PLANETDATA R&D SHOWCASES EVENT REGISTRY

NOVEMBER 2013

Event detection, research journalism, media monitoring, visualization, news bias analysis

SEVENTH FRAMEWORK PROGRAMME

TABLE OF CONTENTS

Contents

SUMMARY

Summary

“Event Registry” – Global media observatory

“Event Registry” is a system developed primarily in FP7 projects PlanetData and X-LIKE with the aim to create a real-time global media observatory. It consists of a series of software components, from data sourcing to visualization and interoperable interfaces. Among others, it includes also a cross-lingual component connecting textual information across over 100 languages. To the best of our knowledge, cross-lingual functionality on such a scale makes the system unique.

The “Event Registry” system is developed as a prototype to support a standardization working group at the IPTC level (publishers' standardization organization – IPTC.org). The aim is to release recommendations to collect, annotate and interoperate information on global events and storylines across languages, domains and granularities.

November 15, 2013

Usage Scenarios

The "Event Registry" system is designed as an end-user application (web based interface) and as a middle-ware component (API and export).

> The original motivation was to build a system helping publishers to (a) search across the media space, and (b) align their own information in the space of other published materials (annotation of news with eventids).

Through connecting textual information across many languages a user can monitor diffusion of information and social dynamics on a global level.

Scenarios which could be covered by the system include the following applications having in most cases overlapping functionalities:

- **Event Registry:** The core of the system is a database of events with extracted locations and time of an event. This covers past, current and future events (i.e. reports with references to future events). The event registry we built is global and connects events (i.e. which are beyond events in the interest of an individual publisher). This kind of global event registry can be integrated with the already existing internal event registries allowing the companies to better plan logistics of sending reporters and coverage of topics in focus. The interest for such a service was expressed by several publishers and news agencies.

- **Research Journalism:** The area of research journalism includes exploratory approach to massive information of different kinds, along many dimensions and on different levels of granularity. The Event Registry system at the present stage allows exploratory analysis of textual news content across different languages on global scale in real-time. Since it includes rich post-processing of collected information and structuring into events and storylines, it can significantly reduce the time to comprehend large amounts of information in a short time which is needed by most publishers with analytic publications. The key features are real-timeness, cross-linguality, dealing with information overload and providing multi-dimensional exploratory interface to the collected information.

- **Event Detection in the long tail:** One of the nontrivial problems each news agency is facing is early detection of events, especially in the long tail of events which are generally below the radar, and in languages outside the focus (i.e.

local languages and main stream languages). Several publishers expressed interest for a system to cover such a gap.

- **Press Clipping and Media Monitoring:** Press clipping is a service provided by agencies to monitor media presence of entities of various kinds. Such services are typically too expensive for a general audience. Possible application of the "Event Registry" system is to provide an inexpensive press clipping service affordable for regular users (for which such a service is otherwise unaffordable).

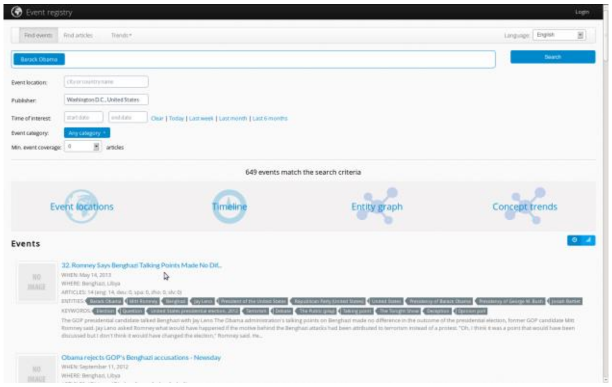
- **Visualization:** The system provides collected and enriched data in a form suitable for various news visualizations (some of which are already included in the present version of the system). The service is relevant for publishers to extend their publications with aggregated visual views from the selected information being collected along several dimensions exposed by the system. In particular, the key dimensions include topical view, social view, temporal view, and opinion view which can be combined in different ways.

- **News bias analysis:** The system collects multiple reports on the same event which allows analysis of diversity in reporting. This includes diversity in opinions, sentiment, coverage, vocabulary in relation to geography, language, publisher, time. The service could be useful for media monitoring and profiling, specifically for media clipping agencies and social science researchers.

- **Plagiarism Detection:** The system allows detection of news stories being copied across data sources. This includes approximate or exact copies. Using cross-lingual functionality, the system allows detection also of translated news (being an innovation). Plagiarism in news publishing is an identified and not generally resolved problem relevant for news agencies and publishers.

Technical Description

The "Event Registry" system is available from the address http://eventregistry.org including web front end (see figures 1 and 2), API and structured data export (RDF, JSON). At the present, the system is in a prototype stage with all major components functioning, but in a process of improvements on several stages.



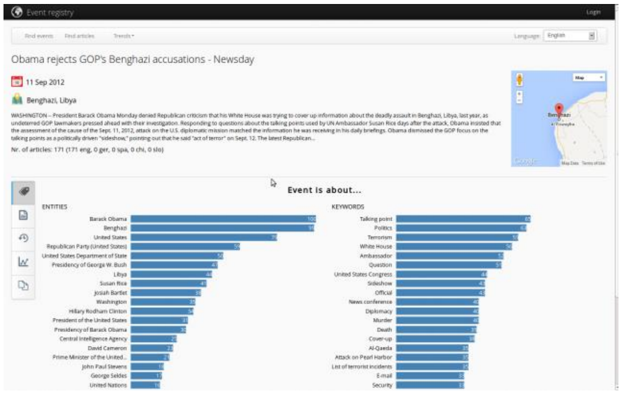
649 events match the search criteria

32. Romney Says Benghazi Talking Points Made No Dif...
WHEN: May 14, 2013
WHERE: Benghazi, Libya
ARTICLES: 14 (eng: 14, deu: 0, spa: 0, zho: 0, slv: 0)

The GOP presidential candidate talked Benghazi with Jay Leno The Obama administration's talking points on Benghazi made no difference in the outcome of the presidential election, former GOP candidate Mitt Romney said. Jay Leno asked Romney what would have happened if the motive behind the Benghazi attacks had been attributed to terrorism instead of a protest. "Oh, I think it was a point that would have been discussed but I don't think it would have changed the election," Romney said. He...

Obama rejects GOP's Benghazi accusations - Newsday
WHEN: September 11, 2012
WHERE: Benghazi, Libya

Figure 1: Event Registry web front end – event search



Obama rejects GOP's Benghazi accusations - Newsday

11 Sep 2012

Benghazi, Libya

WASHINGTON – President Barack Obama Monday denied Republican criticism that his White House was trying to cover up information about the deadly assault in Benghazi, Libya, last year, as undeterred GOP lawmakers pressed ahead with their investigation. Responding to questions about the talking points used by UN Ambassador Susan Rice days after the attack, Obama insisted that the assessment of the cause of the Sept. 11, 2012, attack on the U.S. diplomatic mission matched the information he was receiving in his daily briefings. Obama dismissed the GOP focus on the talking points as a politically driven "sideshow," pointing out that he said "act of terror" on Sept. 12. The latest Republican...

Nr. of articles: 171 (171 eng, 0 ger, 0 spa, 0 chi, 0 slo)

Figure 2: Event Registry web front end - event view

The "Event Registry" operational pipeline consists from several software components implemented as a set of interconnected web services. The key components are the following:

NewsFeed (http://newsfeed.ijs.si) – developed as a part of PlanetData. NewsFeed is sourcing data from few 100 thousands sources including main stream media, blogs, and Twitter. On average it collects between 300.00 and 500.000 documents (excluding Twitter) per day. Each document is stored copied, cleaned and stored in the database and put available for the follow-up omponents.

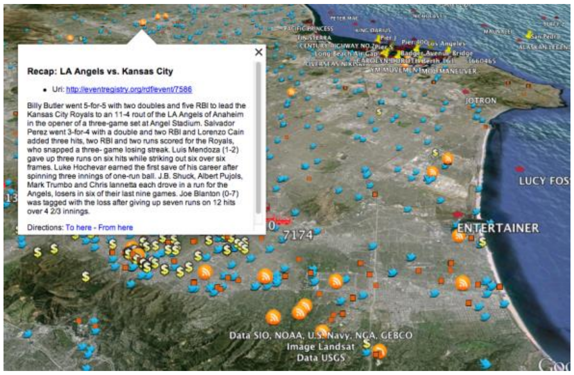
Enrycher (http://enrycher.ijs.si) – developed as a part of ACTIVE, RENDER, PlanetData and X-LIKE projects. Erycher consists from a series of sub-components for linguistic and light-weight semantic annotation for multiple languages. Its main function is transform a ASCII/UNICODE string into a structured XML including tokenization, Part-of-Speech tagging, Name-Entity-Recognition, dependency parsing, entity disambiguation and linking to external sources (Linked Open Data sources like DBpedia, OpenCyc, Freebase), extracting sentiment and summarization. Not all functionalities are implemented for all supported languages.

XLing (http://xling.ijs.si) – developed as a part of X-LIKE project. XLingimplements a statistical cross-lingual matching across 100 languages (in the upcoming version) and classifies documents from all supported languages into a common taxonomic schema (currently used DMoz.org taxonomy). In other words, the key function is measuring topical similarity among documents written in different languages providing information if documents are translations of each other, topically related or topically unrelated.

Event detection – developed as a part of X-LIKE project. This module forms events from documents retrieved from NewsFeed, enriched with Enrycher, and linked across languages with XLing. Each event is a cluster of documents represented with three major components: social (entities), content (keywords, concepts) and temporal (time information). In the future, the goal is to extract structured versions of events in a form of event types.

Storyline detection and visualization (http://eventregistry.org/) – developed as a part of PlanetData project. In this component the goal is to connect events into storylines and visualize them through user interface. Currently, the events are connected in a subgraphs based on the multicomponent similarity. User interface allows temporal browsing through storyline events.

Export to interoperable formats – developed as a part of PlanetData project. The system exports the collected and processed information on events and storylines into multiple formats including JSON and RDF. Both formats are available via REST API polling. A search feature is implemented which can be queried with DBpedia entity URIs and an RDF document with URIs of events where the entities are relevant are returned (example of query: eventregistry.org/rdf/search?keywords=http://dbpedia.org/resource/Barack_Obama). RDF data is annotated using well established ontologies such as Press Association's Storyline ontology (http://www.bbc.co.uk/ontologies/storyline/2013-05-01.html) and IPTC's rNews annotation (http://dev.iptc.org/rNews).



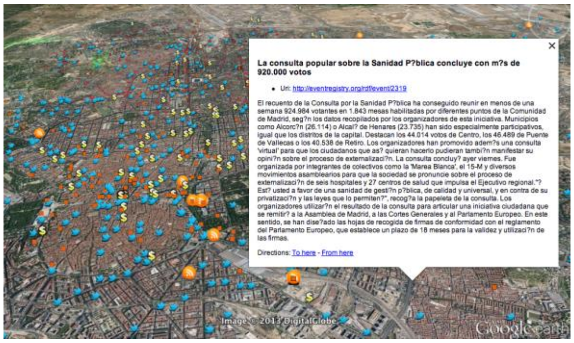
Recap: LA Angels vs. Kansas City

- Uri: http://eventregistry.org/rdf/event/7586

Billy Butler went 5-for-5 with two doubles and five RBI to lead the Kansas City Royals to an 11-4 rout of the LA Angels of Anaheim in the opener of a three-game set at Angel Stadium. Salvador Perez went 3-for-4 with a double and two RBI and Lorenzo Cain added three hits, two RBI and two runs scored for the Royals, who snapped a three- game losing streak. Luis Mendoza (1-2) gave up three runs on six hits while striking out six over six frames. Luke Hochevar earned the first save of his career after spinning three innings of one-run ball. J.B. Shuck, Albert Pujols, Mark Trumbo and Chris Iannetta each drove in a run for the Angels, losers in six of their last nine games. Joe Blanton (0-7) was tagged with the loss after giving up seven runs on 12 hits over 4 2/3 innings.

Directions: To here - From here

Figure 3: Smart City demonstrator - inclusion of Event Registry data



La consulta popular sobre la Sanidad P?blica concluye con m?s de 920.000 votos

- Uri: http://eventregistry.org/rdf/event/2319

El recuento de la Consulta por la Sanidad P?blica ha conseguido reunir en menos de una semana 924.984 votantes en 1.843 mesas habilitadas por diferentes puntos de la Comunidad de Madrid, seg?n los datos recopilados por los organizadores de esta iniciativa. Municipios como Alcorc?n (26.114) o Alcal? de Henares (23.735) han sido especialmente participativos, igual que los distritos de la capital. Destacan los 44.014 votos de Centro, los 46.489 de Puente de Vallecas o los 40.538 de Retiro. Los organizadores han promovido adem?s una consulta 'virtual' para que los ciudadanos que as? quieran hacerlo pudieran tambi?n manifestar su opini?n sobre el proceso de externalizaci?n. La consulta concluy? ayer viernes. Fue organizada por integrantes de colectivos como la 'Marea Blanca', el 15-M y diversos movimientos asamblearios para que la sociedad se pronuncie sobre el proceso de externalizaci?n de seis hospitales y 27 centros de salud que impulsa el Ejecutivo regional."? Est? usted a favor de una sanidad de gesti?n p?blica, de calidad y universal, y en contra de su privatizaci?n y las leyes que lo permiten?", recog?a la papeleta de la consulta. Los organizadores utilizar?n el resultado de la consulta para articular una iniciativa ciudadana que se remitir? a la Asamblea de Madrid, a las Cortes Generales y al Parlamento Europeo. En este sentido, se han dise?ado las hojas de recogida de firmas de conformidad con el reglamento del Parlamento Europeo, que establece un plazo de 18 meses para la validez y utilizaci?n de las firmas.

Directions: To here - From here

Figure 4: Smart City demonstrator - inclusion of Event Registry data

KIT worked with JSI on an RDF-based REST API to access the Event Registry dataset. The collaboration involved designing a URI scheme to identify relevant entities from the Event Registry dataset. Also, we worked together to identify and select relevant vocabularies for the representation of Event

TECHNICAL DESCRIPTION

Registry data. Based on the API access provided to JSI's infrastructure, KIT included event information from Event Registry in the real-time Smart City demonstrator. In Figures 3 and 4, RSS icons represent events from the Event Registry; the pop-up window shows details about the referenced event.

CONTACT INFORMATION

Contact Information

Aljaz Kosmerlj (Jozef Stefan Institute) aljaz.kosmerlj@ijs.si

Andreas Harth (Karlsruhe Institute of Technology) harth@kit.edu

Carolina Fortuna (Jozef Stefan Institute) carolina.fortuna@ijs.si

Marko Grobelnik (Jozef Stefan Institute) marko.grobelnik@ijs.si

PROJECT INFORMATION

Funded under: 7th FWP (Seventh Framework Programme)

Area: Intelligent Information Management (ICT-2009.4.3)

Project Reference: 257641

Contract type: Network of Excellence

Contact: Alice Carpentier alice.carpentier@sti2.at

Website: www.planet-data.eu