This is the Information Age — everybody can be informed about anything and everything. There is no secret, therefore there is no sacredness. Life is going to become an open book. When your computer is more loyal, truthful, informed and excellent than you, you will be challenged. You do not have to compete with anybody. You have to compete with yourself.

[Y. Bhajan, 2000]

Real-Time Cross-Lingual Global Media Monitor

Dunja Mladenić
Jozef Stefan Institute, Ljubljana, Slovenia
http://ailab.ijs.si/
Outline

- Introduction
- Collecting Media Data
- Document Enrichment
- Cross-linguality
- Event Representation
- Event Visualization
- Event API
- Future Directions
INTRODUCTION
The overall goal

Establish a real-time system based on ML and NLP enabling to:

- collect data from global media in real-time
- identify events and track evolving topics
- assign stable identifiers to events
- identify events across languages
- detect diversity of reporting along several dimensions
- provide rich exploratory visualizations
- provide interoperable data export

More: Leban, G., Fortuna, B., Brank, J., Grobelnik, M.,
Event Registry: Learning About World Events from News,
Real-Time Cross-lingual News collection:

- 75,000 news sources
- 350,000 news stories per day,
- 10,000 news events identified per day

Support for 100 languages


Cross-Lingual service for 100 languages: [http://XLing.ijs.si](http://XLing.ijs.si)

...open API
...rich visualizations
...analytic capabilities
Systems/Demos used within the presentation

- NewsFeed (http://newsfeed.ijs.si/)
  - News and social media crawler
- Enrycher (http://enrycher.ijs.si/)
  - Language and Semantic annotation
- XLing (http://xling.ijs.si/)
  - Cross-linguual document linking and categorization
- DiversiNews (http://aidemo.ijs.si/diversinews/)
  - News Diversity Explorer
- Event Registry (http://eventregistry.org/)
  - Event detection and topic tracking
Global Media Monitoring pipeline

Input data: Mainstream news, Blogs

Pre-processing steps:
- Article semantic annotation
- Extraction of date references
- Cross-lingual article matching
- Detection of article duplicates

Event construction:
- Article clustering
- Cross-lingual cluster matching
- Event formation
- Event info. extraction
- Identifying related events

Event storage & maintenance:
- Event registry
  - API Interface
  - GUI/Visualizations

Global Media Monitoring pipeline

http://EventRegistry.org
Global Media Monitoring pipeline

Preprocessing steps:
- Article semantic annotation
- Extraction of date references
- Cross-lingual article matching
- Detection of article duplicates

Event construction:
- Article clustering
- Cross-lingual cluster matching
- Event formation
- Event info. extraction
- Identifying related events

Event storage & maintenance:
- Event registry
- API Interface
- GUI/Visualizations

Input data:
- Mainstream news
- Blogs

Preprocessed and annotated enabling cross-lingual article matching
Global Media Monitoring pipeline

**Input data**
- Mainstream news
- Blogs

**Pre-processing steps**
- Article semantic annotation
- Extraction of date references
- Cross-lingual article matching
- Detection of article duplicates
- Cross-lingual clustering

**Event construction**
- Article clustering
- Cross-lingual cluster matching
- Event formation
- Event info. extraction
- Identifying related events

**Event storage & maintenance**
- Event registry
- API Interface
- GUI/Visualizations

Global Media Monitoring pipeline

http://EventRegistry.org

Cross-lingual clustering
Global Media Monitoring pipeline

Input data

Main stream news

Pre-processing steps

Article semantic annotation
Extraction of date references
Cross-lingual article matching
Detection of article duplicates

Event construction

Article clustering
Cross-lingual cluster matching
Event formation
Event info. extraction
Identifying related events

Event storage & maintenance

Event registry

API Interface

GUI/Visualizations

Advanced search and rich visualization

http://EventRegistry.org
COLLECTING MEDIA DATA
Get references to news publishers

- Good start is Wikipedia list of newspapers:

List of newspapers

List of newspapers in Italy

- La Repubblica—website
- Corriere della Sera—website
- La Gazzetta dello Sport—website
- Punto Informatico—website
- La Stampa—website
- LINKIESTA—website
- Huffington Post—website
- Il Sole 24 Ore—website
- Il Giornale—website
- Il Manifesto—website
- Il Secolo XIX—website
- Corriere dello Sport—Stadio—website
- Wall Street Italia—website
- Italy24—website
From a newspaper home-page to an article

http://www.nytimes.com/  HTML

RSS Feed (list of articles)

Article to be retrieved
Collecting global media data

- Data collection service News-Feed
  - [http://newsfeed.ijs.si/](http://newsfeed.ijs.si/)
  - …crawling global main-stream and social media

- Monitoring
  - ~70k 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%),…
Downloading the news stream (1/2)

- The stream is accessible at http://newsfeed.ijs.si/stream/
- To download the whole stream continuously, you can use the python script (http://newsfeed.ijs.si/http2fs.py)
- The script does the following:

```python
timestamp = [when you want to start downloading, e.g. now() minus 1 hour]
while True:
    fetch http://newsfeed.ijs.si/stream/?after=timestamp
    if [404 error]:
        # there is no new data
        pause 1 minute
    else:
        save data
    timestamp = [extract it from the Content-Disposition HTTP header]
```
News Stream Contents and Format

- The root element, `<article-set>`, contains zero or more articles in the following XML format:

```
<article id="internal article ID; consistent across streams">
  <source>
    <hostname> Publisher hostname </hostname>
    <title> Name of the publisher; failing that, title of the RSS feed </title>
    <location>
      <longitude> publisher longitude in degrees </longitude>
      <latitude> publisher latitude in degrees </latitude>
      <city> publisher city </city>
      <country> publisher country </country>
    </location>
    <tags>
      <tag> a tag for the publisher; the vocabulary is not controlled </tag>
    </tags>
  </source>
  <feed>
    <uri> URL from which the article was discovered; typically the RSS feed </uri>
    <uri> URL from which the article was downloaded </uri>
    <publish-date> The publication time and date. </publish-date>
    <retrieve-date> The retrieval time and date. </retrieve-date>
    <lang> 3-letter ISO 639-2 language code </lang>
    <story_id> story cluster this article was grouped into (at download time) </lang>
    <location>
      <longitude> story content longitude in degrees </longitude>
      <latitude> story content latitude in degrees </latitude>
      <city> story city </city>
      <country> story country </country>
    </location>
    <tags>
      <tag> a tag for the article; the vocabulary is not controlled </tag>
    </tags>
    <img> The URL of a related image, usually a thumbnail. </img>
    <title> Title. Can be empty if we fail to identify it. </title>
    <body-cleartext>
      Clear text body of the article, formatted only with &lt; &gt; tags
    </body-cleartext>
    <body-rych>?, only English, Slovene, 
      Enriched article body; an XML subtree as returned by Enrycher.
    </body-rych>
    <body-xlike>?, only English, Spanish, Catalan 
      Enriched article body; an XML subtree as returned by iSOCO; experimental.
    </body-xlike>
  </feed>
</article>
```

More details:

Mitja Trampus, Blaz Novak:

*The Internals Of An Aggregated Web News Feed.*
DOCUMENT ENRICHMENT

http://enrycher.ijs.si/
### How can we annotate a document?

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

Extracted graph of triples from text

"Enrycher" is available as a web-service generating Semantic Graph, LOD links, Entities, Keywords, Categories, Text Summarization, Sentiment
Enrycher is a web service consisting of a set of interlinked modules…
…covering lexical, linguistic and semantic annotations
…exporting data in XML or RDF
To execute the service, one should send an HTTP POST request, with the raw text in the body:

curl -d "Enrycher was developed at JSI, a research institute in Ljubljana. Ljubljana is the capital of Slovenia." http://enrycher.ijs.si/run
Anaphora resolution (1)

- Link pronouns with their references
- Assume that pronouns refer only to named entities
- Can be a difficult problem
  - Examples of difficult sentences:
    - Tom wrote a letter to Bill. He told him ...
    - One passenger in King's car said they had been drinking liquor..
  - ... 
  - He refers to a different person than him

- We link only 5 different pronouns: he (his, him, himself), she (her,...), I (me,...), they (them,...) and who

- Simple resolution procedure:
  - For each pronoun search backward (and forward) in text to find candidate name entities of correct type
  - Score each candidate name entity
    - Score is based on distance from pronoun, part of speech, other parser information (proper name, name of the county), ...
  - Pick a named entity with the best score
Anaphora resolution (2)

- Common mistakes:
  - Quoted speech: *John said: “He is sick.”*
  - But: “I hope so,” he replies after a pause.
  - *Tom wrote a letter to Bill. He told him …*
  - The relationship between active volcanoes and the communities that surround them is not always confrontational.
  - *Jordan's King Hussein and Yasser Arafat's open sympathy for Iraq has strained their relations with the U.S.*
  - The most fatal case is when we wrongly resolve first occurrence of a pronoun and then follow many sentences using only the pronoun to refer to a person

  - Error: *he == John*
    - he == I
  - Error: *him == Tom*
  - We don’t link *them*
  - Can’t link *their*
Anaphora resolution evaluation

- We manually labeled 91 articles
  - Containing 1506 pronouns
  - 1024 (68%) pronouns are *he, she, I, they, who*
    - We try to link all of them
  - Other 482 (32%) pronouns are: *it, you, we, what, …*
# Anaphora resolution evaluation

<table>
<thead>
<tr>
<th>Pronoun</th>
<th>Frequency</th>
<th>Frequency [%]</th>
<th>Accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>He</td>
<td>681</td>
<td>45.22</td>
<td>86.9</td>
</tr>
<tr>
<td>They</td>
<td>244</td>
<td>16.20</td>
<td>67.2</td>
</tr>
<tr>
<td>It</td>
<td>204</td>
<td>13.55</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>64</td>
<td>4.25</td>
<td>82.8</td>
</tr>
<tr>
<td>You</td>
<td>50</td>
<td>3.32</td>
<td></td>
</tr>
<tr>
<td>We</td>
<td>44</td>
<td>2.92</td>
<td></td>
</tr>
<tr>
<td>That</td>
<td>44</td>
<td>2.92</td>
<td></td>
</tr>
<tr>
<td>What</td>
<td>27</td>
<td>1.79</td>
<td></td>
</tr>
<tr>
<td>She</td>
<td>24</td>
<td>1.59</td>
<td>62.5</td>
</tr>
<tr>
<td>This</td>
<td>22</td>
<td>1.46</td>
<td></td>
</tr>
<tr>
<td>Who</td>
<td>11</td>
<td>0.73</td>
<td>63.6</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1506</td>
<td>100</td>
<td>81.2</td>
</tr>
</tbody>
</table>

**Accuracy on 5 selected 81.2% (55.2% if counting all pronouns)**
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
  - avoid machine translation to gain speed and scale

- The key building block is the function for comparing and categorization of documents in different languages
  - XLing.ijs.si is an open web service to bridge information across 100 languages
Languages covered by XLing (top 100 Wikipedia languages)

- With machine learning techniques we can learn “language neutral document representation”…

- …for over 100 Wikipedia languages each having over 10 000 articles

- [http://en.wikipedia.org/wiki/List_of_Wikipe dias#100_000.2B_articles](http://en.wikipedia.org/wiki/List_of_Wikipe dias#100_000.2B_articles)

More:
Jan Rupnik, Andrej Muhic, Primož Skraba,
With machine learning techniques we can learn “language neutral document representation”…

…for over 100 Wikipedia languages each having over 10 000 articles

http://en.wikipedia.org/wiki/List_of_Wikipe dias#100_000.2B_articles

More:

Jan Rupnik, Andrej Muhic, Primož Skraba,
Cross-lingual document retrieval through hub languages. In Workshop book : NIPS 2012, 
Neural Information Processing Systems 
Workshop, Lake Tahoe, USA
Cross-lingual article matching

- Collected articles are written in various languages
- Using CCA we can identify articles in other languages that contain similar content
- Used to determine if articles in different languages are about the same event

documents as vectors

sparse vectors

common semantic space

500 dimensions
XLing (**XLing.ijs.si**) service for comparing and categorization of documents across 100 languages

**Automatically Extracted Keywords**

**Similarity Between Two Documents**

**Chinese Text**

**Selection Of 100 Languages**

**English Text**

World Bank cuts China and Thailand's 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 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 context.

We believe economic challenges remain, IMF managing director Christine Lagarde said.

In addition, China’s growth rate continues to slow - a concern that has been a feature of the 2014 forecast.

IMF forecast growth of 7.1% last October.

But in a report published earlier this month, the World Bank’s chief economist,投猜姆（Andrew Burns）, warned that the Chinese economy is likely to continue to slow.
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 the users we are sending away coming back?
  - …evaluation shows they all come back
Example: Social Media Recommendation (http://aidemo.ijs.si/xlike/hsdemo/)

- Usual problem publishers have what and when to publish to which social media channel?
  - …the demo shows what Bloomberg articles should be the most relevant based on what is most ‘popular’ in a particular geographical area
Example: Article clustering

Identify articles that describe a single event
- Online clustering algorithm
- Grouping based on article title + content + named entities

Procedure:
- Each new article is assigned to the closest cluster
- Every once in a while check if some clusters need to be split or merged
- Old clusters are removed
Cross-lingual cluster linking

- Clusters in different languages can describe the same event
- Consider similarity of relevant concepts and date of articles

English articles

Spanish articles

Most similar articles in Spanish
EVENT REPRESENTATION

http://eventregistry.org/
What is an event? (abstract description)

...more practical question: what definition of event is computationally feasible?

In general, an event is something which “sticks out” of the average in some kind of (high dimensional) data space
- ...could be interpreted as an “anomaly”
- ...densification of data points (e.g. many similar documents)
- ...significant change of distribution (e.g. a trend on Twitter)

In practice, the event could be:
- A cluster of documents / change of a distribution in data
  - Detected in an unsupervised way
- A fit to a pre-built model
  - Detected in a supervised way
How to represent an event?

- Baseline data for a news event is usually a cluster of documents
  - …with some preprocessing we extract linguistic and semantic annotations
  - …semantic annotations are linked to ontologies providing possibility for multiresolution annotations

- Three levels of event representation:
  - **Feature vector event representation:**
    - light weight representation that can be easily represented as a set of feature vectors augmented with external ontologies – suitable for scalable ML analysis
  - **Structured event representation:**
    - Infobox representation (slots filling) using open schema or event taxonomy
  - **Deep event representation**
    - Semantic representation linked to a world-model (e.g. CycKB common sense knowledge) – suitable for reasoning and diagnostics
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”

Where? (geography)

When? (temporal distribution)

Who? (named entities)

What? (keyword topics)
Structured event representation

- Structured event representation describes an event by its “Event Type” and corresponding information slots to be filled.
- Event Types should be taken from “Event Taxonomy”.
- …at this stage of development this level of representation still requires human intervention to achieve high accuracy (Precision/Recall) extraction.

Example on the right – Wikipedia event infobox:
- 2011 Tōhoku earthquake and tsunami
“Event Taxonomy” – preview to the current development
“Event Taxonomy” – preview to the current development
Prototype for event Infobox extraction: semi-automatic annotation service

The goal is to build a system for economically viable extraction of event infoboxes:

- ...using crowdsourcing
- ...aiming at high Precision & Recall for a small cost
Event sequences & Hierarchical events

- Once having events identifies and represented we can connect events into “event sequences” (also called storylines)
- “Event sequences” include events which are supposedly related and constitute larger story
- Collection of interrelated events can be also organized in hierarchies (e.g. World Cup event consists from a series of smaller events)
An example event: Microsoft Windows 9

Windows 9 release date, price, features: why Microsoft can't wait to launch Windows 9 in 2015

WHEN: April 23, 2014

CATEGORIES: Shopping → Home and Garden → Windows, Home → Home Improvement → Windows and Doors, Computers → Software → Freeware

ENTITIES: Microsoft, Twitter, Google, Apple Inc., Computerworld, Gartner, Facebook, Universal Studios, Nokia, United States

KEYWORDS: Microsoft Windows, History of Microsoft Windows, Patch (computing), Software release life cycle, Operating system, Desktop computer, Electronics, Company, NeXT, Software build

ARTICLES: 239 (239 eng, 0 ger, 0 spa, 0 chi, 0 slo)

We outline what we know about Windows 9 release date, Windows 9 price and Windows 9 features. It's fair to say that Windows 8 has not been a universal success. In many ways a necessary step not terribly well executed, Windows 8’s attempt to stitch together desktop and mobile has been way too radical to encourage new users. And as the longer life of existing PCs and laptops combines with the growth of smartphones and tablets to continually slow PC sales, Window 8 has come to be seen as a dud....
## Similar events example: similar events to Microsoft Windows 9 event

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Event Description</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>83%</td>
<td>Windows 8.1 Update: Five features we love</td>
<td>2014/04/04</td>
</tr>
<tr>
<td>82%</td>
<td>Windows 8.2 Rumors: New Start Menu, Apps on Desktop, More</td>
<td>2013/12/18</td>
</tr>
<tr>
<td>81%</td>
<td>Windows 9 release date, price, features: new Start menu and no Charms bar</td>
<td>2014/08/13</td>
</tr>
<tr>
<td>80%</td>
<td>Microsoft's new, new Windows cadence: Once a year is not enough</td>
<td>2014/04/26</td>
</tr>
<tr>
<td>80%</td>
<td>Microsoft making Windows free on devices with screens under nine inches</td>
<td>2014/04/03</td>
</tr>
<tr>
<td>77%</td>
<td>Here are the latest details on Windows 9 and Microsoft's other major Windows updates</td>
<td>2014/06/25</td>
</tr>
<tr>
<td>77%</td>
<td>Mouse and keyboard-friendly Windows 8.1 update arrives April 8</td>
<td>2014/04/02</td>
</tr>
<tr>
<td>77%</td>
<td>Five Places to Learn How to Use Windows 8.1 Better</td>
<td>2014/01/18</td>
</tr>
<tr>
<td>77%</td>
<td>Microsoft's 'Project Siena' lets anyone create Windows 8.1 apps</td>
<td>2013/12/19</td>
</tr>
<tr>
<td>76%</td>
<td>Microsoft strips some Windows 7 users of IE11 patch privileges</td>
<td>2014/07/11</td>
</tr>
<tr>
<td>76%</td>
<td>No Microsoft Start Menu for Windows 8 until 2015</td>
<td>2014/06/02</td>
</tr>
<tr>
<td>76%</td>
<td>Intel plans a CES coup: Android and Windows in the same computer</td>
<td>2014/01/03</td>
</tr>
<tr>
<td>76%</td>
<td>Why Microsoft Is Already Dropping Support for Windows 8.1</td>
<td>2014/04/25</td>
</tr>
<tr>
<td>76%</td>
<td>New Windows 'Threshold' rumor claims virtual desktop support is being tested</td>
<td>2014/08/06</td>
</tr>
<tr>
<td>75%</td>
<td>Microsoft Hopes Windows 9 Will Win Over Desktop Users</td>
<td>2014/07/03</td>
</tr>
<tr>
<td>75%</td>
<td>Microsoft showcases latest updates to Windows, opportunities for developers</td>
<td>2014/04/03</td>
</tr>
<tr>
<td>75%</td>
<td>Windows 8.1 update sweeps 'Metro' under the rug, boots to desktop by default</td>
<td>2014/02/02</td>
</tr>
<tr>
<td>75%</td>
<td>In Six Months, Microsoft Will Pull The Plug On Some Parts Of Windows 7 Support</td>
<td>2014/07/09</td>
</tr>
<tr>
<td>75%</td>
<td>Microsoft talks Windows 8.1 Update 1 features, April 8 release, teases future Start menu</td>
<td>2014/04/05</td>
</tr>
</tbody>
</table>
Event sequence identification
Example Microsoft hierarchy of events
Zoom-in Example
Microsoft hierarchy of events
Live Event tracking with http://EventRegistry.org/

Obama, Harper discuss NATO summit, Iraq in phone call:
White House
Location: Washington D.C., United States
Articles about event: 17

Lt Gen Tijendar Singh to be arrested in connection with Tatra truck deal
Location: New Delhi, India
Articles about event: 11
Event description through entities and Sem
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 Agency access to 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.

The White House...

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

Events identified across languages
Collection of events described through Reporting dynamics
Collection of events described through a story-line of related events.

September 11, 2011

Obama to unveil NSA reforms, response to Snowden

WASHINGTON: President Barack Obama presents Americans on Friday with long-awaited reforms of spy agency phone and Internet data collection sweeps, prompted by the damaging torrent of leaks unleashed by Edward Snowden. Caught between civil liberties campaigners and a resistant intelligence...

Obama announces changes to data collection

Obama on NSA Spying Program: Someone Else, Not

Obama feels reform on some but not all NSA spying
EVENT REGISTRY API
Python code to access "Event Registry"

https://github.com/gregorleban/event-registry-python
Searching for events using Python

```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}
    ], ...]}`
Getting info for a particular event

# get information about event with ID 123

```python
>>> q = QueryEvent("123");
```

# return concept labels in 3 languages

```python
>>> q.addRequestedResult(RequestEventInfo(["eng", "spa", "slv"]));
```

# get 10 most central articles

```python
>>> q.addRequestedResult(RequestEventArticles(0, 10));
```

# get information how articles about the event were trending

```python
>>> q.addRequestedResult(RequestEventArticleTrend());
```

# get top keywords

```python
>>> q.addRequestedResult(RequestEventKeywordAggr());
```

```python
>>> eventRes = er.execQuery(q);
```
Searching for articles

```python
>>> q = QueryArticles();
# articles should be from a particular time period
>>> q.setDateLimit(datetime.date(2014, 4, 16),
    datetime.date(2014, 4, 28))
# they should mention apple
>>> q.addKeyword("apple")
# they should also mention iphone
>>> q.addKeyword("iphone")
# get top 30 articles that match criteria
>>> q.addRequestedResult(RequestArticlesInfo(page=0,
    count = 30));
>>> res = er.execQuery(q)
```
FUTURE: CHALLENGES AND OPPORTUNITIES
Summary

- Combining (light) natural language (pre)processing and data analytics – document enrichment
- Language-neutral text representation by applying statistical methods based on comparable corpora
  - Enables cross-lingual problem solving (information retrieval, document classification and clustering, sentiment detection, event extraction)
  - Similar statistical approaches used for cross-modal data analytics describing the target entity (concept, object, named entity…) by its textual description, photo, related entities from an ontology
- Extracting events from news stream
Scientific Challenges

- Deep understanding of global social dynamics
  - ...what is happening in the World, where, why, who, ...?
  - ...can we predict future events and event consequences?
  - ...what are the drivers of influence and manipulation?
  - ...identifying causality in global event dynamics
  - ...societal tipping points and complex events

- Understanding collected multilingual information
  - ...using actionable semantic representation
  - ...in a language neutral way (semantic cross-linguality!)
  - ...micro-reading (deep understanding of individual documents)
Business/Innovation opportunities

- **Financial/Business sector**
  - prediction of events and their market moving consequences

- **Media sector**
  - …how to report faster, more accurate, more balanced

- **Policy makers**
  - Human Rights / Environment / Research Policy / …
  - What are effects of policy changes?

- **Health**
  - Can we detect Ebola sooner?

- **Security**
  - …problematic trends of different kinds in society

“The outer education provided by the information revolution must be matched by an inner education in wisdom, self-control, intuition and the use of the neutral mind.”

[Y. Bhajan]