



**Extracting intelligence from multilingual
SMS, IM, e-mails...**

Agenda

<http://scanandtarget.com/> - contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



Scan & Target presentation

Mass interception issues

Specificities for Arabic, Dialects and Arabish

Recommended approach

What's happening in 60 s on the web?

<http://scanandtarget.com/> - contact@scanandtarget.com

Scan & Target



Bla Bla Bla

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



Conversations represent a big chunk of this traffic



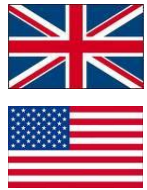
Help, Natural Language processing required!

<http://scanandtarget.com/>

contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



U don't got da jack but remember we got da
screenin 2mro at 8



C vré ke C pa + facil ! G mi $2x + 2$ tan a lir C 2
post en langaj SMS ke 2 posts ékri normleman



Hexo x ti y xa ti, tú pones las reglas



Sda7med ya 5ouya Ma chba3tech biiik allah
ghaleb...nchallah kol 3aam wenti 7ay b5iir



Who is Scan & Target?

<http://scanandtarget.com/>

contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



Scan & Target analyzes **digital communications in real time** to provide **actionable intelligence** to software vendors, brands, service publishers, marketing agencies, governments...



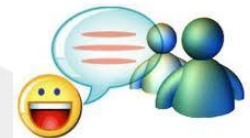
Social
networks



Forums, blogs



E-mails



Instant Messaging

Our text Meaning Technology is smart enough to look in real time at an incoming text User Generated Content data stream, **see patterns of interest**, and **alert the right people** or trigger the appropriate action-- **all without being queried**

Customers

Real-Time Text Meaning



<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target





Scan & Target technology

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Unlike solutions based on simple keywords or semantic, our technology takes into account the **different alterations and variants** of expressions to analyze the content:

- Small/ capital letters use
- Letters repetition (vvviiiagrrra for example)
- Orthographical variations (vi@gra, vIagra, v1@gra, v149r4)
- Missing letters in some cases (v|agra, v agra...)
- Word alteration whatever the use of non alpha symbol (v.i.a.g.r.a, v_i°ag#r:a, v-iagra, viagr"a...)
- Phonetic alterations
- SMS and IM languages
- And the combination of these variations

The solution is available in **English** and **French** and **Spanish** and **Arabic (MSA + dialects, Arabic alphabet + transliteration)**.



Scan & Target technology

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

The solution is based on a smart engine that rates not just single words but the entire content as it passes through the filtering engine. **Words** are therefore **placed in context** to **extract meaning**

The solution applies detailed thematic thesauruses - our **Smart Wordbooks**. Filters are categorized to allow customers to fine-tune the analysis (Terrorism/Drugs/Violence, etc.) according to their needs

Additional analysis layers: **sentiment analysis, questions detection...**

Proprietary scoring technology tailored to short digital text contents

Using a powerful and accurate conditional analysis system, our customers experience a **very low level of false positives** (between 0,05% to 0,001% in average)

What can we find for you?

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

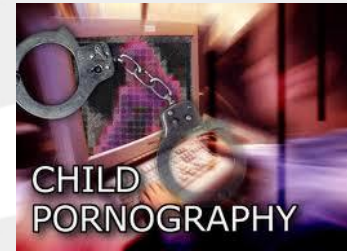
Real-Time Text Meaning



Drugs traffic



Incitement of violence



Corruption



Online prostitution



Smuggling



Big Data? No problem.

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



- For homeland security, our API is distributed using IBM hardware (to be hosted on your premises)
- Thanks to our connector, it's very easy to implement our API into your own applications
- You choose how to display our analysis results into your interfaces
- Capacity to deal in real time with Big Data
 - All of Twitter's traffic (10 TB / day, average 1200 Tweets per second)* could be analyzed in real time using one IBM blade center (for one language)
 - *Source - Twitter

Agenda

<http://scanandtarget.com/> - contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



Scan & Target presentation

Mass interception issues

Specificities for Arabic, Dialects and Arabish

Recommended approach

Mass interception issues

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



- Mass interception of digital text communications, (OSINT or COMINT like SMS, e-mails, IM...) is now technically available
- Issues for intelligence or law enforcement agencies:
 - How to deal with the volume (flow never stops)
 - How to find the needle in the digital haystack

“Finding the needle” strategies

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning





Benefits	Identified Suspects	Interception on keywords	Indexation and search	Text Meaning
Real time information		+	-	+
Fuzzy search	-	-		+
Advanced analysis	-	-	+	+
False positive ratio		-		+
Unknown threat detection	-		+	+
Required analyst time		-	-	+



Strategies comparison on OSINT

<http://scanandtarget.com/>
- contact@scanandtarget.com
Scan & Target

Service / % alerts	Keywords	Indexing	Text Meaning
BlueLight.ru Drugs forum 	13%	6.5%	<1%
Gaia Online 	19%	11%	<2%

Agenda

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



Scan & Target presentation

Mass interception issues

Specificities for Arabic, Dialects and Arabish

Recommended approach

Arabic usage

<http://scanandtarget.com/>

contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



Top Ten Languages Used in the Web
(Number of Internet Users by Language)

TOP TEN LANGUAGES IN THE INTERNET	Internet Users by Language	Internet Penetration by Language	Growth in Internet (2000 - 2001)	Internet Users by Language	World Population for this Language (2010 Estimate)
English	536,564,837	42.0 %			1,277,528,133
Chinese	444,948,013	32.6 %			1,365,524,982
Spanish	153,309,074	36.5 %			420,469,703
Japanese	99,143,700	78.2 %			126,804,433
Portuguese	82,548,200	33.0 %	96.2 %		250,372,925
German	75,158,584	78.6 %	173.1 %	3.8 %	95,637,049
Arabic	65,365,400	18.8 %	2,501.2 %	3.3 %	347,002,991
French	59,779,500	17.2 %	398.2 %	3.0 %	347,932,305
Russian	50,000,000	42.8 %	1,825.8 %	3.0 %	139,390,205
Korean	40,000,000	55.2 %	107.1 %	2.0 %	71,393,343

Arabic is the fastest growing language in the Web

With one of the lowest penetration rate

Arabic principles

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



- Arabic is used to describe 3 different forms of the same language:
 - Classical Arabic: used in the Qur'an and classical literature
 - Modern Standard Arabic (MSA):
 - ✓ no one's native spoken language any more
 - ✓ Form of Arabic taught in schools and used in newspapers, books, sermons, TV...
 - ✓ The most widely understood type of Arabic used in conversation between educated Arabs from different countries
 - Colloquial or Dialectal Arabic: national or regional varieties derived from Classical Arabic, which constitute the everyday spoken language

<http://scanandtarget.com/> - contact@scanandtarget.com



-

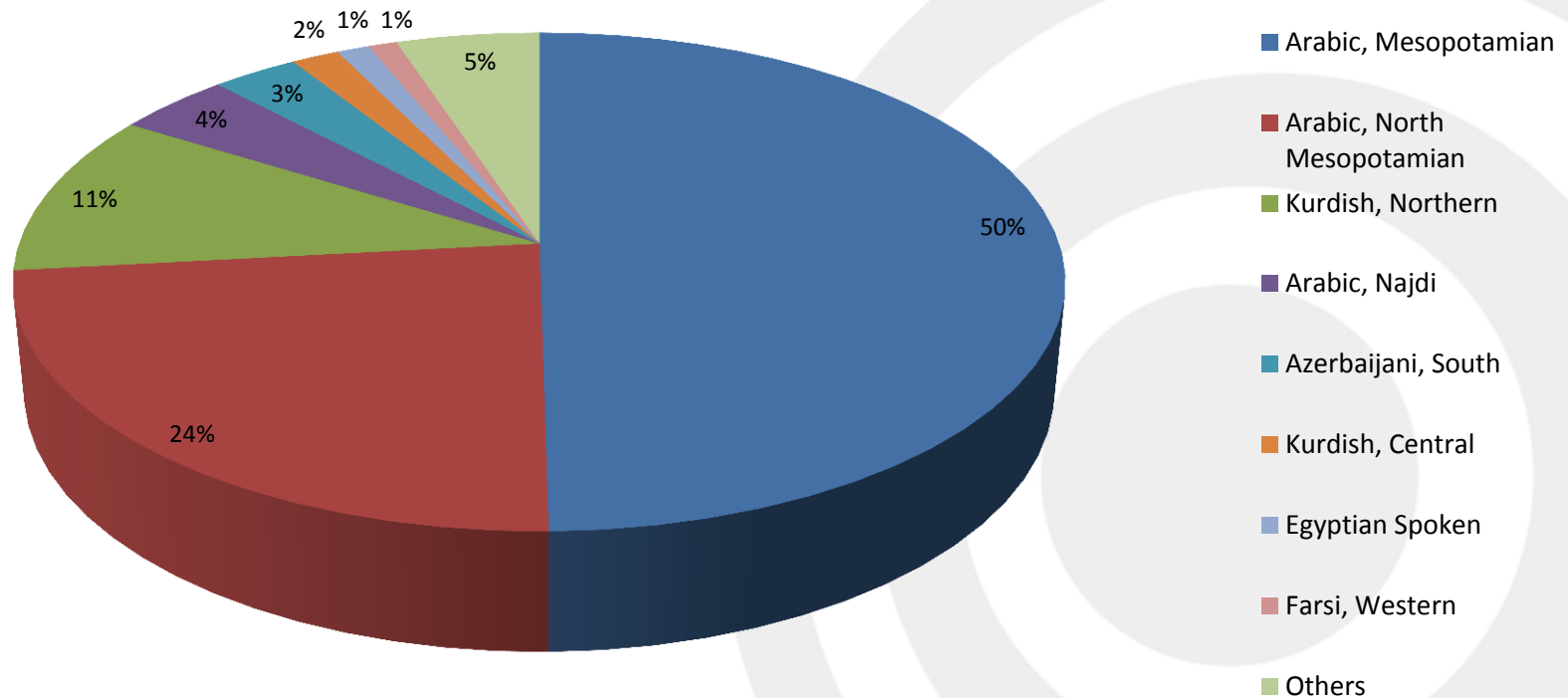
Iraq languages

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



Dialects example

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



English Sentence:	I want	to drink	water
Standard Arabic Transliteration	Ureedu	an ashrafa	ma'an
Egyptian Transliteration:	Awez	ashrab	mayya
Syrian Transliteration:	Beddy	eshrab	Mayy
Saudi Transliteration:	Abgha / Areed	Ashrab	Mayyeh
Moroccan Transliteration:	Bghit	Neshrab	Elma

Transliteration

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



- Transliteration is the romanization of Arabic
 - From قهوة to Gahwa (Coffee)
- Problem: written Arabic is normally unvocalized , i.e., the vowels are not written out, and must be supplied by a reader familiar with the language

Arabic chat alphabet

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



- The Arabic chat alphabet (Arabish or Arabizi) is used to communicate in the Arabic language over the Internet or for sending messages via mobile phones when the Arabic alphabet is unavailable
- Arabic letters are replaced by letters that are phonetically equivalent
- Arabic letters that have no Latin phonetic counterpart are represented by numbers, or numbers in conjunction with an accent mark

Issues with Arabic compared to latin languages

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



- Language identification issue:
 - MSA, dialects, mix of languages
- Transliteration issue (notably for names)
 - ABD AL-WADOUB
 - ABD EL OUADOUD
 - ABD-AL-WADUD
 - ABDEL EL-WADOUD
- Use of Arabish / Arabizi
 - bri6ania al3o'6ma / britanya al 3ozma = Great Britain for example

Our Text Meaning
Technology handles
all these issues

Agenda

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



Scan & Target presentation

Mass interception issues

Specificities for Arabic, Dialects and Arabish

Recommended approach

Text meaning mission

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



- To identify and destroy terrorist / criminal networks, you must detect the mistakes / errors they will make
- This is the job of text meaning : bringing actionable intelligence to the analyst for investigation

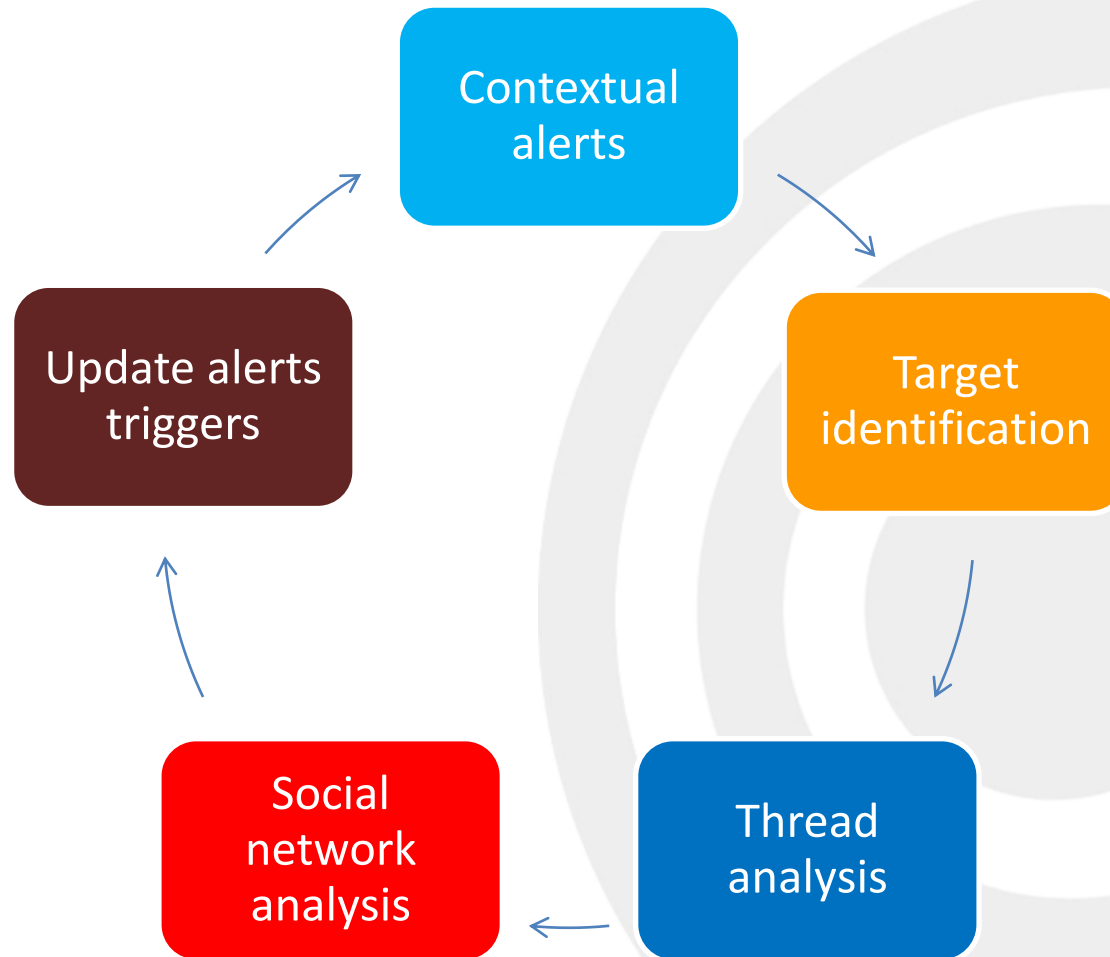


New threat detection

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target





Messages vs thread

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

- A web or mobile conversation is a thread of messages between 2 or more persons
- Analysis is first performed at message level for contextual alerts
- When an alert is detected, the associated discussion thread is again analyzed to:
 - Increase accuracy and precision
 - Extract investigation elements (names, places, nationality, places...)

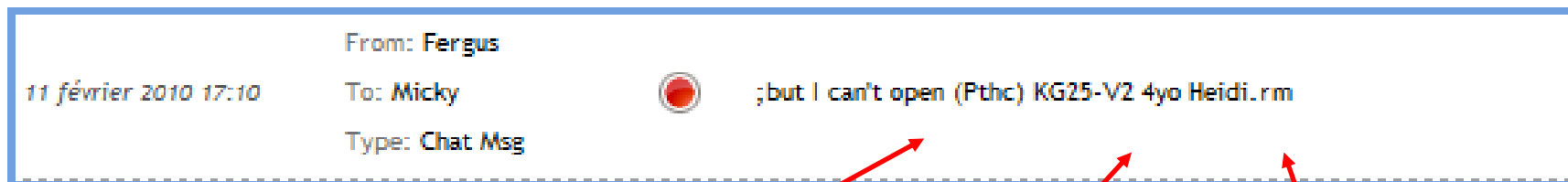
Message identification: paedophilia

<http://scanandtarget.com/>

contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



PTHC =
Pre Teen Hard Core

Age detection

Multimedia content
extension detection

= automatic contextual
alert sent for potential
child pornography



Thread expansion: paedophilia

<http://scanandtarget.com/> - contact@scanandtarget.com

Scan & Target

(11 février 2010 17:10) Fergus:  ;but I can't open (Pthc) KG25-V2 4yo Heidi.rm

(11 février 2010 17:10) Fergus: ;unknown format

(11 février 2010 17:10) Micky: ;you need real player

(11 février 2010 17:10) Fergus: ;ah

(11 février 2010 17:10) Micky: ;it is good

(11 février 2010 17:10) Micky: ;kg 25 Regina shows her ass

(11 février 2010 17:10) Micky: ;all still critical error :(

(11 février 2010 17:10) Fergus: ;yes :(

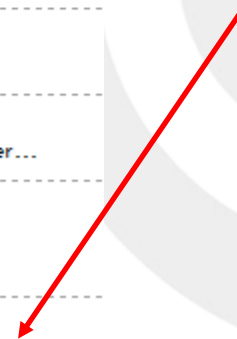
(11 février 2010 17:10) Micky: ;:'(

(11 février 2010 17:10) Fergus: ;don't worry, it will work tommorow... or later...

(11 février 2010 17:10) Micky:  ;other non pic forum works

(11 février 2010 17:10) Fergus:  ;Join this forum: <http://shafeclub.forum24.ru/>

Investigation element:
Forum to be
investigated



Use case: drugs traffic detection

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



- Mass Surveillance of SMS communications (20 to 30 millions per day with a lot of different languages, English, Arabic, dialects...)
- Contextual alerts sent to analysts using conditional analysis on:
 - Substance related discussions,
 - Transaction related discussions (quantities, money...)
 - Middle men related discussions (dealers, luggage handler, docker, customs...)
 - Smuggling related discussions (places like ports, airports and smuggling tricks)
- Investigation by analyst (conversation thread analysis, social network analysis...) identifies:
 - Dealers' ring (pseudo, IP address...)
 - Coded language detection (use of culinary vocabulary for example)
- High precision: 40 alerts per million SMS

Recommended solution

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target

Real-Time Text Meaning



- Scan & Target text meaning technology is a very efficient tool to detect previously **unknown terrorist or criminal threats** on the Internet or wireless networks
- Main benefits:
 - Ability to deal with **huge volumes** in **real time**
 - Multilingual and ability to **manage fuzzy languages like IM or arabizi**
 - **Actionable intelligence** with message & thread analysis
 - **Low level of false positive** thanks to advanced analysis
- To be integrated into your existing monitoring system

Contact Information

<http://scanandtarget.com/>

- contact@scanandtarget.com

Scan & Target



Bastien Hillen, CEO

[Phone] + 33 6 11 25 53 80

b.hillen@scanandtarget.com

Scan & Target

80 rue des haies

75020 Paris

France

www.scanandtarget.com

www.oorook.com

