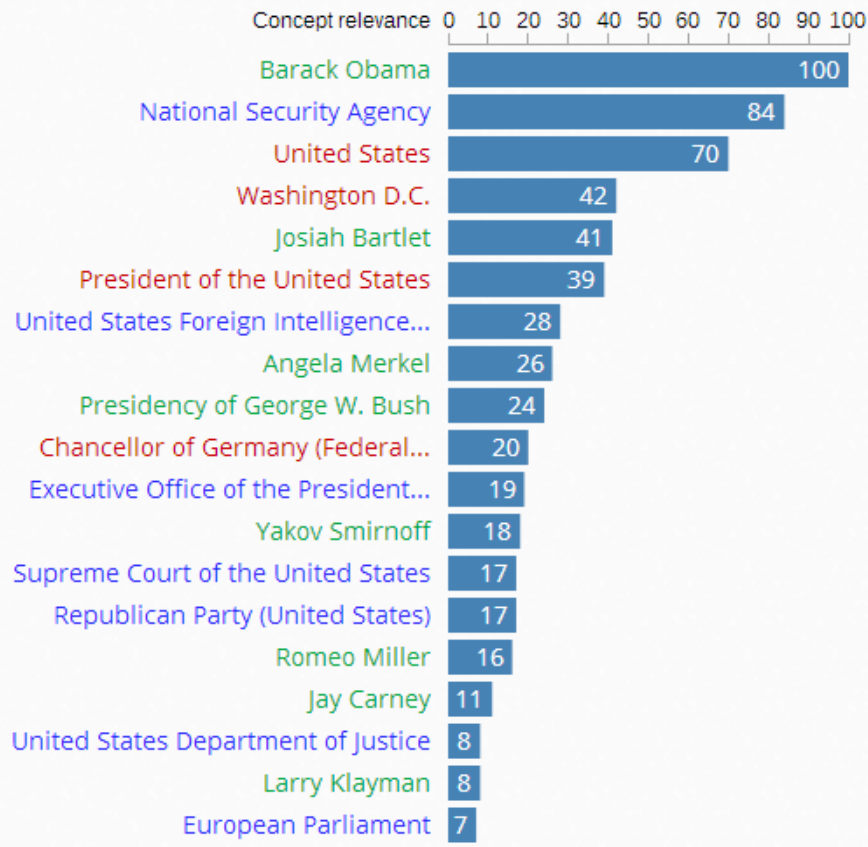
The White House...

Nr. of articles: 105 (89 eng, 5 ger, 11 spa, 0 chi, 0 slo)

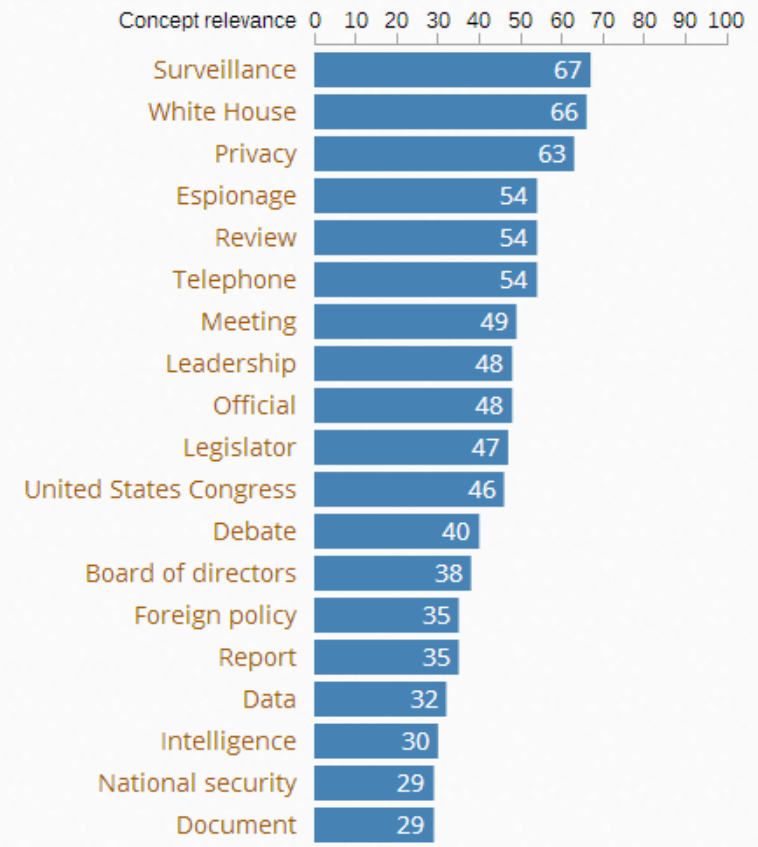
Event description through entities and Semantic keywords



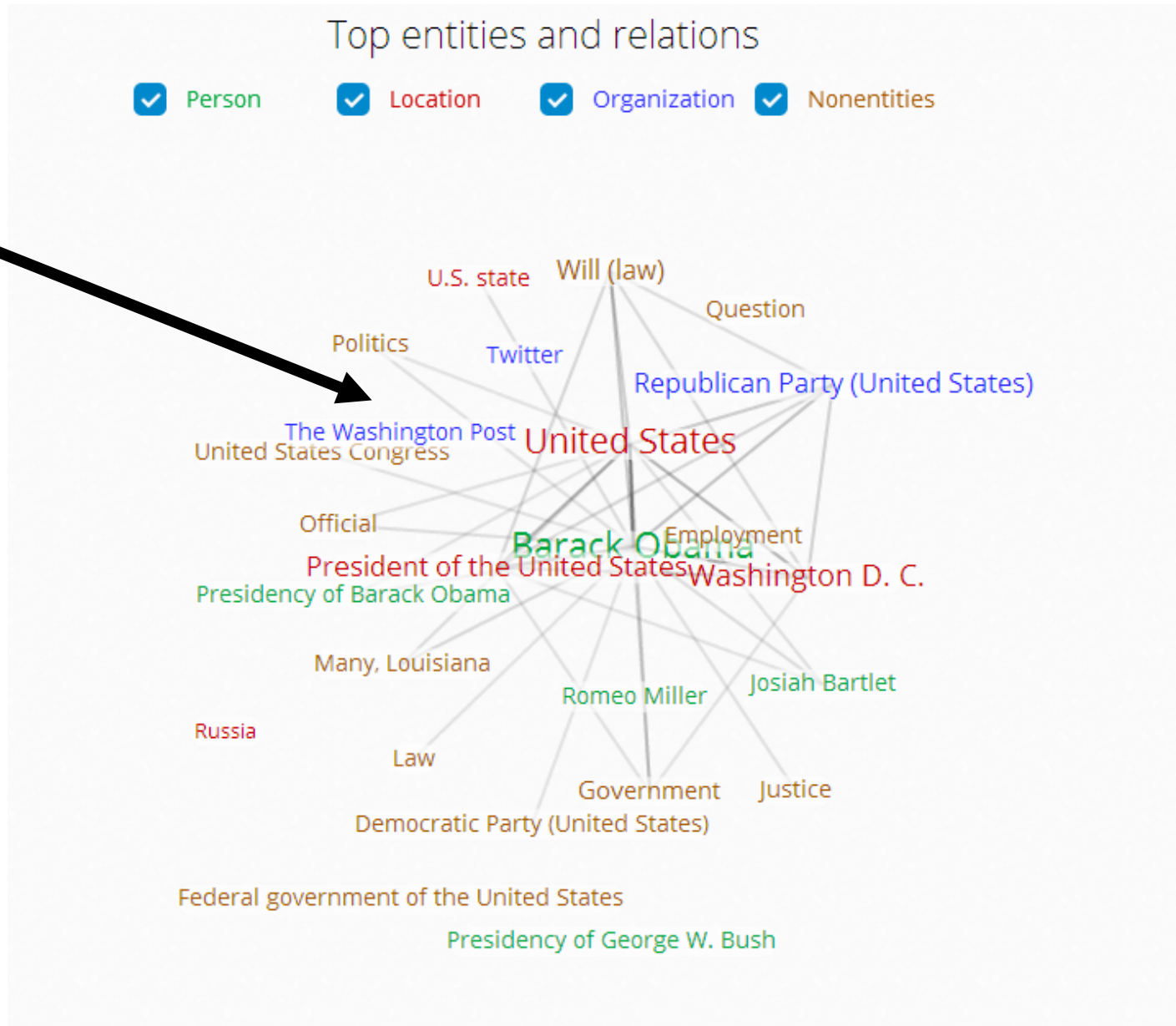
ENTITIES



KEYWORDS



Collection of events described through Entity relatedness



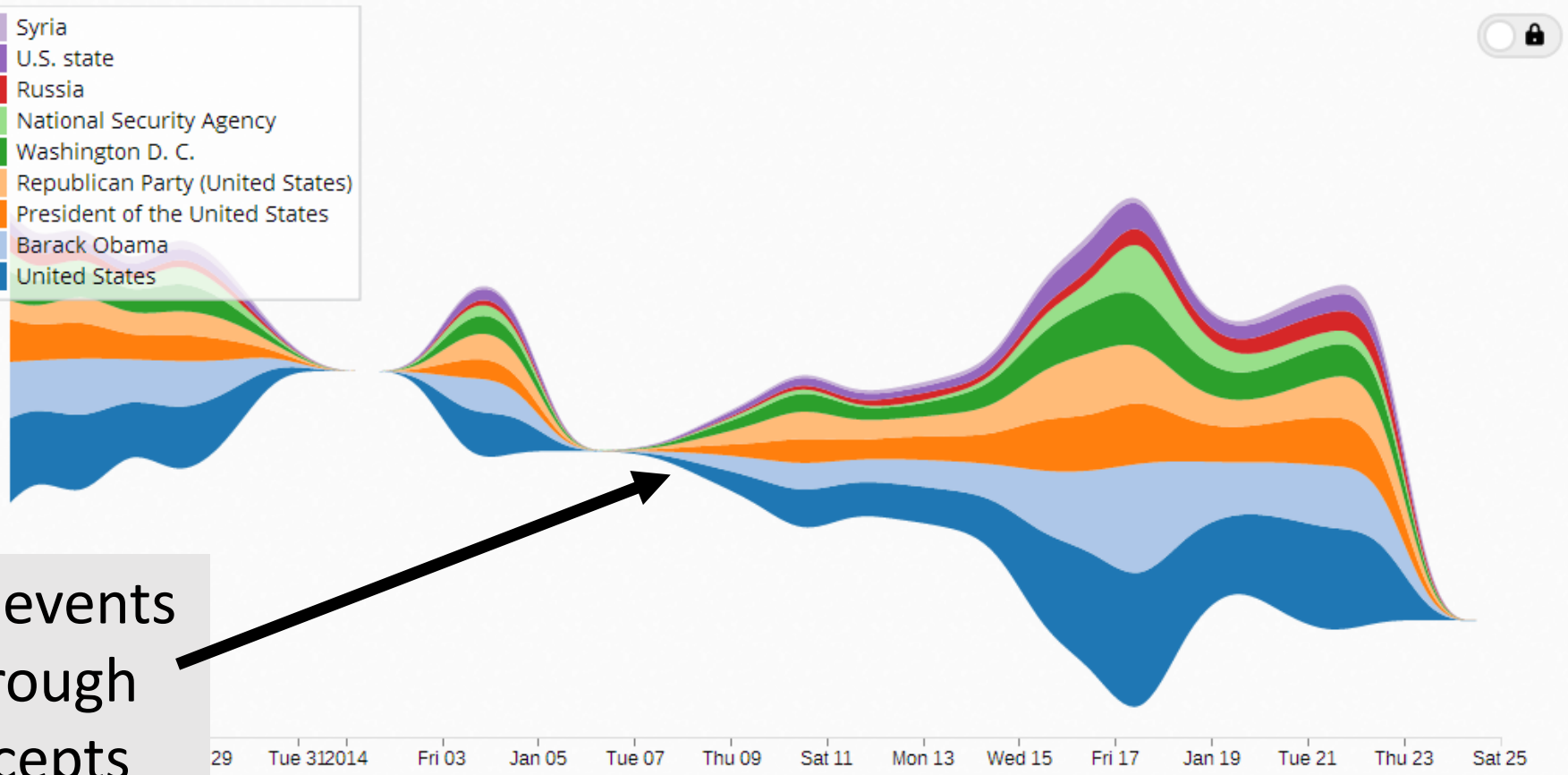
Trending of concepts in search results

Person Location Organization Nonentities

Normalize by article count

United States Barack Obama President of the United States Republican Party (United States) Washington D. C. National Security Agency
Russia U.S. state Syria

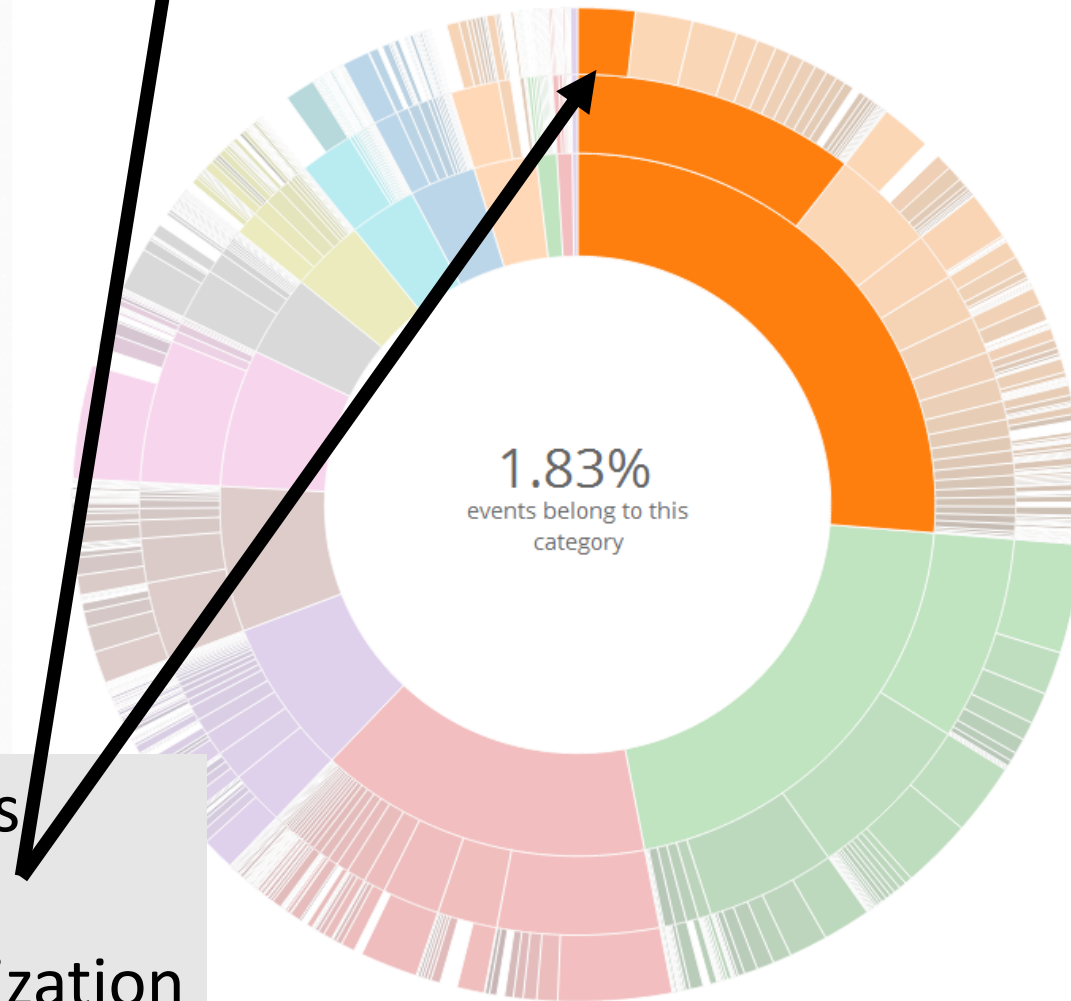
- Syria
- U.S. state
- Russia
- National Security Agency
- Washington D. C.
- Republican Party (United States)
- President of the United States
- Barack Obama
- United States



Collection of events described through trending concepts

Categories of events

Society > Issues > Labor 1.83%



Collection of events
described through
three level categorization

Obama ponders limiting NSA access to phone records

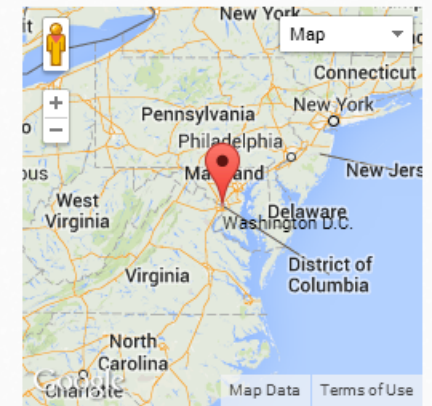
15 9 Jan 2014

Washington D.C., United States

Washington -- President Barack Obama is expected to rein in spying on foreign leaders and is considering restricting National Security Agency access to the government's surveillance program. Obama could unveil his highly anticipated decisions as early as next week. On Thursday, the president is expected to discuss his review with congressional lawmakers, while his top lawyer plans to meet with privacy groups. Representatives from tech companies are meeting with White House staff on Friday.

The White House...

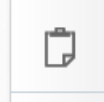
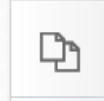
Events identified across languages



Nr. of articles: 105 (89 eng, 5 ger, 11 spa, 0 chi, 0 slo)



Articles



Obama ponders limiting N

Washington -- President Barack Obama is expected to restrict National Security Agency access to the government's surveillance program. Obama could unveil his highly anticipated decisions as early as next week. On Thursday, the president is expected to discuss his review with congressional lawmakers, while his top lawyer plans to meet with privacy groups. Representatives from tech companies are meeting with White House staff on Friday.

89 eng 11 spa 5 ger

Jan. 9, 12:51
WWW.DETROITNEWS.COM

view with

Obama likely to adjust US spying on foreign leaders, nearing decision on intelligence changes

WASHINGTON - President Barack Obama is expected to restrict National Security Agency access to Americans' phone records and rein in spying on foreign leaders, according to people familiar with a White House review of the government's surveillance programs. Obama could unveil his highly anticipated decisions as early as next week. On Thursday, the president is expected to discuss his review with congressional lawmakers, while his top lawyer plans to meet with privacy groups. Representatives from tech companies are meeting with White House staff on Friday.

Jan. 9, 07:57
CKGL AM 570

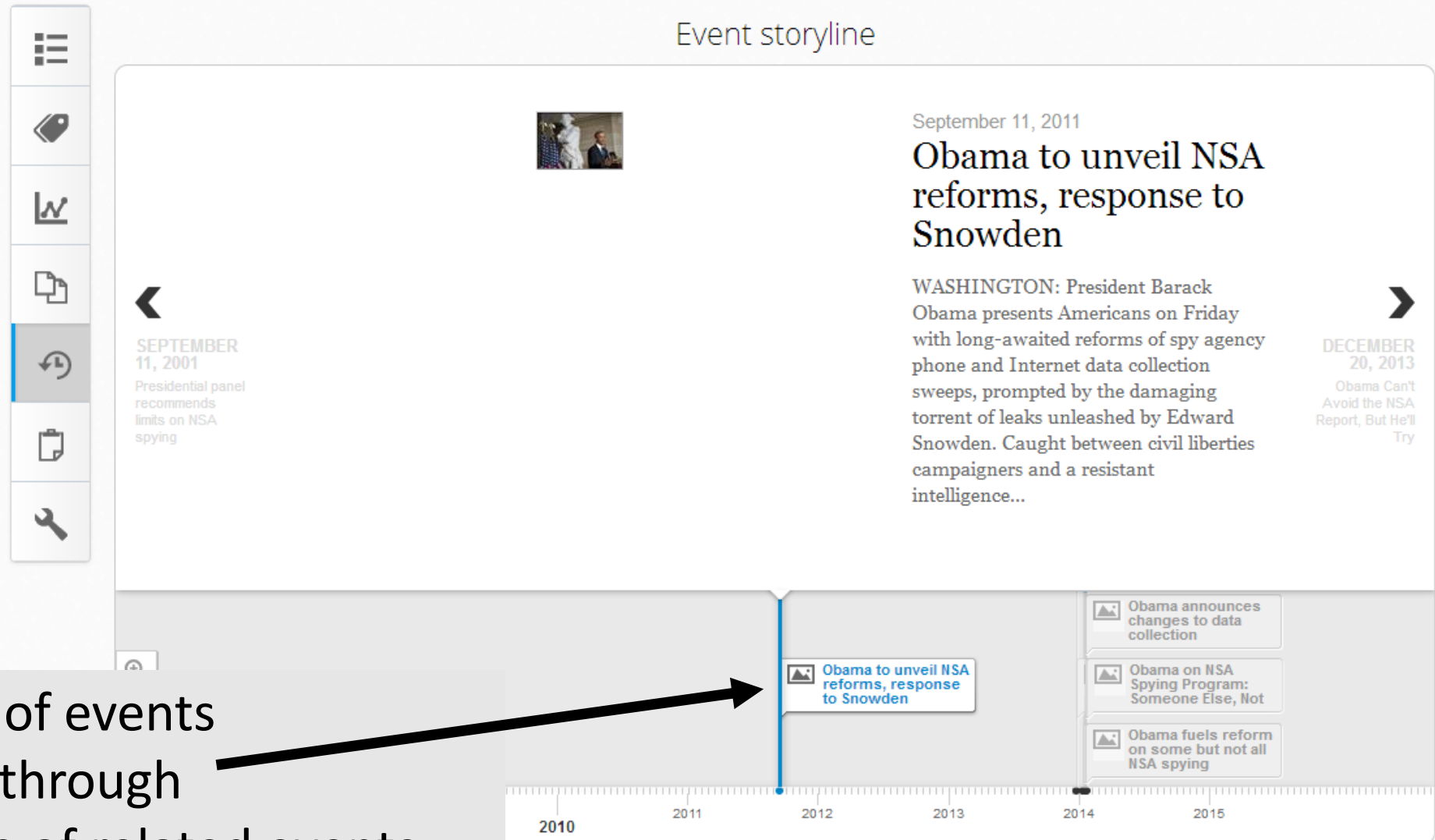
Obama might limit how much the NSA can spy on you

President Barack Obama speaks at a press conference on Dec. 20, 2013. (Pablo Martinez Monsivais/AP)

Jan. 10, 14:04
WWW.IMPACTOUSA.COM

Nr. of articles: 105 (89 eng, 5 ger, 11 spa, 0 chi, 0 slo)

Event storyline



Collection of events described through a story-line of related events

Event Registry API

...to be used as a research platform

Searching for events using Python

```
>>> from EventRegistry import *
>>> er = EventRegistry()
>>> q = QueryEvents()
# get events related to Barack Obama
>>> q.addConcept(er.getConceptUri("Obama"))
# and are related to issues in society
>>> q.addCategory(er.getCategoryUri("society issues"))
# and have been reported by the BBC
>>> q.addNewsSource(er.getNewsSourceUri("bbc"))
# return event details for first 30 events
>>> q.addRequestedResult(RequestEventsInfo(page = 0, count = 30))
# execute query and obtain results
>>> res = er.execQuery(q)
```

Result of the query

```
'events': { 'resultCount': 122,  
  'results': [  
    { 'articleCounts': { 'eng': 54.0, 'total': 54.0 },  
      'categories': [{...}],  
      'concepts': [{...}, ...],  
      'eventDate': '2014-08-29',  
      'eventDateEnd': '',  
      'multiLingInfo': { 'eng': {  
        'title': ..., 'summary': ... }},  
      'uri': '1211229', 'wgt': 9.0 }  
  ], ...  
}]
```

Future / Follow-up projects

What next?

- Understanding global social dynamics
 - How global society functions?
- Integrating text-based media with TV channels
 - ...requires speech recognition, video processing, visual object recognition, face recognition, ...
- Event prediction / Event-Consequence prediction
 - ...requires understanding of causality in the social dynamics and much more
- Micro-reading / Machine-reading
 - ...full understanding of individual documents – the goal for 10+ years

Who is using “Event Registry”?

- Bloomberg
- IBM Watson
- Stratfor
- Datarama (risk management)
- Robinhood (personal investments)
- IDV solutions (risk management)
- Oxford University Press (dictionaries)
- Sobey Digital Technology Co.,Ltd. (TV)
- Qatar News Agency



Datarama.

Bloomberg



IBM Watson



**QATAR
news**

- Taykey
- udu, Inc.
- 911 International LLC
- PM, poslovni mediji d.o.o.
- Argo Group
- Social Strategy, LLC
- OneQube
- Wildtrails Technologies
- Big-datext.com
- ITRI
- Fanlig
- Wholecrowd
- Colorado State University
- SYVERSE LTD

Some statistics

- Over 200 million main-stream news articles imported
 - ...in 100+ languages
- 8000 events identified per day
 - ...archive of over 7 million interconnected events
- Over 40 million search requests in the last year
- 120+ thousands unique users

Systems/Demos used within the presentation

- NewsFeed (<http://newsfeed.ijs.si/>)
 - News and social media crawler
- Enrycher (<http://enrycher.ijs.si/>)
 - Complete Linguistic Stack as a web service
- Wikifier (<http://wikifier.org/>)
 - Language and Semantic annotation
- XLing (<http://xling.ijs.si/>)
 - Cross-lingual document linking and categorization
- Event Registry (<http://eventregistry.org/>)
 - Event detection and topic tracking

