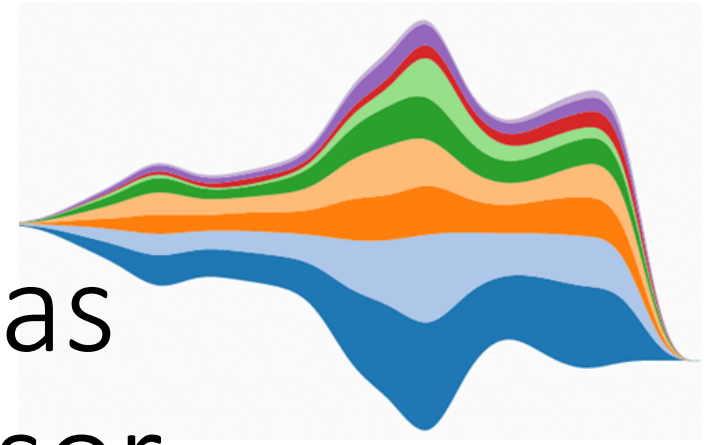


Cross-Linguality as Global Social Sensor

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

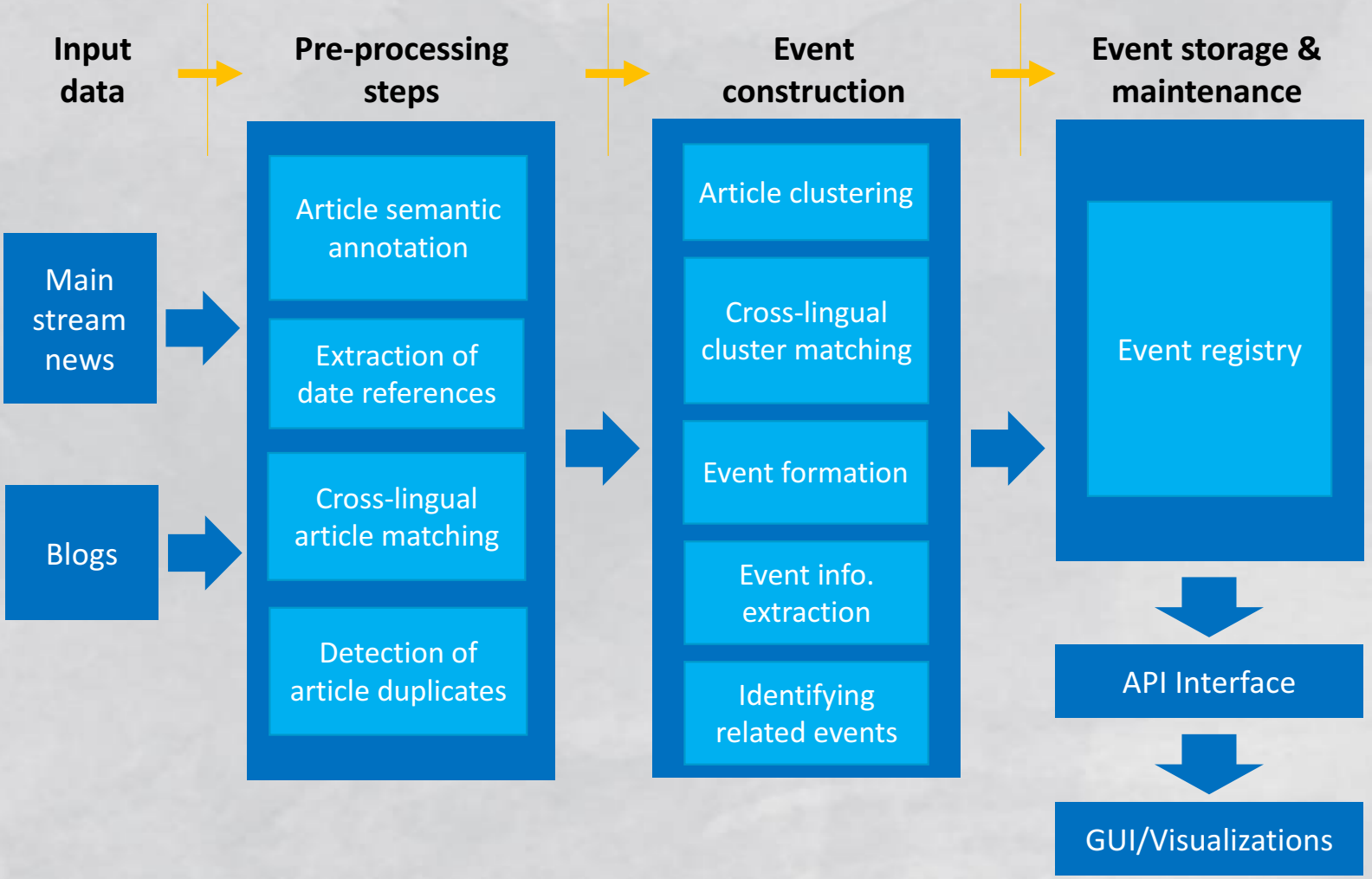
What questions are we addressing?

- 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/

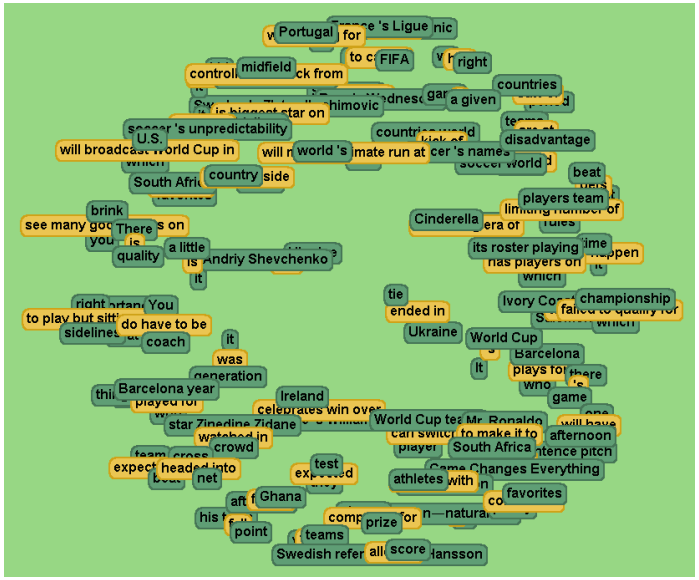
Collecting global media data

- 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 Derniere Heure
  - La Libre Belgique
  - La Premiere
  - 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/

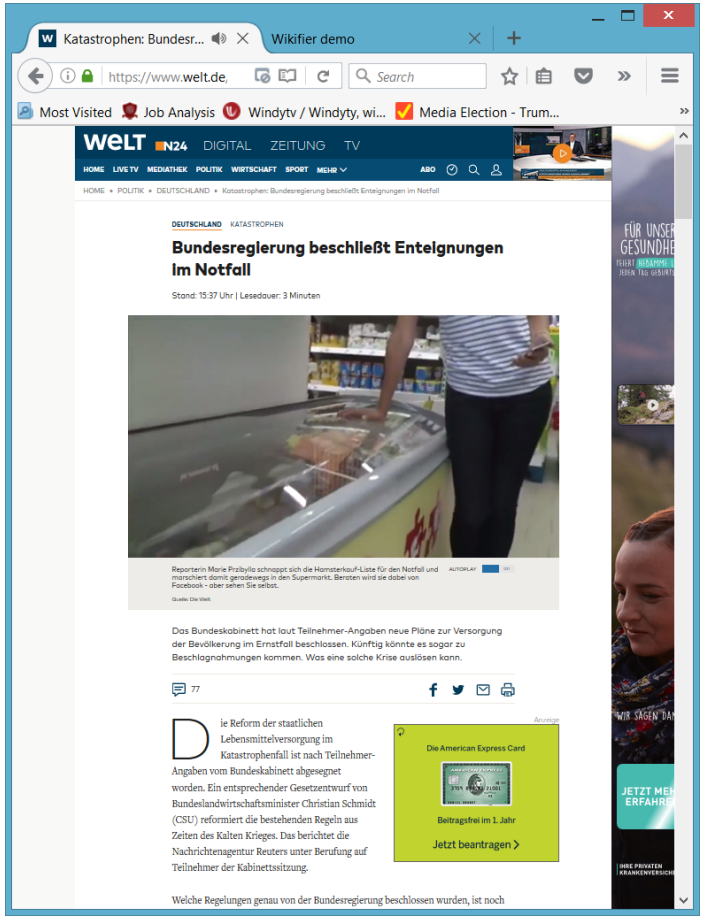
What can extracted from a document?

- 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

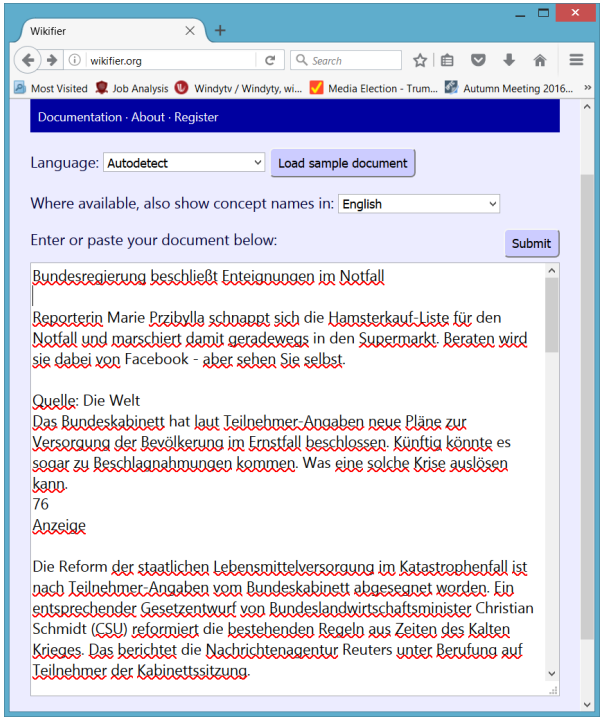
Semantic annotation with Wikifier.Org

(operates for 100 languages)

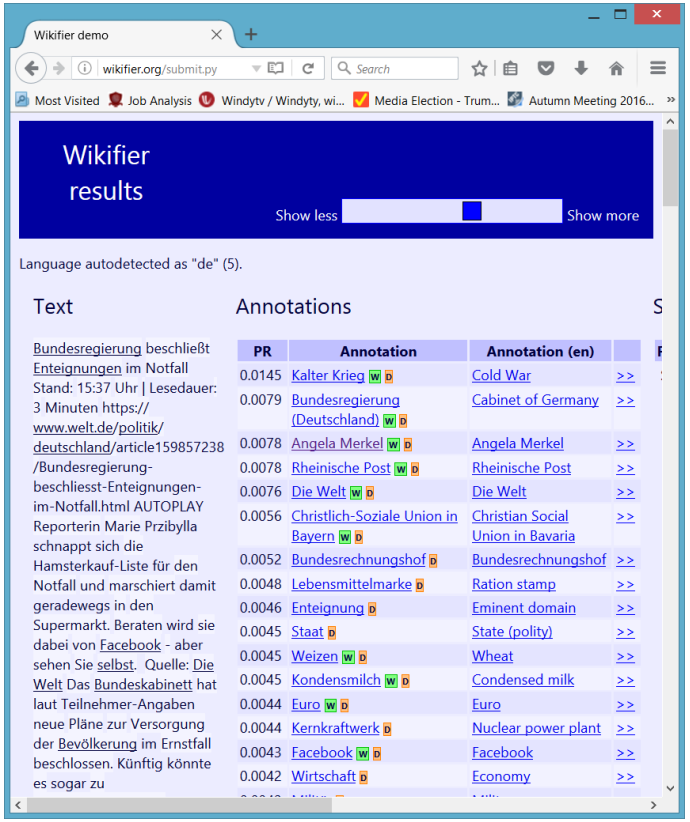
Original text



Plain clean text



Annotated text in original language and English



Key problem in semantic annotation (wikification) is concept disambiguation

process "that is the best opportuni
achieve the goals of the Syrian pe
Mr Kerry
opposition and the regime" would
in any future transitional governm
coalition told reporters earlier this
warned privately that Britain and t
support if it did not attend the pea

Many Candidates

Target page
Kerry GAA
County Kerry
John Kerry
Kerry (UK Parliament constituency)
Kerry, Powys
Kerry (Dáil Éireann constituency)
Kerry (Parliament of Ireland constituency)
Kerry Von Erich
Kerry Airport
Kerry, Queensland
Kerry County Council
Roman Catholic Diocese of Kerry
List of Neighbours characters (2006)
Kerry Earnhardt
Kerry Bishop
Kerry Kennedy
Kerry
Kerry Packer
Lloyd Kerry

disambiguation

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

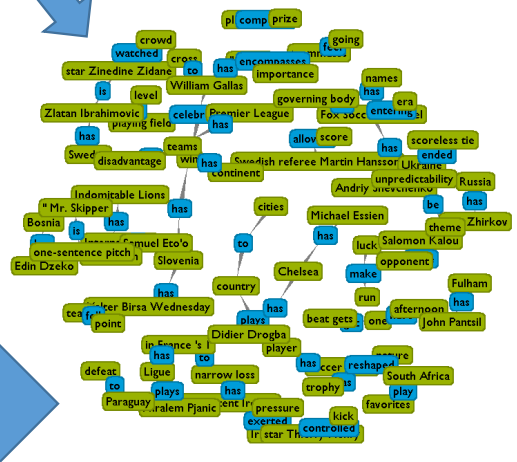
Plain text

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 wor coveted trophy.

Russia's Yuri Zhirkov, right, fights for the ball with Slovenia's 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

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/ArgentinaInternationalFootballers
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

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

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

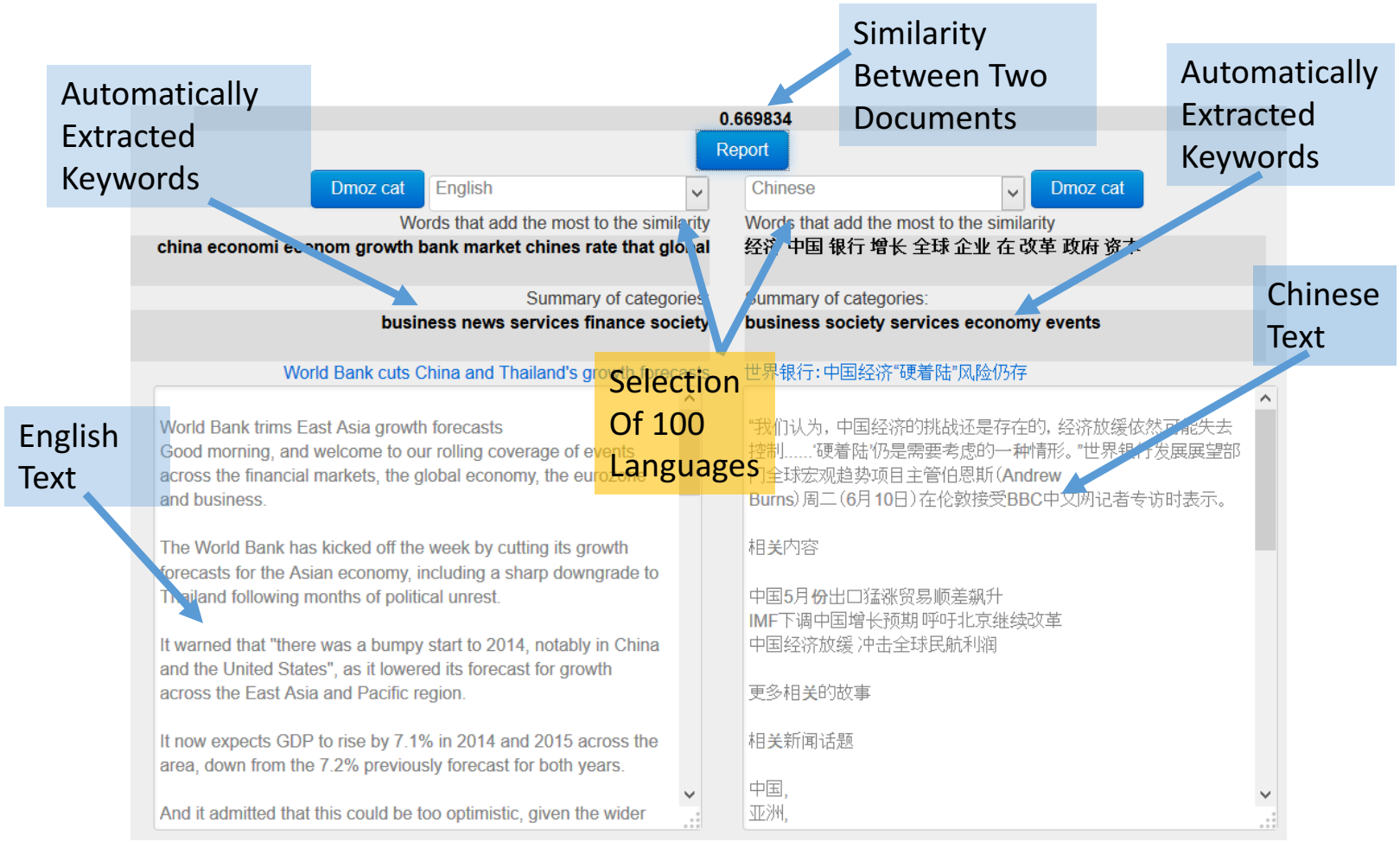
- The key building block is the function for comparing and categorization of documents in different languages
  - http://XLing.ijs.si is an open web service to bridge information across 100 languages

Languages covered by XLing (top 100 Wikipedia languages)

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

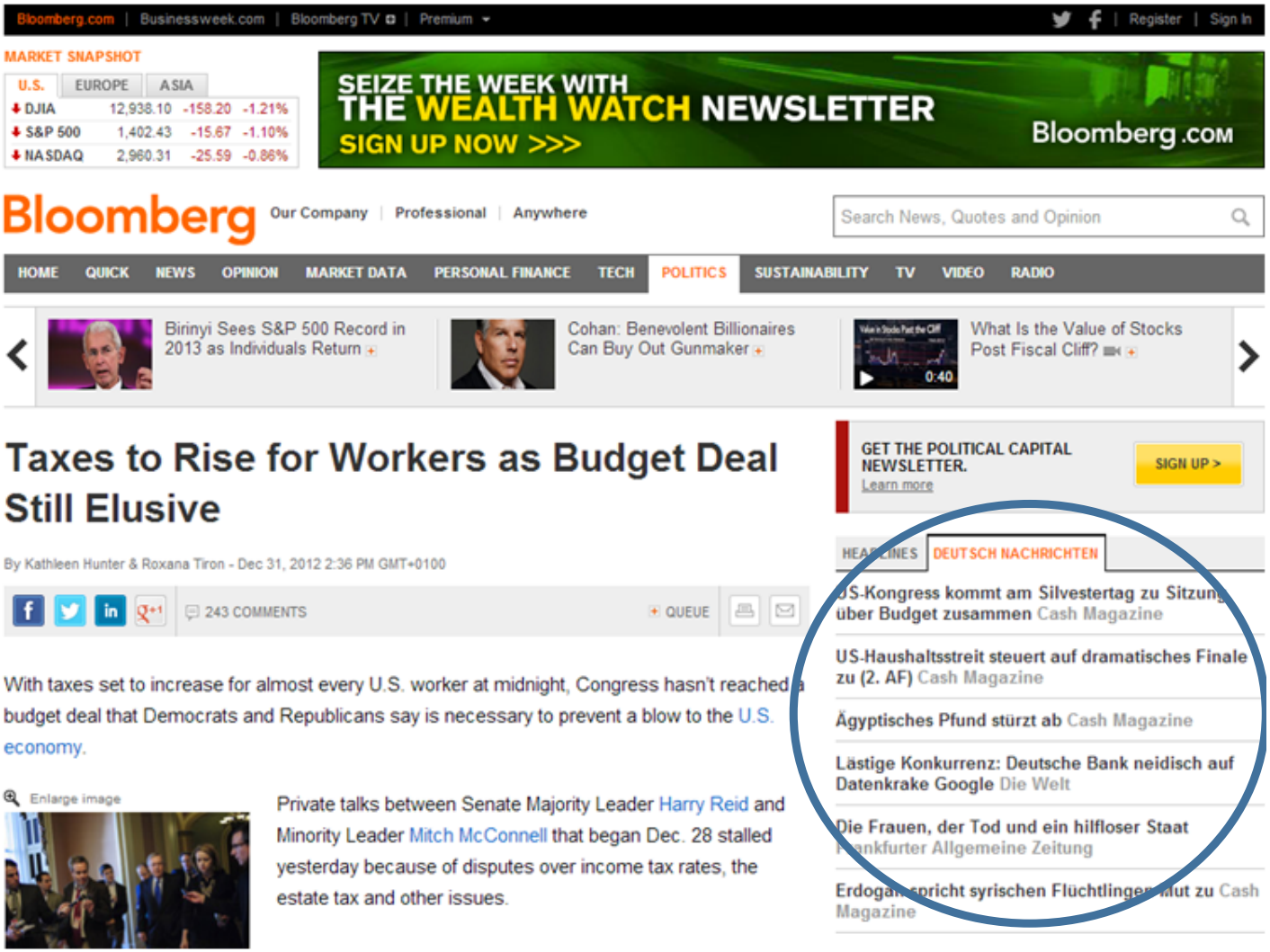
XLing (http://XLing.ijs.si)

service for comparing and categorization of documents across 100 languages



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



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/

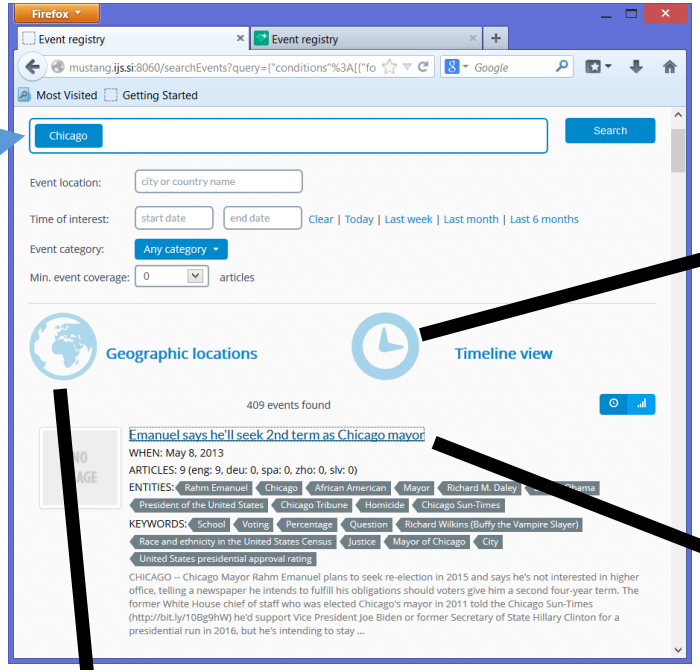
Feature vector event representation

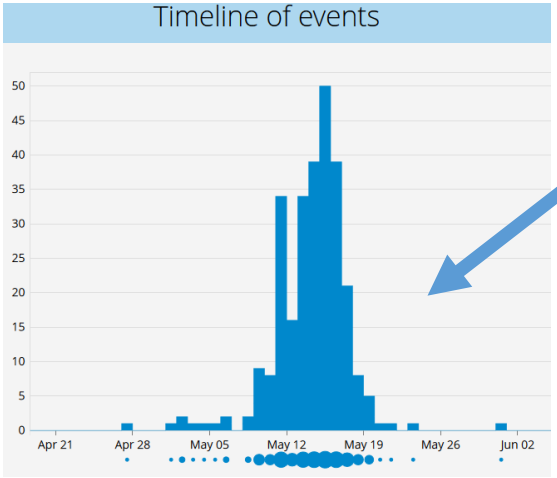
- Feature vectors easily extractable from news documents:
  - **Topical dimension** – what is being talked about? (keywords)
  - **Social dimension** – which entities are mentioned? (named entities)
  - **Temporal aspect** – what is the time of an event? (temporal distribution)
  - **Geographical aspect** – 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

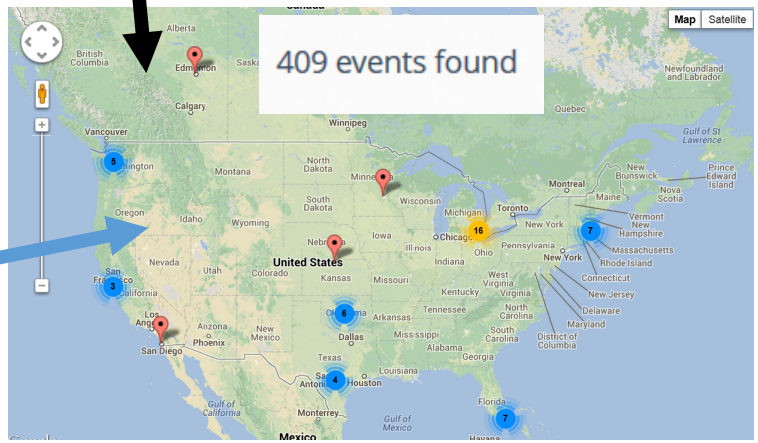
Example of “feature vector” event representation: Event Registry “Chicago” related events

Query:
“Chicago”

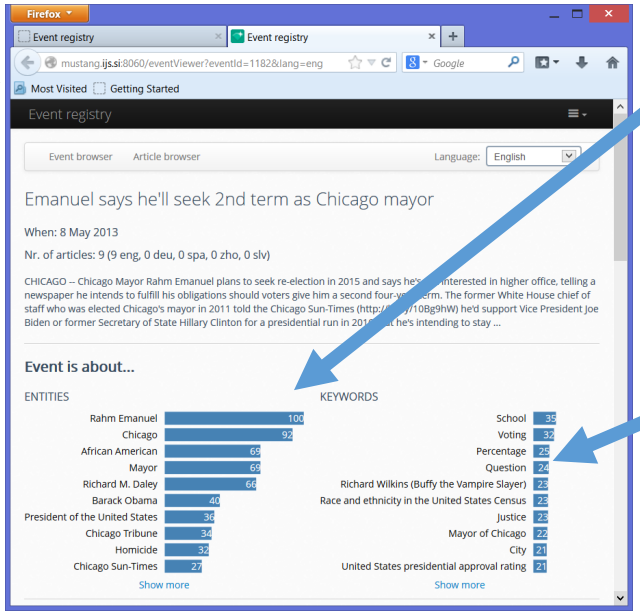




When?
(temporal distribution)



Where?
(geography)



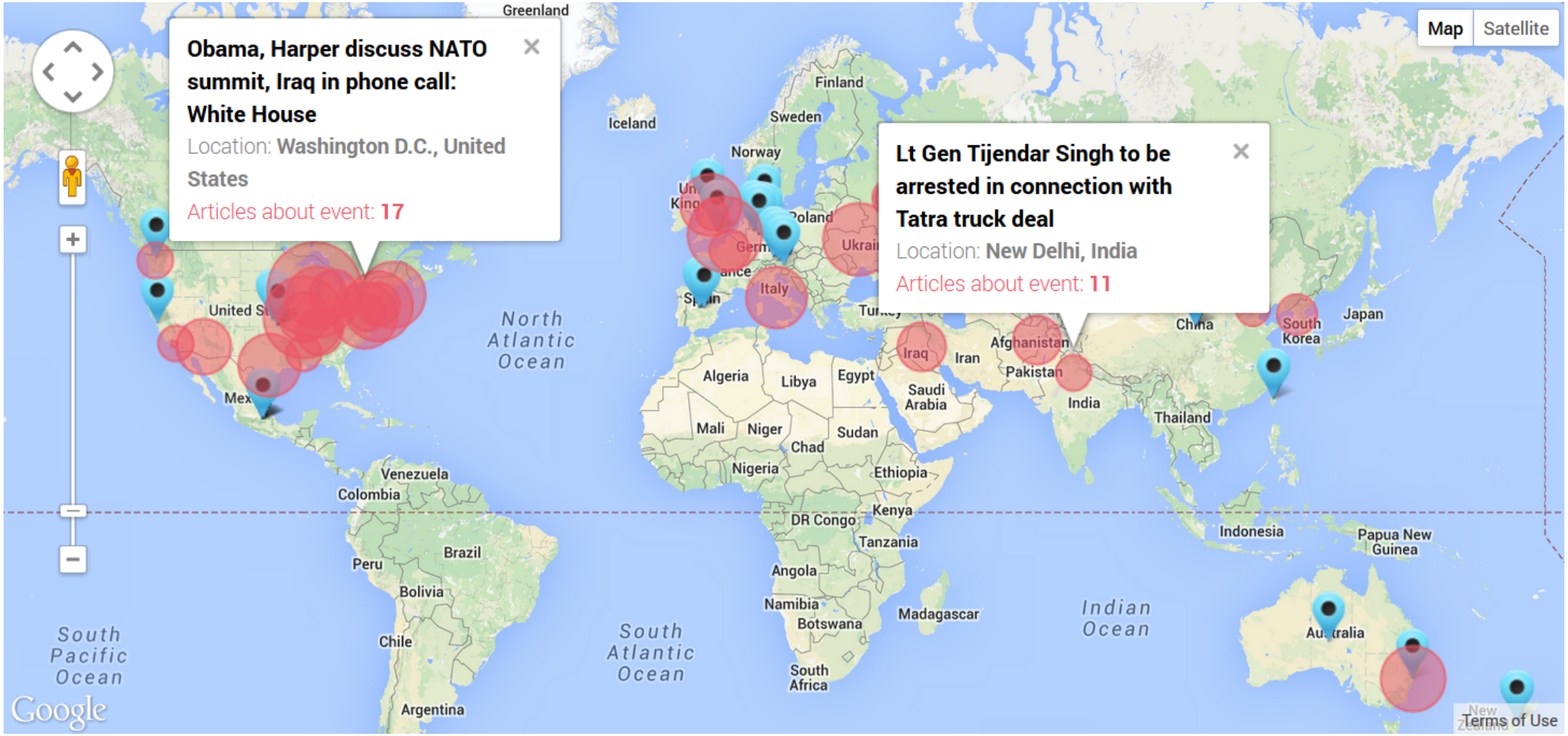
Who?
(named entities)

What?
(keyword/ topics)

Event Visualization

http://eventregistry.org/

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

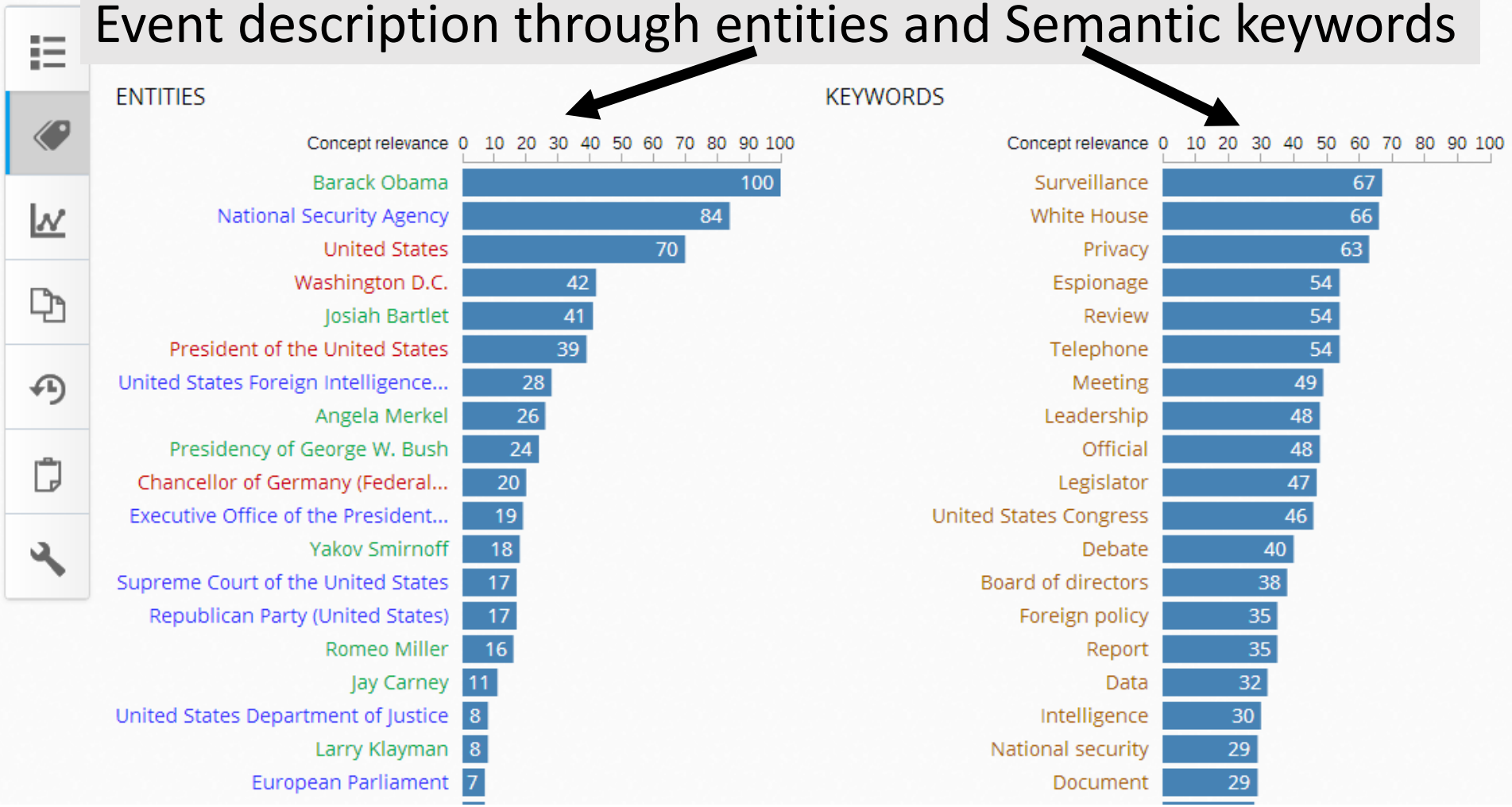


Representatives from tech companies are meeting with White House staff on Friday.

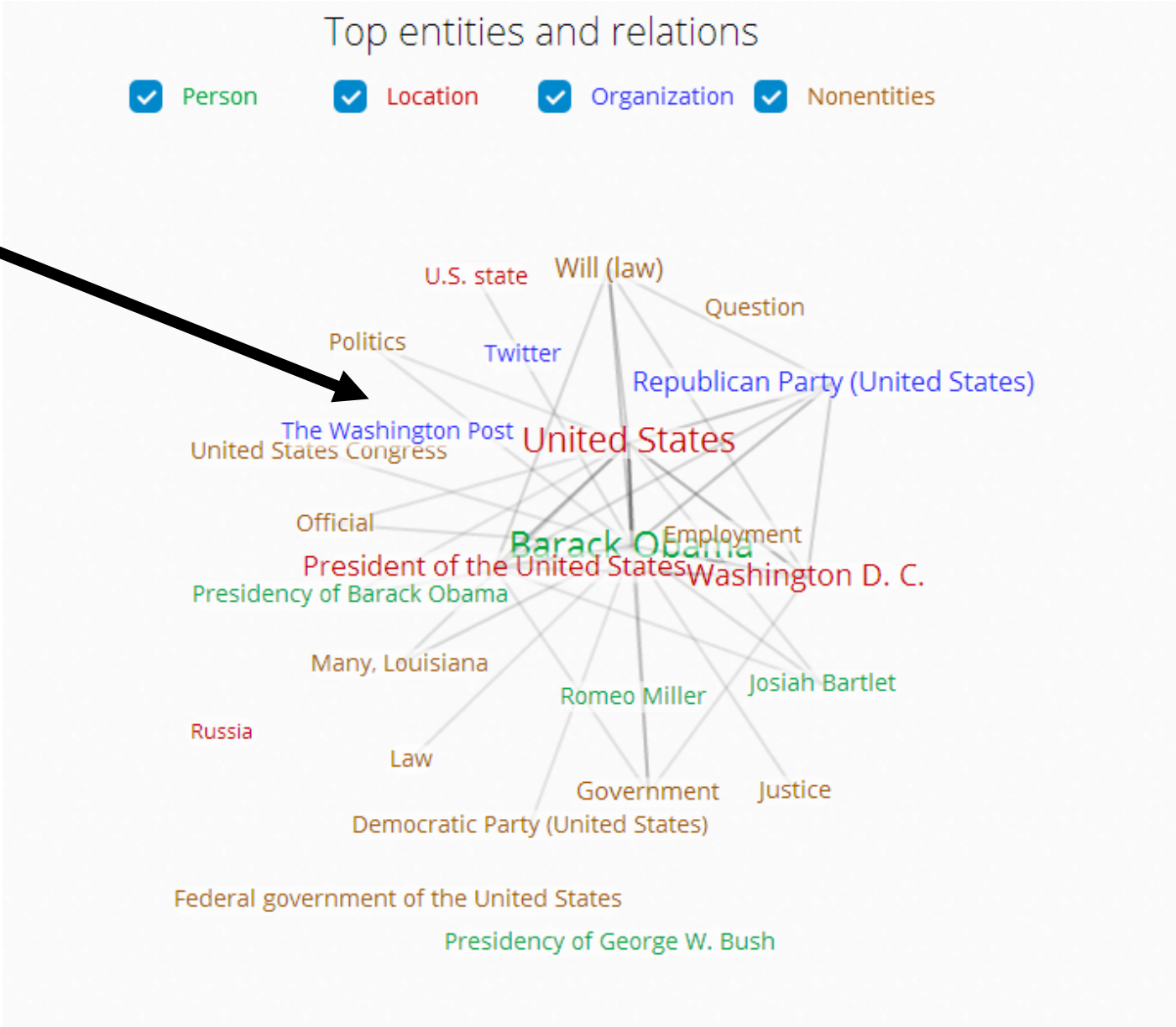
The White House...

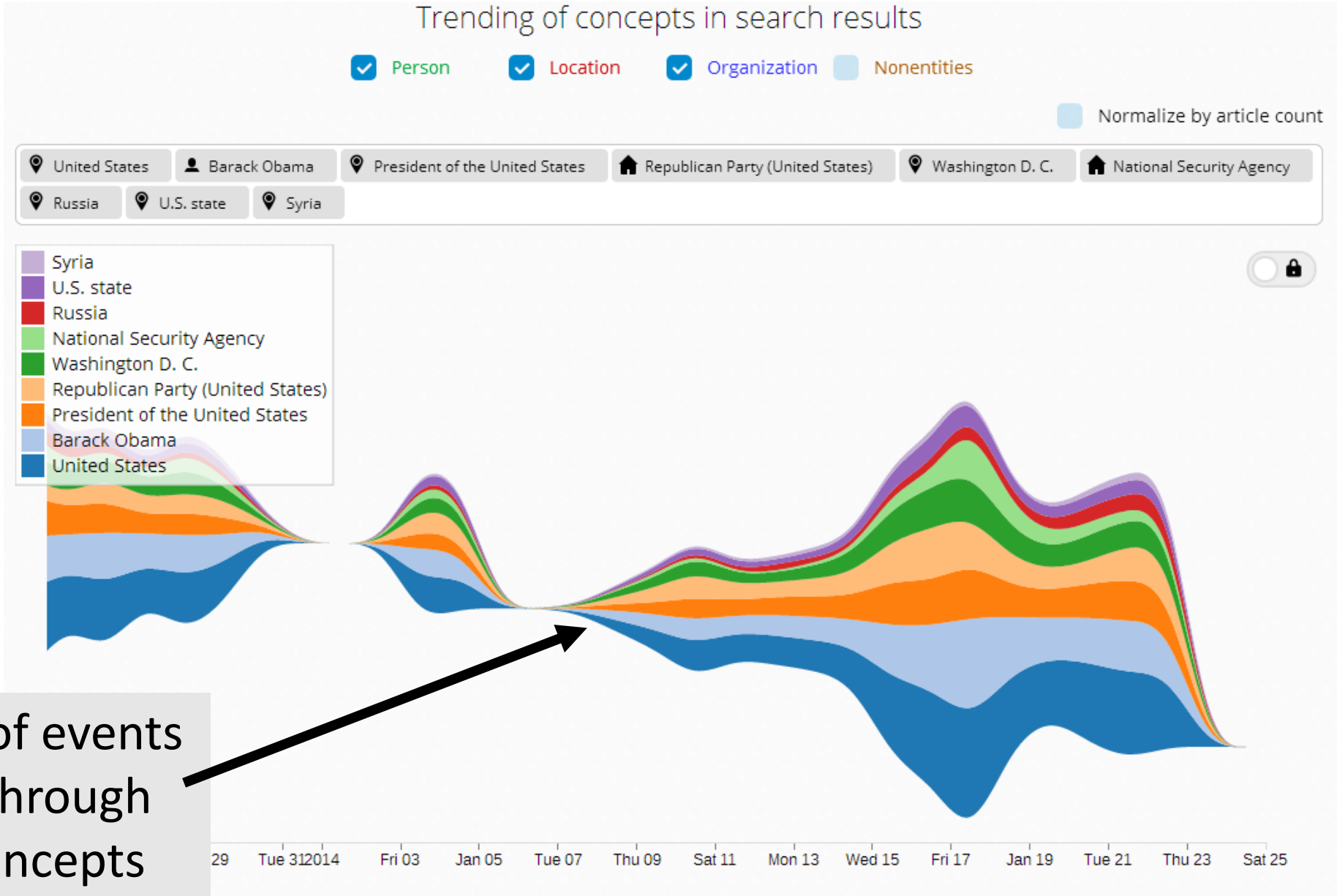
Nr. of articles: 105 (89 eng, 5 ger, 11 spa, 0 chi, 0 slo)

Event description through entities and Semantic keywords

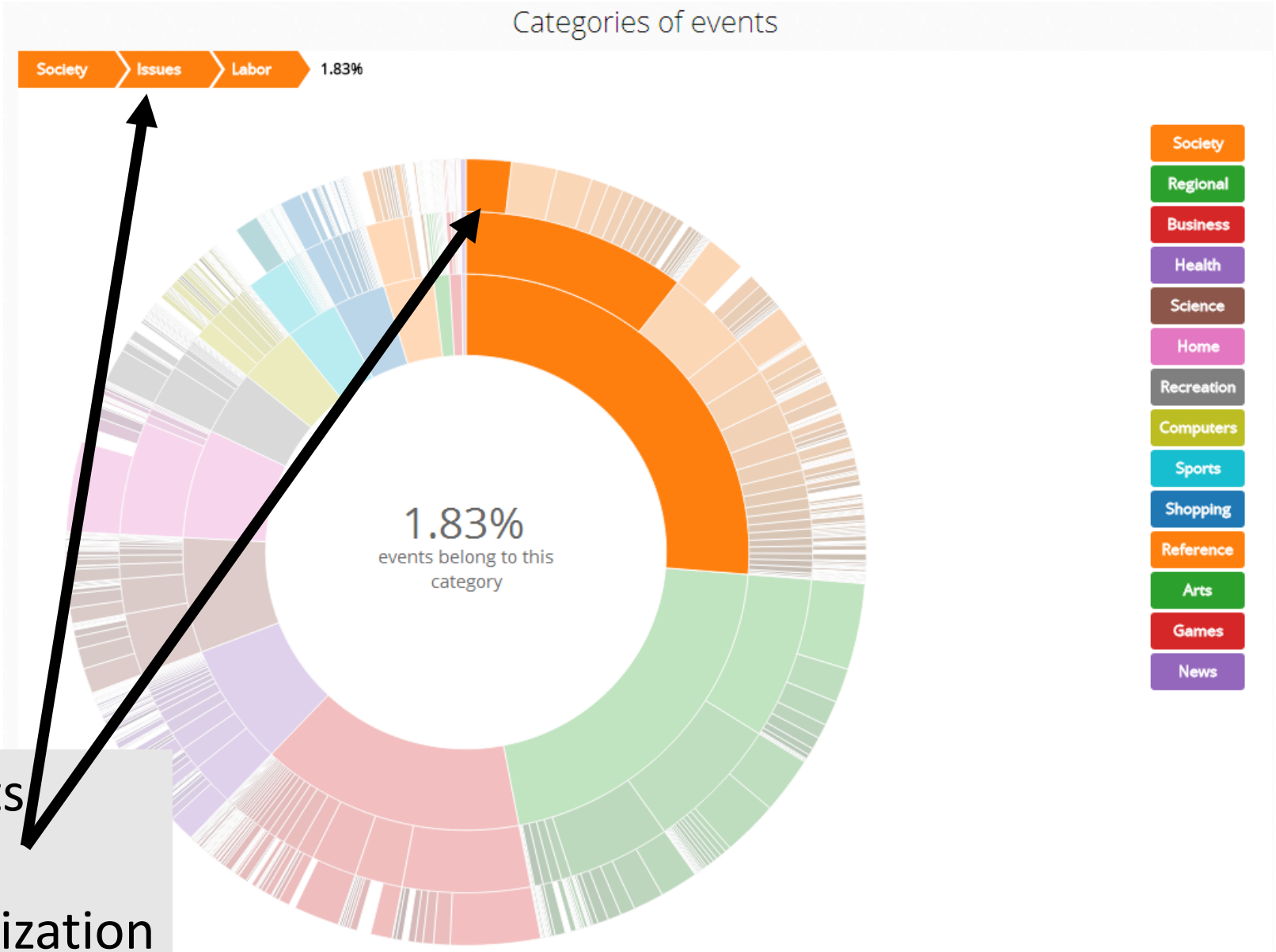


Collection of events described through Entity relatedness

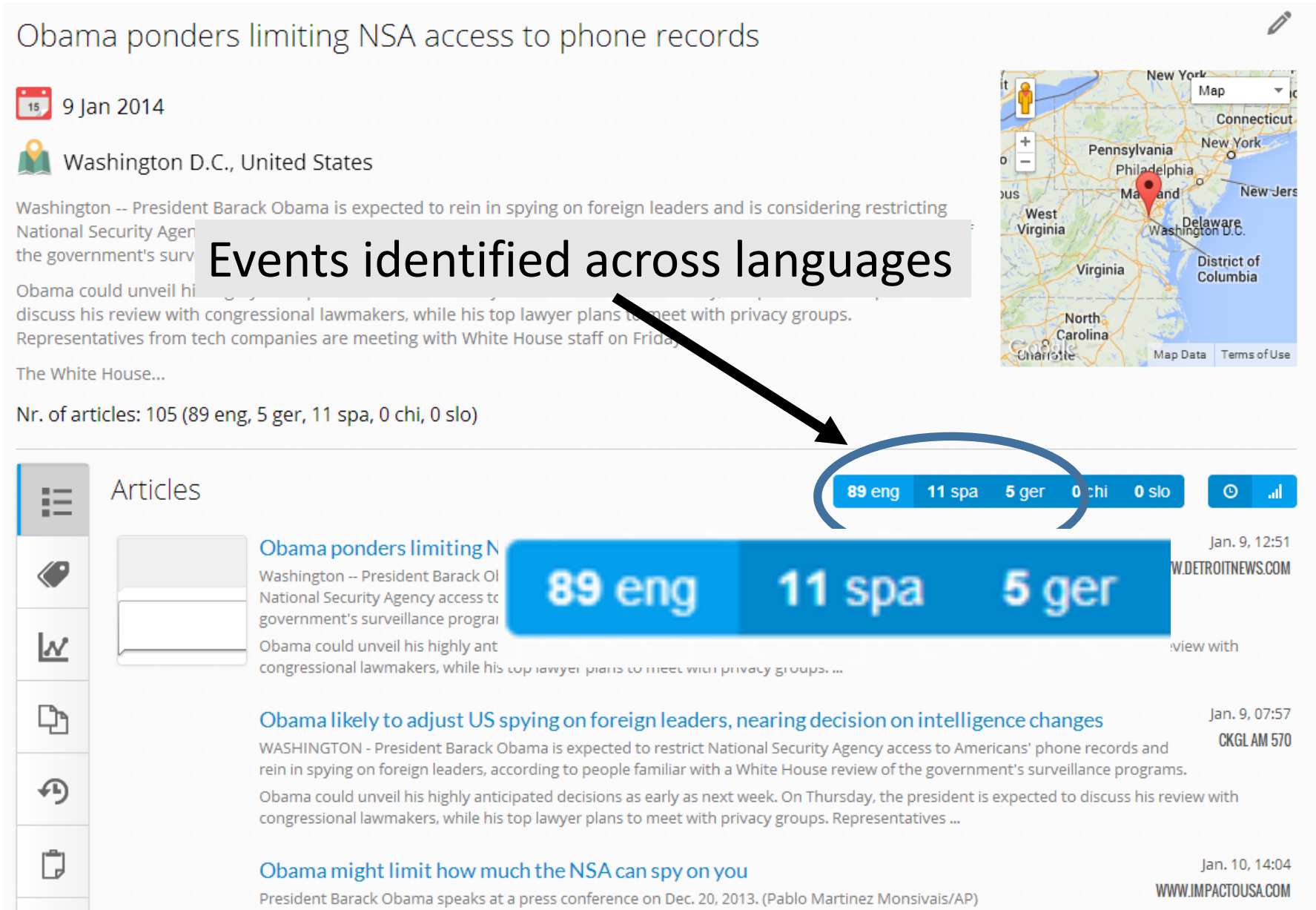


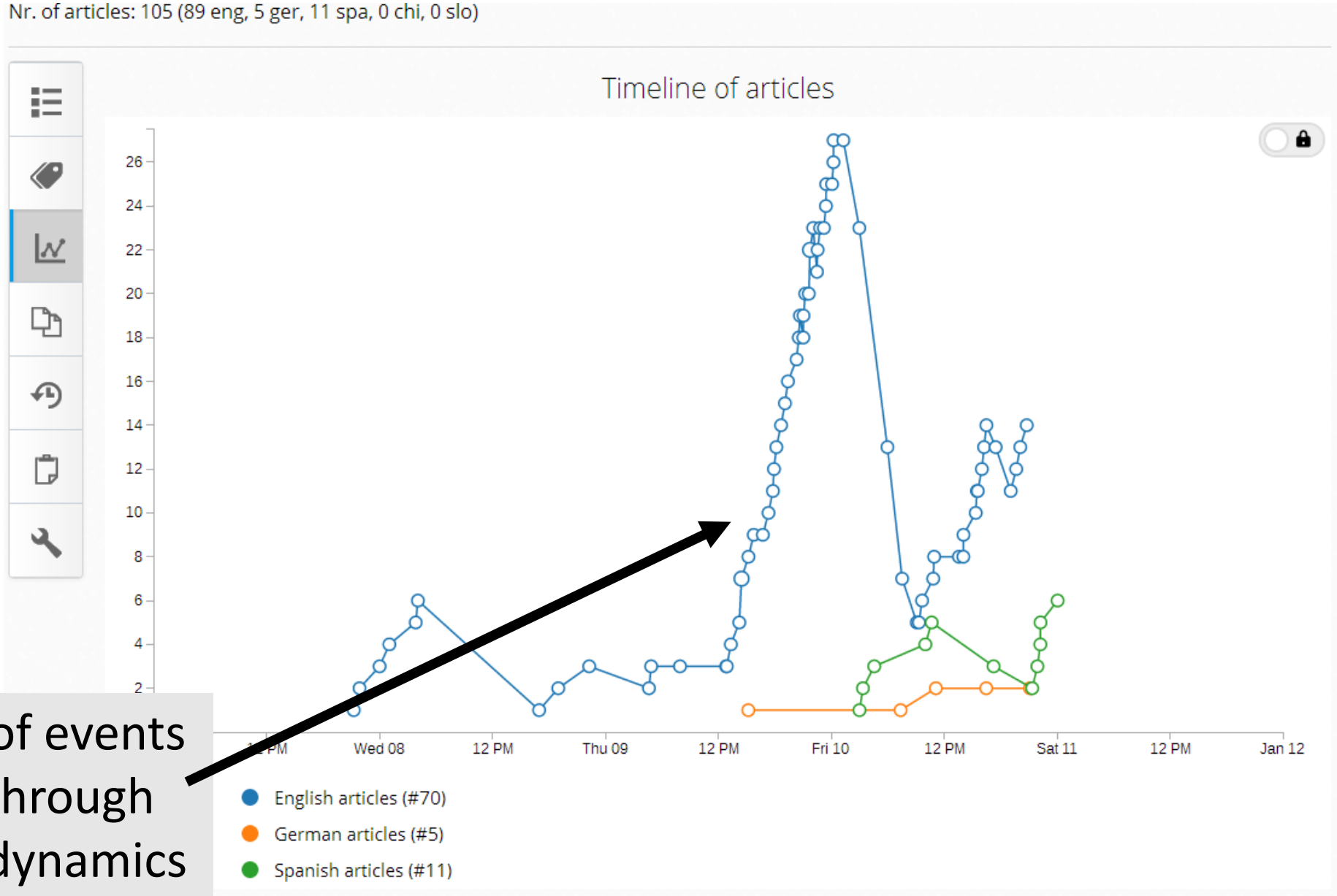


Collection of events described through trending concepts

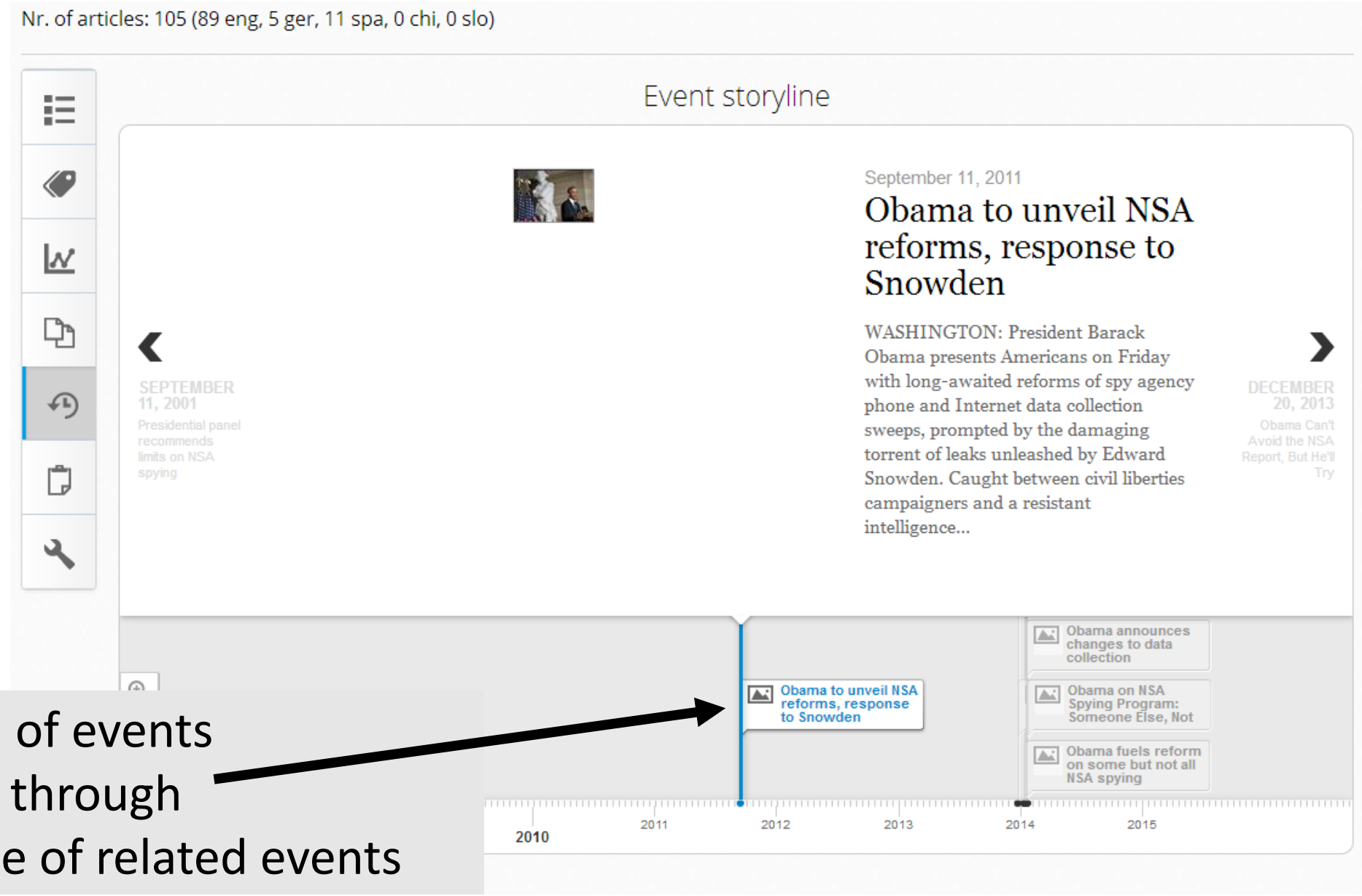


Collection of events described through three level categorization


Obama ponders limiting NSA access to phone records
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 Agen
the government's surv
Obama could unveil hi
discuss his review with congressional lawmakers, while his top lawyer plans t meet with privacy groups.
Representatives from tech companies are meeting with White House staff on Frida
The White House...
Nr. of articles: 105 (89 eng, 5 ger, 11 spa, 0 chi, 0 slo)
Events identified across languages
Map
New York
Connecticut
Pennsylvania
New York
Philadelphia
New Jers
West Virginia
Delaware
Washington D.C.
District of Columbia
Virginia
North Carolina
Map Data
Terms of Use
Articles
89 eng
11 spa
5 ger
0 chi
0 slo
Obama ponders limiting N
Washington -- President Barack Ol
National Security Agency access tc
government's surveillance prograı
Obama could unveil his highly ant
congressional lawmakers, while his
Jan. 9, 12:51
W.DETROITNEWS.COM
89 eng
11 spa
5 ger
view with
Obama likely to adjust US spying on foreign leaders, nearing decision on intelligence changes
Jan. 9, 07:57
CKGL AM 570
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 ...
Obama might limit how much the NSA can spy on you
Jan. 10, 14:04
WWW.IMPACTOUSA.COM
President Barack Obama speaks at a press conference on Dec. 20, 2013. (Pablo Martinez Monsivais/AP)



Collection of events described through Reporting dynamics



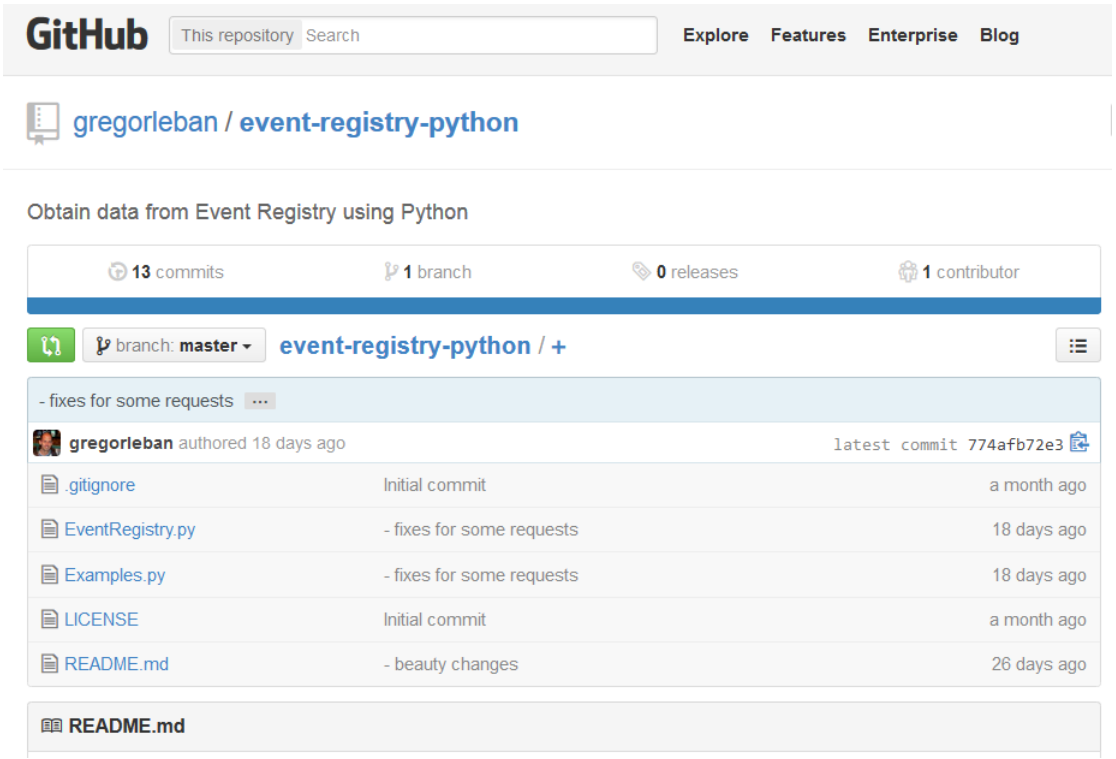
Collection of events
described through
a story-line of related events

Event Registry API

…to be used as a research platform

Python code to access “Event Registry”

- https://github.com/gregorleban/event-registry-python



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

- 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