Improving the Freshness of Web Collections by Integrating Social Web and Focused Web Crawling

Thomas Risse

L3S Research Center/Leibniz Universität Hannover

IFLA International News Media Conference

Hamburg, 21.4.2016

Social Media

Properties

- Important change in the communication on the internet
- Easy to create, share, or exchange information
- Easy to connect with family, friends, colleagues, interesting people
- Everybody is able to contribute
- Can be used everywhere
  - Independent of the location
  - Independent of the medium: Web, Smartphone, Smartwatch,

Societal View

- Good representation of our culture and society
- Valuable insights into individuals, groups, and organizations
- Enable an understanding of the public perception of events, people, products, or companies, including the flow of information
- Detailed insights into the day-to-day process of public communication

Twitter – A News Medium for Event-Following

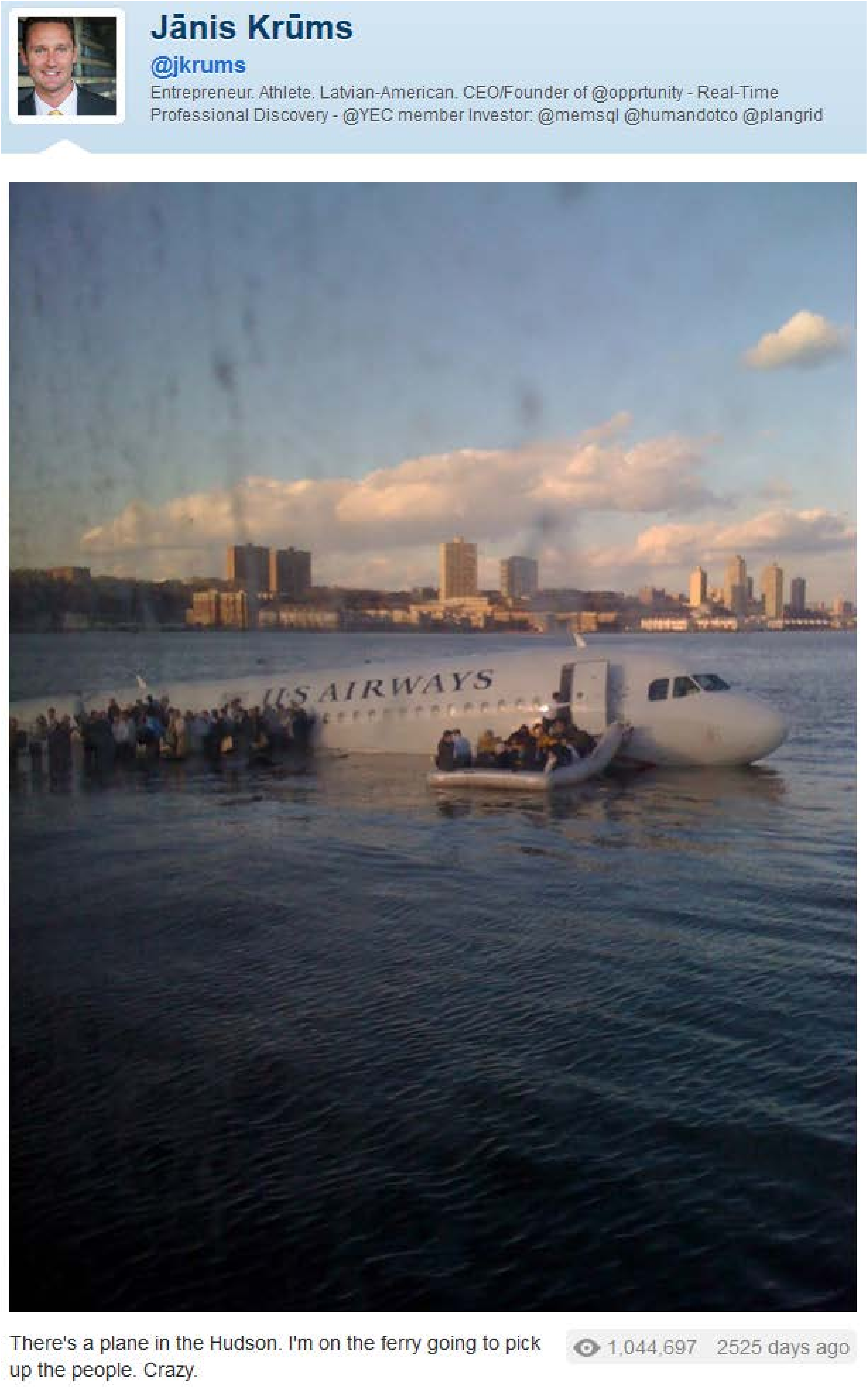
Citizen Journalism

- Everybody can be a journalist by using Smartphone & Twitter
- E.g. Hudson River Plane Crash 2009

Event Discussions

- 2014 FIFA World Cup semi-final between Brazil and Germany on July 8, 2014
  → 35.6 Million tweets

→ Good documentation of the public perception of the event



Jānis Krūms
@jkrums
Entrepreneur. Athlete. Latvian-American. CEO/Founder of @opprtunity - Real-Time Professional Discovery - @YEC member Investor: @memsql @humandotco @plangrid

There's a plane in the Hudson. I'm on the ferry going to pick up the people. Crazy.

1,044,697 2525 days ago

Growing Interest in Web Archive Content

Journalists, Historians, Social Sciences, Law, …

- Relevant content
  - Official Publications (e.g. Government)
  - Journalistic Resources
  - Important topics and events with a high media coverage
  - Multi-cultural or controversial topics
- Observations of topics and events on major sites or Social Media are good starting points
- Metadata / Context (e.g. Author, Organizations and their interests, gender, location)
- Demographic information about social sites
- Provenance: Transparent and detailed documentation of content selection

Derived Requirements

Topical Dimension

- Crawl intention are mainly focused around events and rarely around entities
- What is the intention of the researcher?
- Easy monitoring by the researcher and possibility to correct

Flexible Crawling Strategies

- Shallow observation crawls (Social Media, Web)
- Focused crawls with prioritization (e.g. PageRank and/or semantics)

Social Web Crawling

- General interest with different media focus
- Integrated with Web crawler to capture the **full context**

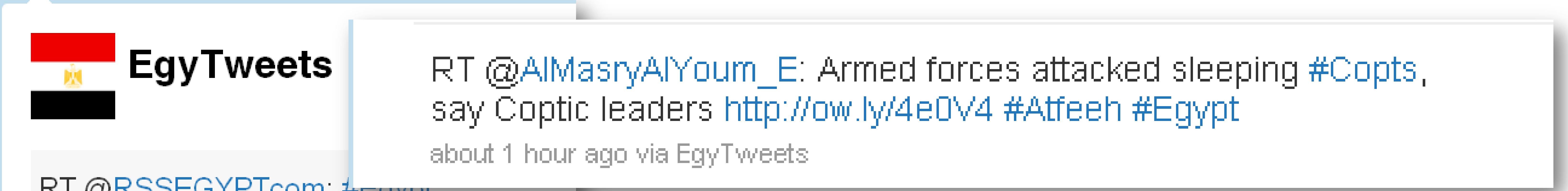
Authenticity

- See a web page as the user saw the page **(e.g. including ads and tweets at that time point)**

Context and Provenance

- Demographics of sites
- Documentation of crawl specification and history

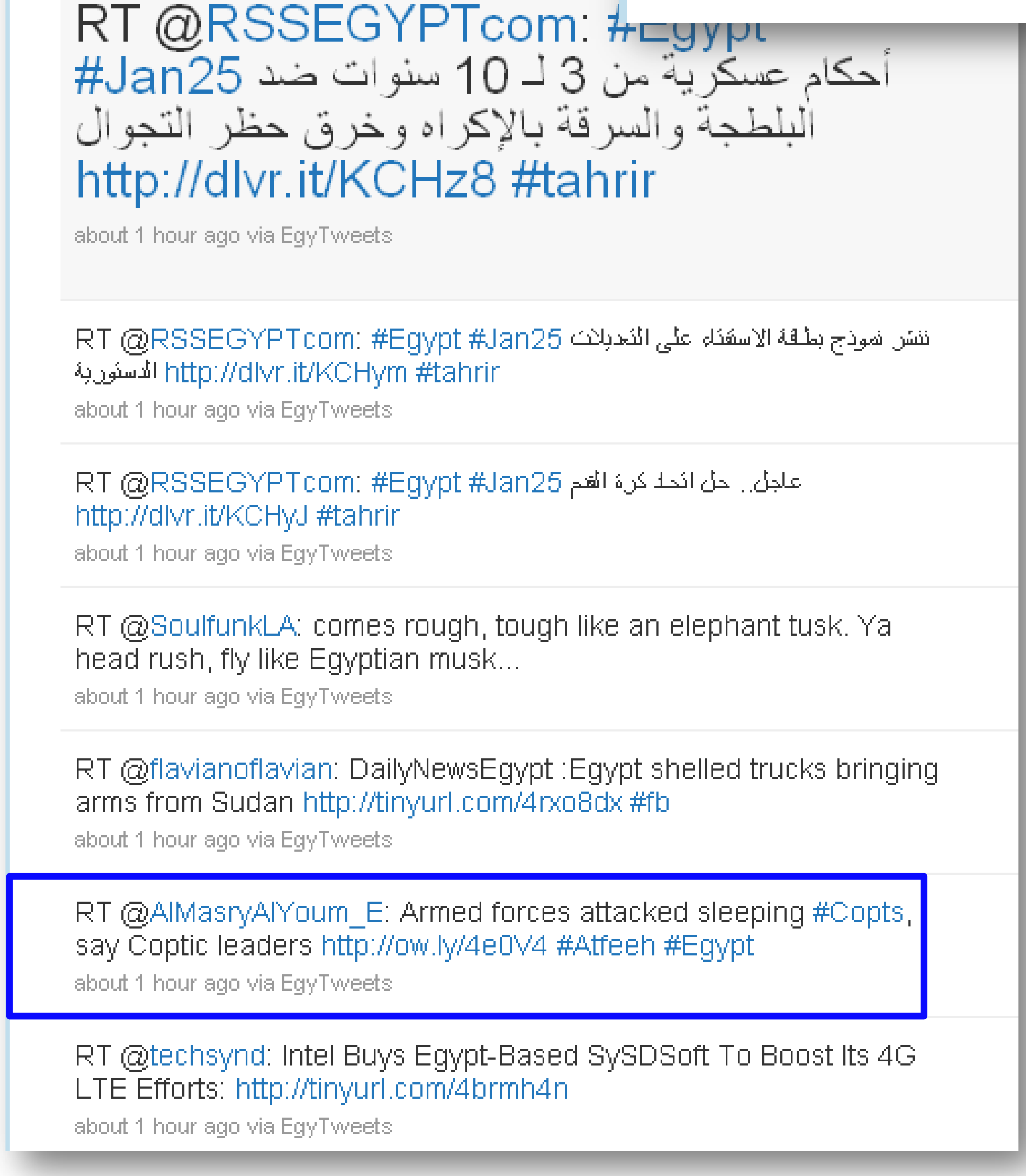
Is Twitter Content enough?



EgyTweets

RT @AlMasryAlYoum_E: Armed forces attacked sleeping #Copts, say Coptic leaders http://ow.ly/4e0V4 #Atfeeh #Egypt

about 1 hour ago via EgyTweets



RT @RSSEGYPTcom: #Egypt #Jan25 أحكام عسكرية من 3 لـ 10 سنوات ضد البلطجة والسرقة بالإكراه وخرق حظر التجوال http://dlvr.it/KCHz8 #tahrir

about 1 hour ago via EgyTweets

RT @RSSEGYPTcom: #Egypt #Jan25 نشر نموذج بطاقة الاستفتاء على التعديلات الدستورية http://dlvr.it/KCHym #tahrir

about 1 hour ago via EgyTweets

RT @RSSEGYPTcom: #Egypt #Jan25 عاجل .. حل اتحاد كرة القدم http://dlvr.it/KCHyJ #tahrir

about 1 hour ago via EgyTweets

RT @SoulfunkLA: comes rough, tough like an elephant tusk. Ya head rush, fly like Egyptian musk...

about 1 hour ago via EgyTweets

RT @flavianoflavian: DailyNewsEgypt :Egypt shelled trucks bringing arms from Sudan http://tinyurl.com/4rxo8dx #fb

about 1 hour ago via EgyTweets

RT @AlMasryAlYoum_E: Armed forces attacked sleeping #Copts, say Coptic leaders http://ow.ly/4e0V4 #Atfeeh #Egypt

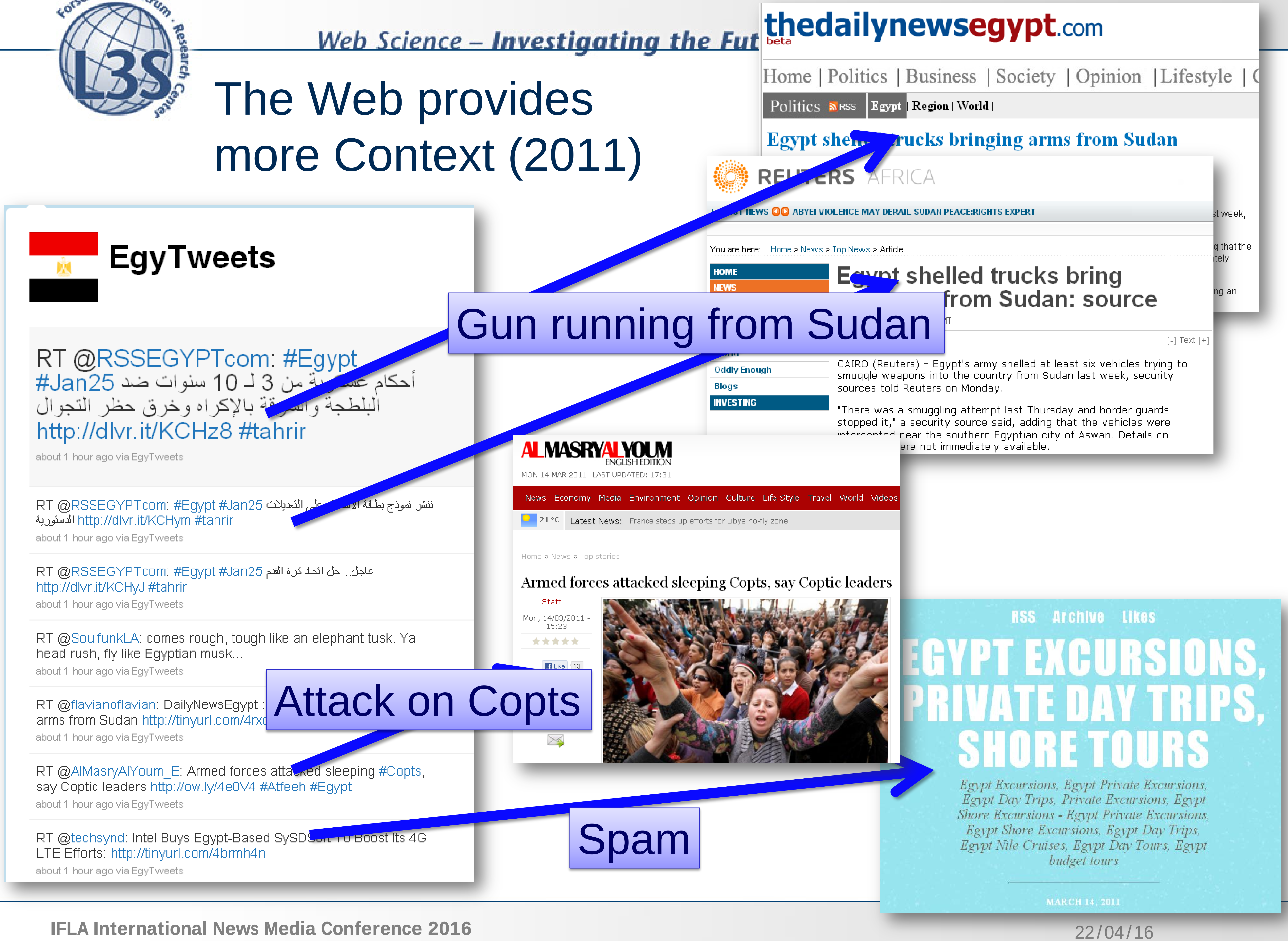
about 1 hour ago via EgyTweets

RT @techsynd: Intel Buys Egypt-Based SySDSoft To Boost Its 4G LTE Efforts: http://tinyurl.com/4brmh4n

about 1 hour ago via EgyTweets

- A tweet is limited to the most important information
- Can we still understand the meaning and the context in the future?
- We need to make use of all hints we can get to ensure the interpretability

The Web provides more Context (2011)



EgyTweets

RT @RSSEGYPTcom: #Egypt #Jan25 أحكام عسكرية من 3 لـ 10 سنوات ضد البلطجة والسرقة بالإكراه وخرق حظر التجوال http://dlvr.it/KCHz8 #tahrir
about 1 hour ago via EgyTweets

RT @RSSEGYPTcom: #Egypt #Jan25 نشر نموذج بطاقة الاستفتاء على التعديلات الدستورية http://dlvr.it/KCHym #tahrir
about 1 hour ago via EgyTweets

RT @RSSEGYPTcom: #Egypt #Jan25 عاجل.. حل اتحاد كرة القدم http://dlvr.it/KCHyJ #tahrir
about 1 hour ago via EgyTweets

RT @SoulfunkLA: comes rough, tough like an elephant tusk. Ya head rush, fly like Egyptian musk...
about 1 hour ago via EgyTweets

RT @flavianoflavian: DailyNewsEgypt : arms from Sudan http://tinyurl.com/4rx
about 1 hour ago via EgyTweets

RT @AlMasryAlYoum_E: Armed forces attacked sleeping #Copts, say Coptic leaders http://ow.ly/4e0V4 #Atfeeh #Egypt
about 1 hour ago via EgyTweets

RT @techsynd: Intel Buys Egypt-Based SySDSoft To Boost Its 4G LTE Efforts: http://tinyurl.com/4brmh4n
about 1 hour ago via EgyTweets

thedailynewsegypt.com beta

Home | Politics | Business | Society | Opinion | Lifestyle |

Politics RSS | Egypt | Region | World |

Egypt shells trucks bringing arms from Sudan

REUTERS AFRICA

NEWS ABYEI VIOLENCE MAY DERAIL SUDAN PEACE:RIGHTS EXPERT

You are here: Home > News > Top News > Article

HOME
NEWS
Oddly Enough
Blogs
INVESTING

Egypt shelled trucks bring from Sudan: source

[-] Text [+]

CAIRO (Reuters) - Egypt's army shelled at least six vehicles trying to smuggle weapons into the country from Sudan last week, security sources told Reuters on Monday.

"There was a smuggling attempt last Thursday and border guards stopped it," a security source said, adding that the vehicles were intercepted near the southern Egyptian city of Aswan. Details on ere not immediately available.

AL MASRY AL YOUM ENGLISH EDITION

MON 14 MAR 2011 LAST UPDATED: 17:31

News Economy Media Environment Opinion Culture Life Style Travel World Videos

21 °C Latest News: France steps up efforts for Libya no-fly zone

Home » News » Top stories

Armed forces attacked sleeping Copts, say Coptic leaders

Staff
Mon, 14/03/2011 - 15:23

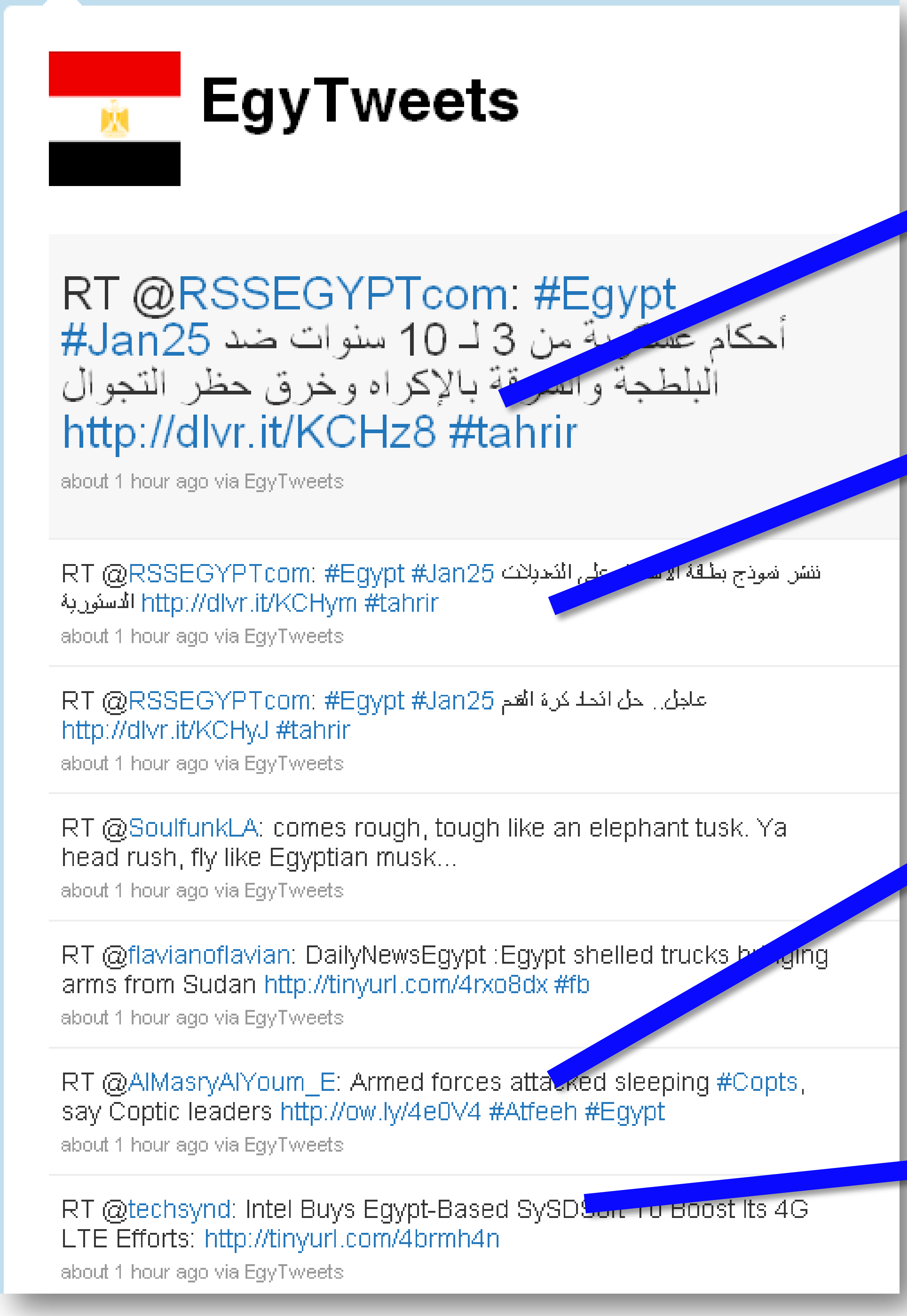
RSS Archive Likes

EGYPT EXCURSIONS, PRIVATE DAY TRIPS, SHORE TOURS

Egypt Excursions, Egypt Private Excursions, Egypt Day Trips, Private Excursions, Egypt Shore Excursions - Egypt Private Excursions, Egypt Shore Excursions, Egypt Day Trips, Egypt Nile Cruises, Egypt Day Tours, Egypt budget tours

MARCH 14, 2011

The Web provides more Context (2016)



EgyTweets

RT @RSSEGYPTcom: #Egypt #Jan25 أحكام عسكرية من 3 لـ 10 سنوات ضد البلطجة والسرقة بالإكراه وخرق حظر التجوال http://dlvr.it/KCHz8 #tahrir

about 1 hour ago via EgyTweets

RT @RSSEGYPTcom: #Egypt #Jan25 نشر نموذج بطاقة الاستفتاء على التعديلات الدستورية http://dlvr.it/KCHym #tahrir

about 1 hour ago via EgyTweets

RT @RSSEGYPTcom: #Egypt #Jan25 عاجل.. حل اتحاد كرة القدم http://dlvr.it/KCHyJ #tahrir

about 1 hour ago via EgyTweets

RT @SoulfunkLA: comes rough, tough like an elephant tusk. Ya head rush, fly like Egyptian musk...

about 1 hour ago via EgyTweets

RT @flavianoflavian: DailyNewsEgypt :Egypt shelled trucks bringing arms from Sudan http://tinyurl.com/4rxo8dx #fb

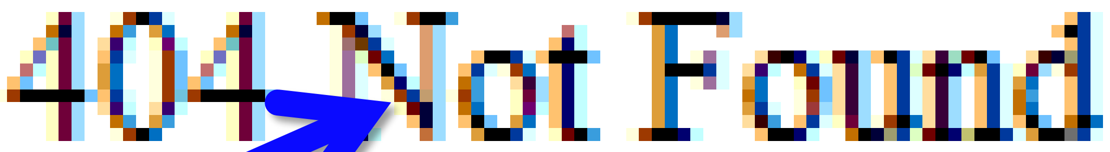
about 1 hour ago via EgyTweets

RT @AlMasryAlYoum_E: Armed forces attacked sleeping #Copts, say Coptic leaders http://ow.ly/4e0V4 #Atfeeh #Egypt

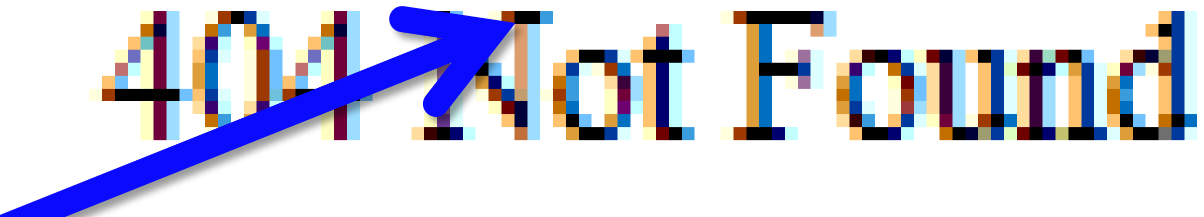
about 1 hour ago via EgyTweets

RT @techsynd: Intel Buys Egypt-Based SySDSoft To Boost Its 4G LTE Efforts: http://tinyurl.com/4brmh4n

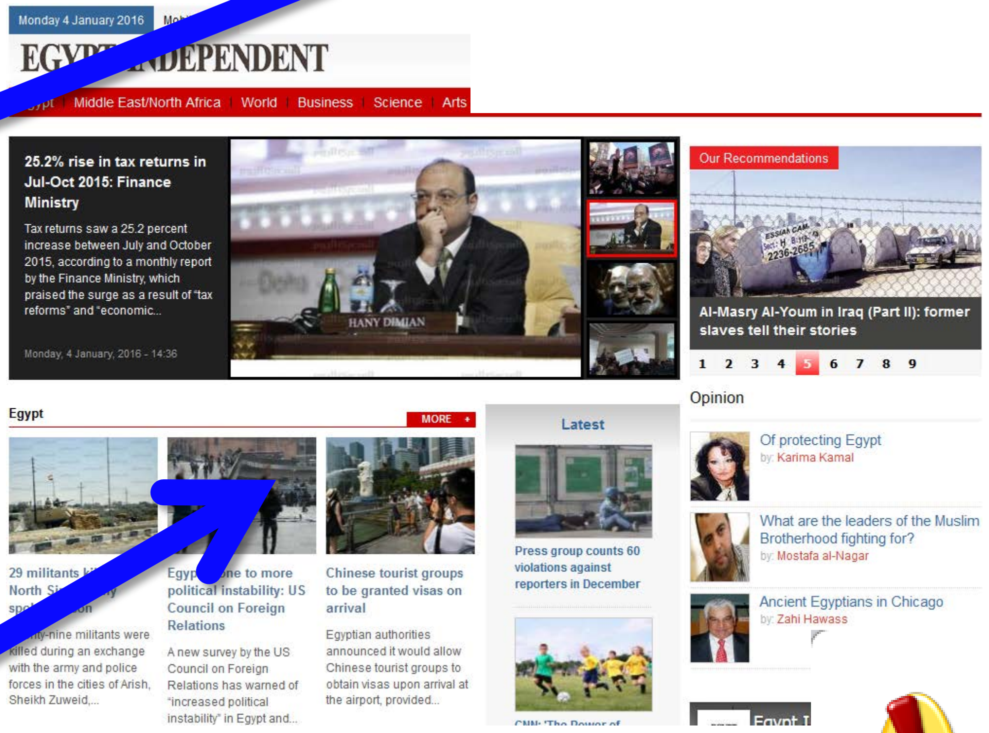
about 1 hour ago via EgyTweets



404 Not Found



404 Not Found



Monday 4 January 2016

EGYPT INDEPENDENT

Egypt | Middle East/North Africa | World | Business | Science | Arts

25.2% rise in tax returns in Jul-Oct 2015: Finance Ministry

Tax returns saw a 25.2 percent increase between July and October 2015, according to a monthly report by the Finance Ministry, which praised the surge as a result of "tax reforms" and "economic...

Monday, 4 January, 2016 - 14:36

Our Recommendations

Al-Masry Al-Youm in Iraq (Part II): former slaves tell their stories

1 2 3 4 5 6 7 8 9

Egypt MORE +

29 militants killed in North Sinai: army spokesperson

Twenty-nine militants were killed during an exchange with the army and police forces in the cities of Arish, Sheikh Zuweid,...

Egypt prone to more political instability: US Council on Foreign Relations

A new survey by the US Council on Foreign Relations has warned of "increased political instability" in Egypt and...

Chinese tourist groups to be granted visas on arrival

Egyptian authorities announced it would allow Chinese tourist groups to obtain visas upon arrival at the airport, provided...

Latest

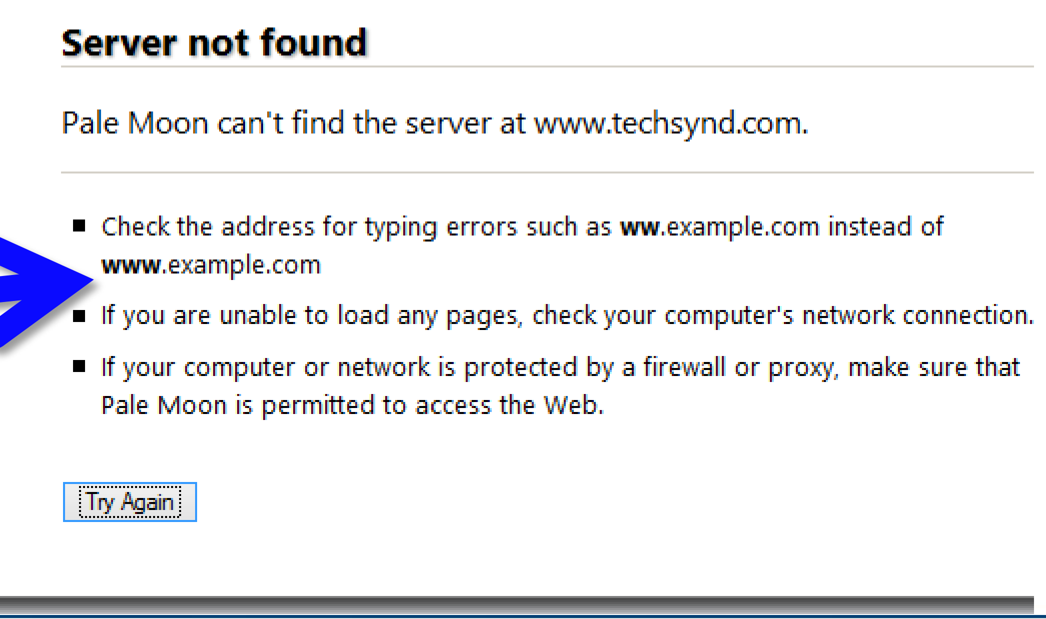
Press group counts 60 violations against reporters in December

Opinion

Of protecting Egypt
by: Karima Kamal

What are the leaders of the Muslim Brotherhood fighting for?
by: Mostafa al-Nagar

Ancient Egyptians in Chicago
by: Zahi Hawass



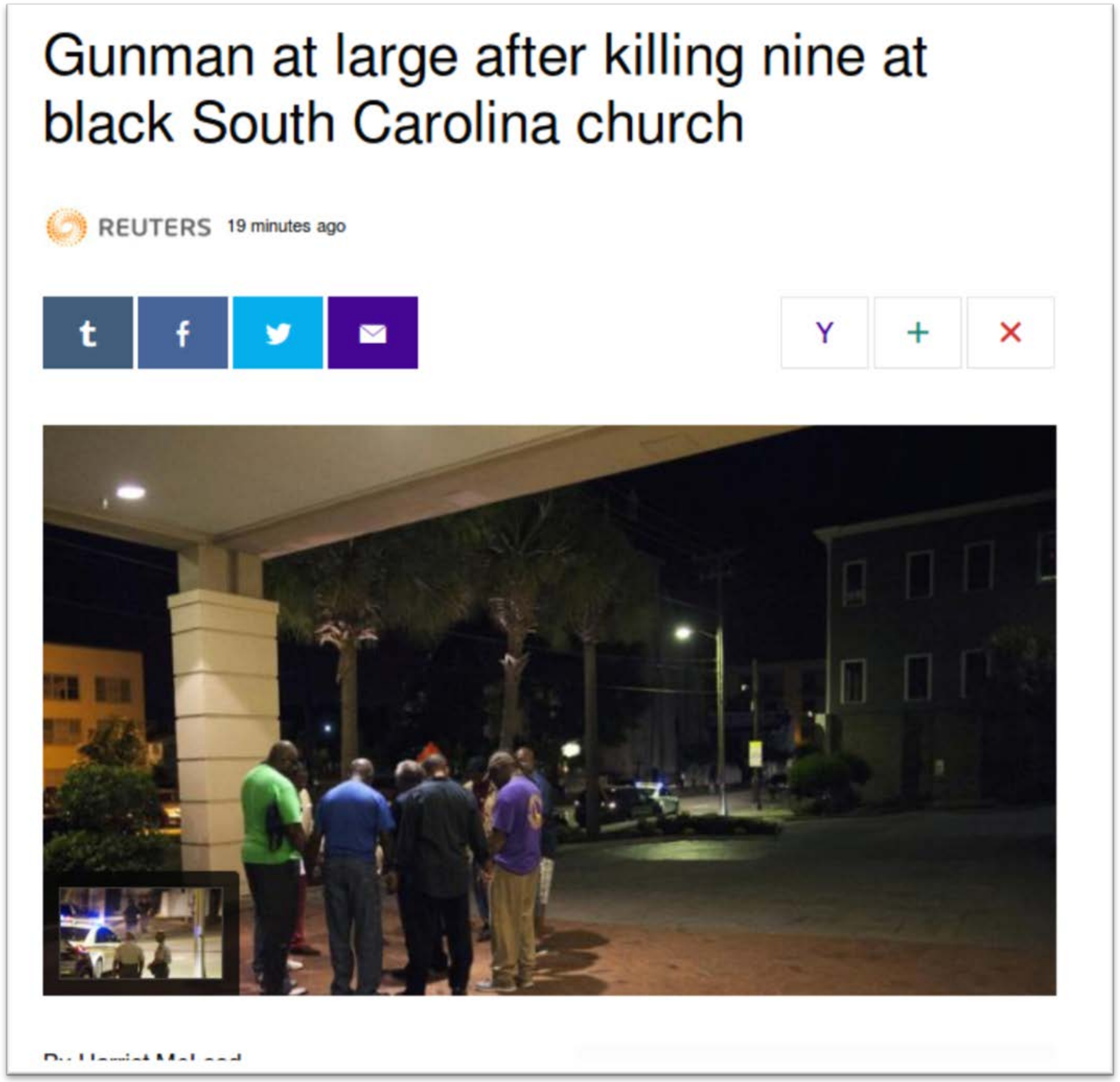
Server not found

Pale Moon can't find the server at www.techsynd.com.

- Check the address for typing errors such as **ww**.example.com instead of **www**.example.com
- If you are unable to load any pages, check your computer's network connection.
- If your computer or network is protected by a firewall or proxy, make sure that Pale Moon is permitted to access the Web.

Web changes in response to current events

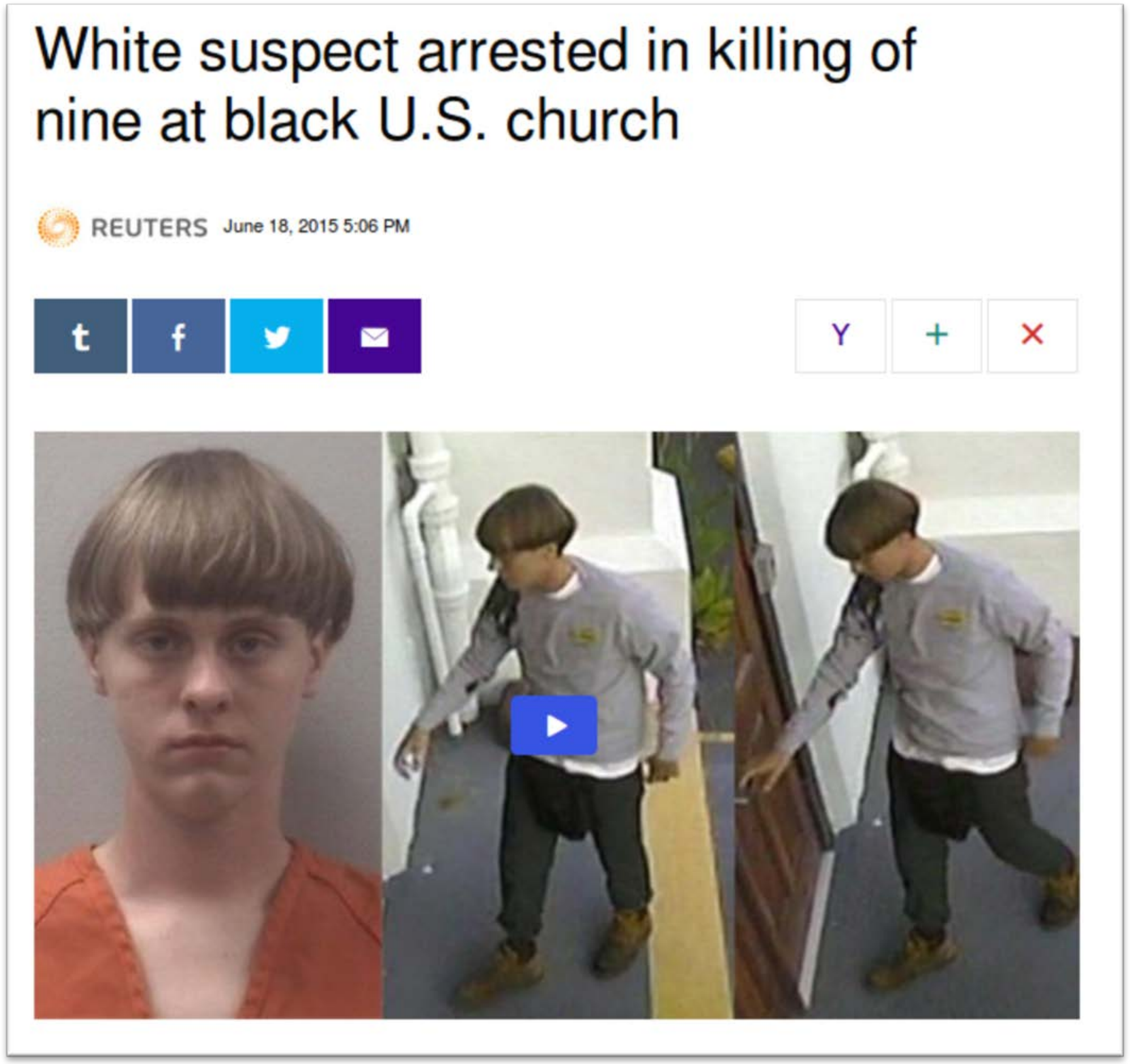
Internet Archive June 18th, 2015,3:17 vs. 17:06 (same day)



Gunman at large after killing nine at black South Carolina church

REUTERS 19 minutes ago

By Harriet McLeod



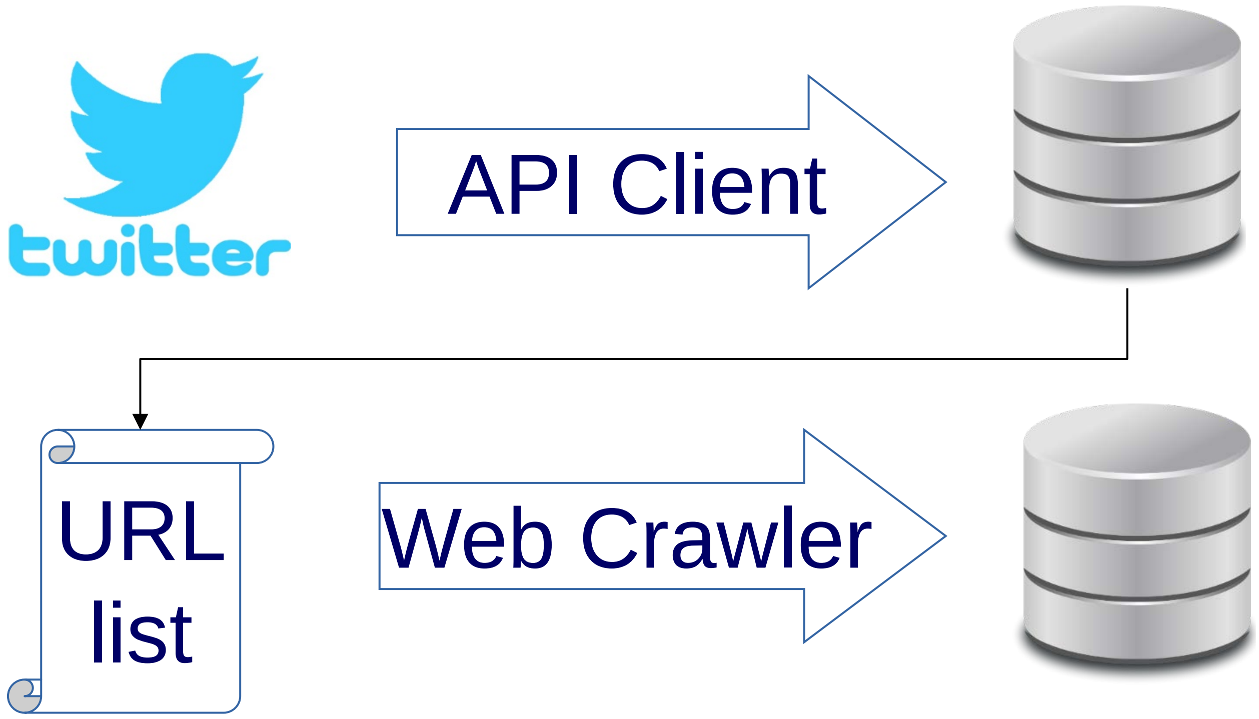
White suspect arrested in killing of nine at black U.S. church

REUTERS June 18, 2015 5:06 PM

Source: http://news.yahoo.com/shooting-erupts-church-charleston-south-carolina-021744448.html,
example by Bergis Jules (https://medium.com/on-archivy/the-narrative-of-terrorism-in-charleston-b8bd79d81741)

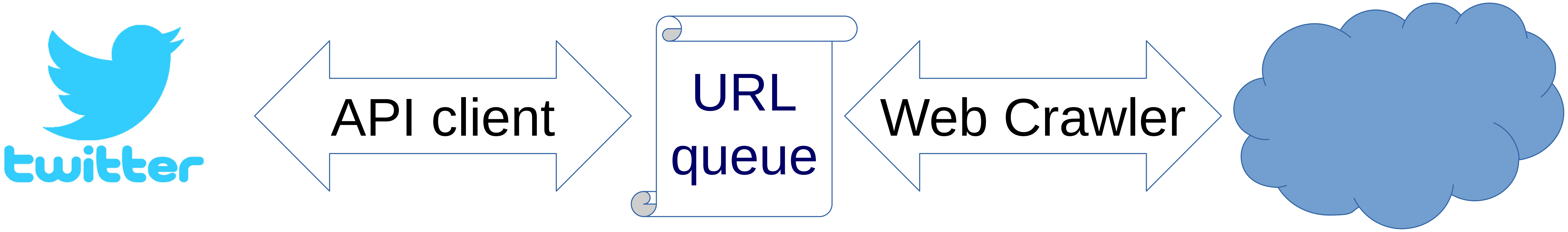
Current approach: Collect, then crawl

- Social Media: scalable access only through API
  - Requires special client programming and maintenance
  - Not supported by typical crawlers
- Workaround Process
  1. Crawling of Social Media content
  2. Extraction of Links
  3. Crawling of Web Pages
- Result
  - Static integration of Social Media
  - Uni-directional Path: Social Media → Web Content
  - Huge delay between time of post and time of crawling!
  - Missing Path: Web Content → Social Media

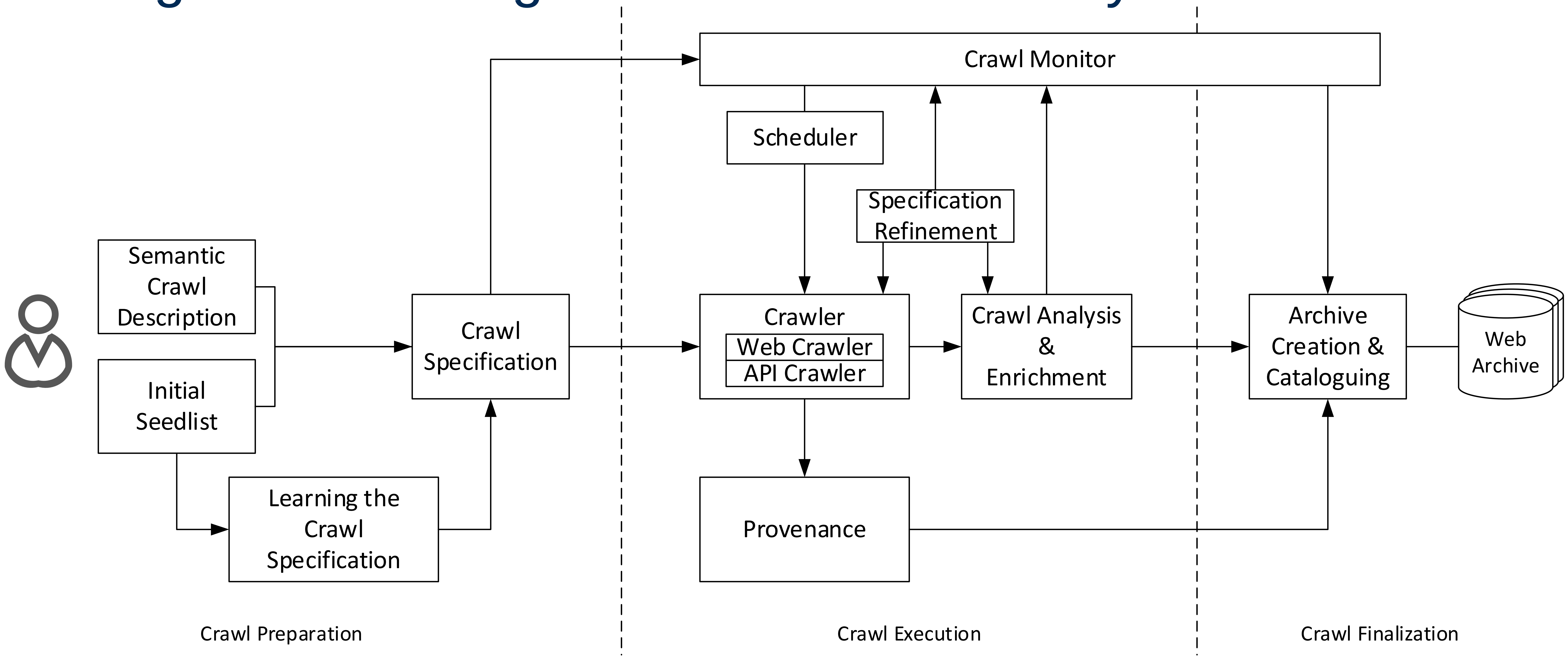


Integrated Crawling approach

- Social Media API
  - convenient query methods + (in Twitter) real-time stream
    → continuous stream of seeds for Web crawler
  - Social media URLs follow changes in topic
    → keeps crawler on topic even when topic evolves
- Integrated Crawling
  - API client and Web crawler cooperate through shared queue
  - URLs in Tweets are inserted early in the queue to ensure timely crawling
  - Suitable prioritization of URLs
  - Crawl continues also from tweeted URLs



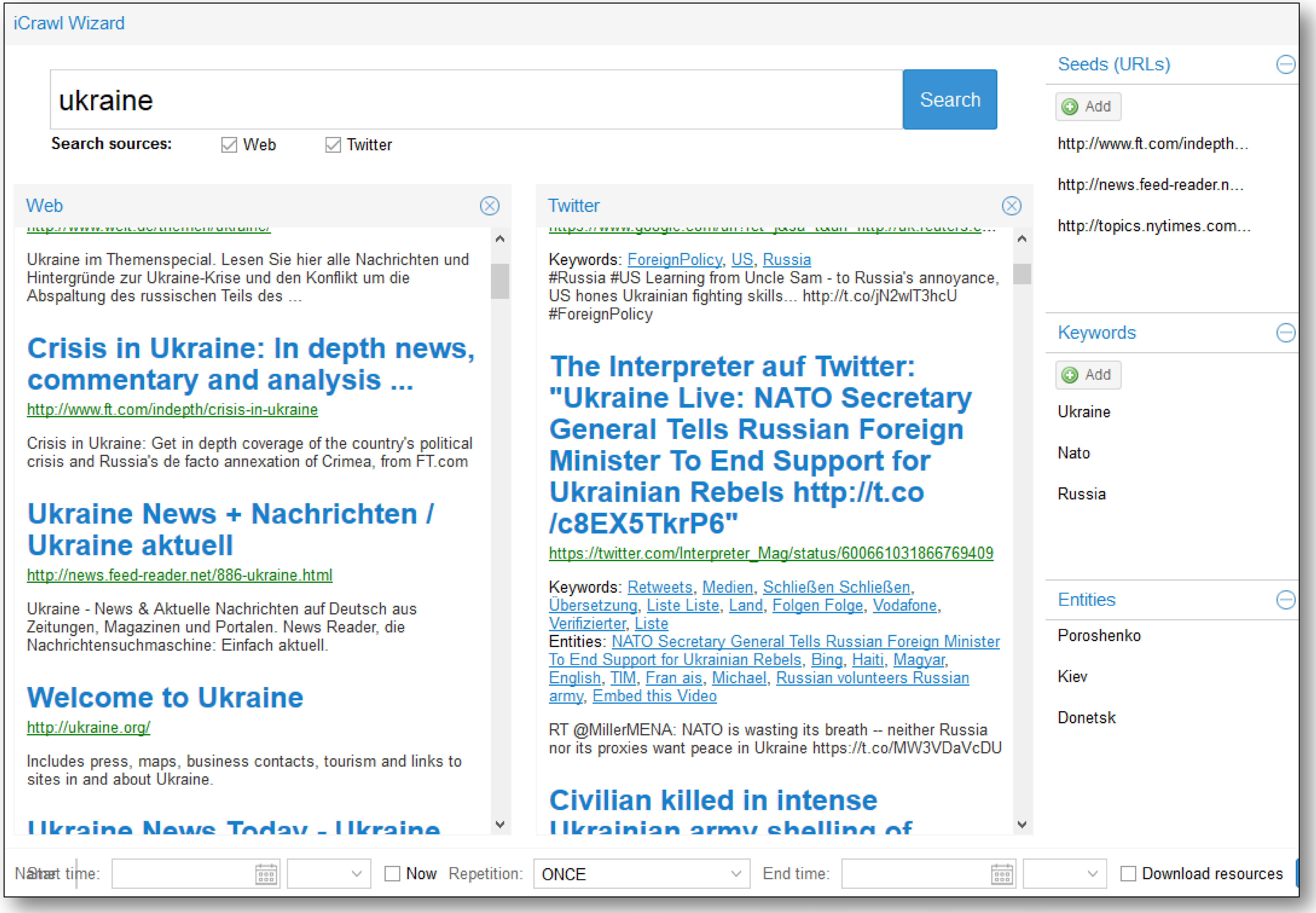
Integrated crawling with the L3S iCrawl System



L3S iCrawl System (under development)

- Learning the intention of the crawl
- Integration of Web and Social Media Crawling
- Content based monitoring of the crawl process

iCrawl Wizard



iCrawl Wizard

ukraine

Search sources: ☑ Web ☑ Twitter

Web

http://www.welt.de/themen/ukraine/

Ukraine im Themenspecial. Lesen Sie hier alle Nachrichten und Hintergründe zur Ukraine-Krise und den Konflikt um die Abspaltung des russischen Teils des ...

Crisis in Ukraine: In depth news, commentary and analysis ...

http://www.ft.com/indepth/crisis-in-ukraine

Crisis in Ukraine: Get in depth coverage of the country's political crisis and Russia's de facto annexation of Crimea, from FT.com

Ukraine News + Nachrichten / Ukraine aktuell

http://news.feed-reader.net/886-ukraine.html

Ukraine - News & Aktuelle Nachrichten auf Deutsch aus Zeitungen, Magazinen und Portalen. News Reader, die Nachrichtensuchmaschine: Einfach aktuell.

Welcome to Ukraine

http://ukraine.org/

Includes press, maps, business contacts, tourism and links to sites in and about Ukraine.

Ukraine News Today - Ukraine

Twitter

https://www.google.com/url?rct=j&sa=t&url=http://uk.reuters.c...

Keywords: ForeignPolicy, US, Russia
#Russia #US Learning from Uncle Sam - to Russia's annoyance, US hones Ukrainian fighting skills... http://t.co/jN2wlT3hcU #ForeignPolicy

The Interpreter auf Twitter: "Ukraine Live: NATO Secretary General Tells Russian Foreign Minister To End Support for Ukrainian Rebels http://t.co /c8EX5TkrP6"

https://twitter.com/Interpreter_Mag/status/600661031866769409

Keywords: Retweets, Medien, Schließen Schließen, Übersetzung, Liste Liste, Land, Folgen Folge, Vodafone, Verifizierter, Liste
Entities: NATO Secretary General Tells Russian Foreign Minister To End Support for Ukrainian Rebels, Bing, Haiti, Magyar, English, TIM, Fran ais, Michael, Russian volunteers Russian army, Embed this Video

RT @MillerMENA: NATO is wasting its breath -- neither Russia nor its proxies want peace in Ukraine https://t.co/MW3VDaVcDU

Civilian killed in intense Ukrainian army shelling of

Seeds (URLs)

http://www.ft.com/indepth...

http://news.feed-reader.n...

http://topics.nytimes.com...

Keywords

Ukraine

Nato

Russia

Entities

Poroshenko

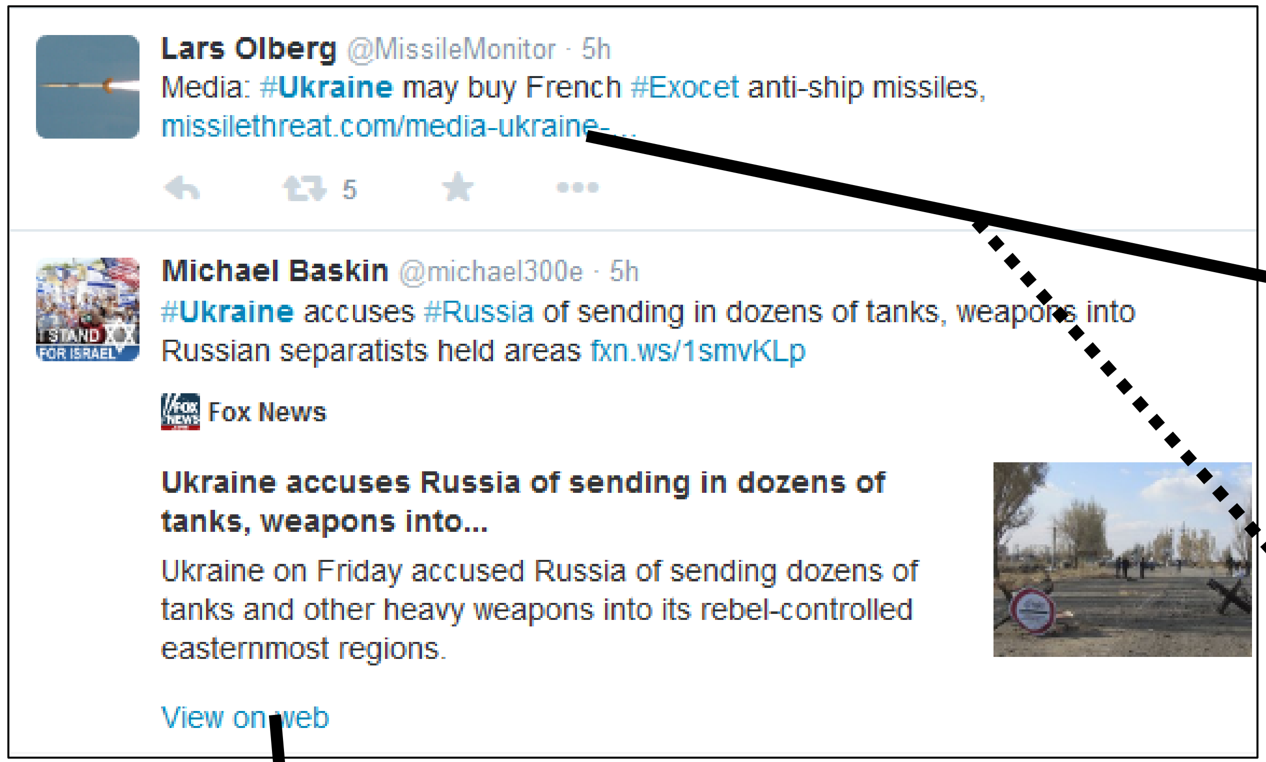
Kiev

Donetsk

Name Start time: ☐ Now Repetition: ONCE End time: ☐ Download resources

Example for Integrated Crawling

Twitter #Ukraine Feed



Lars Olberg @MissileMonitor · 5h
Media: #**Ukraine** may buy French #Exocet anti-ship missiles,
missilethreat.com/media-ukraine-...

Michael Baskin @michael300e · 5h
#**Ukraine** accuses #Russia of sending in dozens of tanks, weapons into Russian separatists held areas fxn.ws/1smvKLp

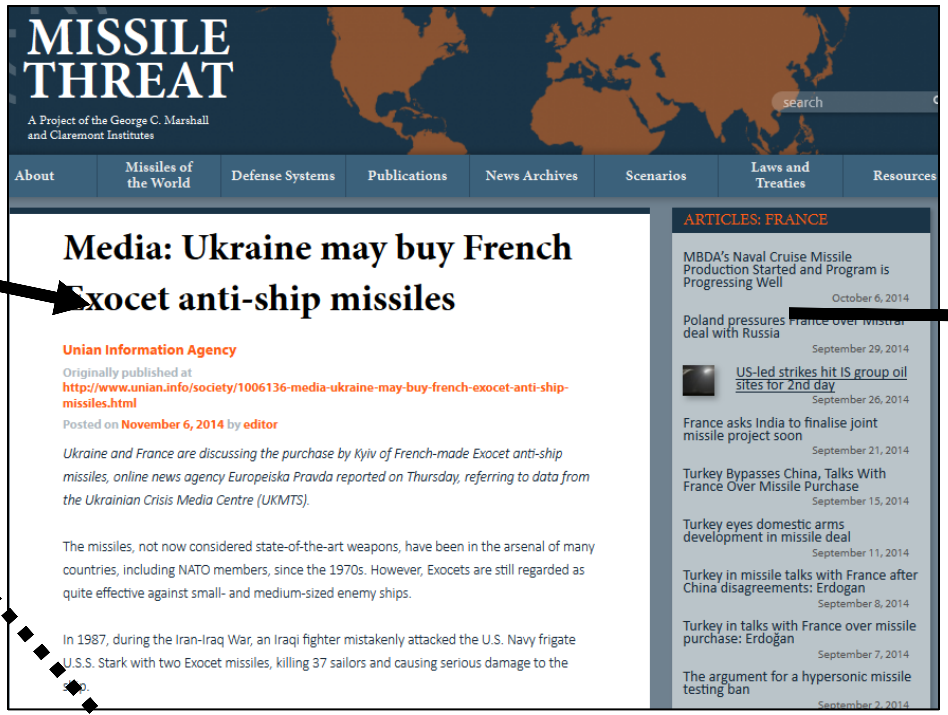
Fox News

Ukraine accuses Russia of sending in dozens of tanks, weapons into...

Ukraine on Friday accused Russia of sending dozens of tanks and other heavy weapons into its rebel-controlled easternmost regions.

View on web

(Medium Page Relevance)



Media: Ukraine may buy French Exocet anti-ship missiles

Unian Information Agency
Originally published at
http://www.unian.info/society/1006136-media-ukraine-may-buy-french-exocet-anti-ship-missiles.html
Posted on November 6, 2014 by editor

Ukraine and France are discussing the purchase by Kyiv of French-made Exocet anti-ship missiles, online news agency Europeiska Pravda reported on Thursday, referring to data from the Ukrainian Crisis Media Centre (UKMTS).

The missiles, not now considered state-of-the-art weapons, have been in the arsenal of many countries, including NATO members, since the 1970s. However, Exocets are still regarded as quite effective against small- and medium-sized enemy ships.

In 1987, during the Iran-Iraq War, an Iraqi fighter mistakenly attacked the U.S. Navy frigate U.S.S. Stark with two Exocet missiles, killing 37 sailors and causing serious damage to the ship.

ARTICLES: FRANCE

MBDA's Naval Cruise Missile Production Started and Program is Progressing Well
October 6, 2014

Poland pressures France over Mistral deal with Russia
September 29, 2014

US-led strikes hit IS group oil sites for 2nd day
September 26, 2014

France asks India to finalise joint missile project soon
September 21, 2014

Turkey Bypasses China, Talks With France Over Missile Purchase
September 15, 2014

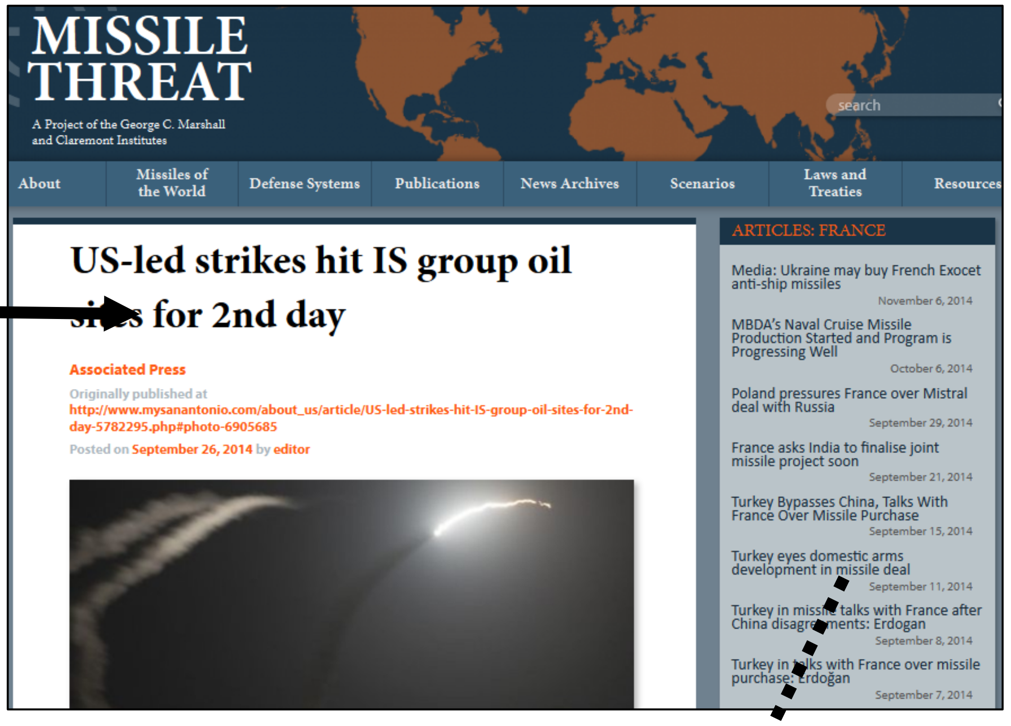
Turkey eyes domestic arms development in missile deal
September 11, 2014

Turkey in missile talks with France after China disagreements: Erdogan
September 8, 2014

Turkey in talks with France over missile purchase: Erdoğan
September 7, 2014

The argument for a hypersonic missile testing ban
September 2, 2014

(Low Page Relevance)



US-led strikes hit IS group oil sites for 2nd day

Associated Press
Originally published at
http://www.mysanantonio.com/about_us/article/US-led-strikes-hit-IS-group-oil-sites-for-2nd-day-5782295.php#photo-6905685
Posted on September 26, 2014 by editor

ARTICLES: FRANCE

Media: Ukraine may buy French Exocet anti-ship missiles
November 6, 2014

MBDA's Naval Cruise Missile Production Started and Program is Progressing Well
October 6, 2014

Poland pressures France over Mistral deal with Russia
September 29, 2014

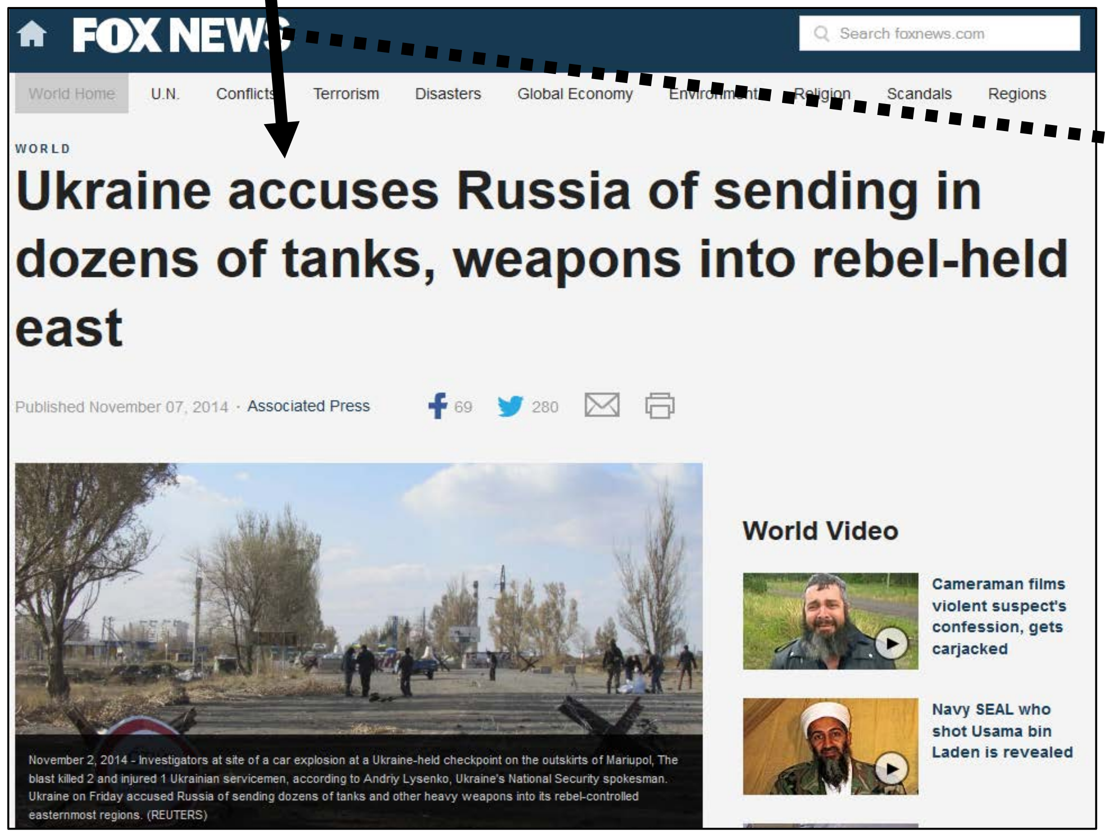
France asks India to finalise joint missile project soon
September 21, 2014

Turkey Bypasses China, Talks With France Over Missile Purchase
September 15, 2014

Turkey eyes domestic arms development in missile deal
September 11, 2014

Turkey in missile talks with France after China disagreements: Erdogan
September 8, 2014

Turkey in talks with France over missile purchase: Erdoğan
September 7, 2014



WORLD

Ukraine accuses Russia of sending in dozens of tanks, weapons into rebel-held east

Published November 07, 2014 · Associated Press

November 2, 2014 - Investigators at site of a car explosion at a Ukraine-held checkpoint on the outskirts of Mariupol, The blast killed 2 and injured 1 Ukrainian servicemen, according to Andriy Lysenko, Ukraine's National Security spokesman. Ukraine on Friday accused Russia of sending dozens of tanks and other heavy weapons into its rebel-controlled easternmost regions. (REUTERS)

World Video

Cameraman films violent suspect's confession, gets carjacked

Navy SEAL who shot Usama bin Laden is revealed

(High Page Relevance)

Crawler Queue

ID	Batch	URL	Priority
UK1	1	http://www.foxnews.com/world/2014/11/07/ukraine-accuses-russia-sending-in-dozens-tanks-other-heavy-weapons-into-rebel/	1.00
UK2	1	http://missilethreat.com/media-ukraine-may-buy-french-exocet-anti-ship-missiles/	1.00
UK3	x	http://missilethreat.com/us-led-strikes-hit-group-oil-sites-2nd-day/	0.40
UK4	y	http://missilethreat.com/turkey-missile-talks-france-china-disagreements-erdogan/	0.05
		…	…

→ Web Link ⋯→ Extracted URL

Conclusions

Social Media Preservation

- Social Media can provide more then short term views
- Social Media preservation enable long term studies

Social Media Crawling

- Twitter crawls should include the context
  - Context of the content
  - Visual presentation

Freshness of Content

- Context of an event can evolve of time
- Social Media might point to the wrong context
- Limiting the time gap between Social Media and Web crawling

iCrawl System

- Under development
- Will be integrated into the SoBigData Research Infrastructure

Thank You!

Dr. Thomas Risse
Forschungszentrum L3S
Leibniz Universität Hannover
Appelstrasse 9a
30167 Hannover, Germany

E-Mail: risse@L3S.de
Telefon: +49-511-762 17764
Telefax: +49-511-762 17779