

# Big Data Triage with Text Analytics

**Steve Kearns**

Director of Product Management

[www.basistech.com](http://www.basistech.com)



# Agenda

- About Basis Technology
- Challenges of Big Bata
- Text Analytics Technology
- Text Analytics for Big Data Triage

# About Basis Technology

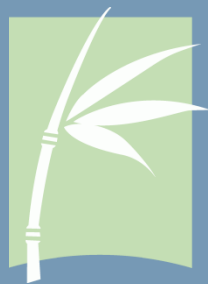
- Specialists in human language technology, as applied to web and enterprise search, OSINT/DOCEX/MEDEX, e-discovery, and digital forensics
- Developers of the most capable, most mature, and most widely used platform for multilingual text analytics
- Solutions for government agencies dealing with multi-source intelligence and large data sets

# Customers

- Central Intelligence Agency (CIA)
- Defense Intelligence Agency (DIA)
- Department of Defense (DOD)
- Federal Bureau of Investigation (FBI)
- National Security Agency (NSA)
- “International police agency”
- French MOD
- Japanese MOD
- Singapore CSIT



# What is Big Data?

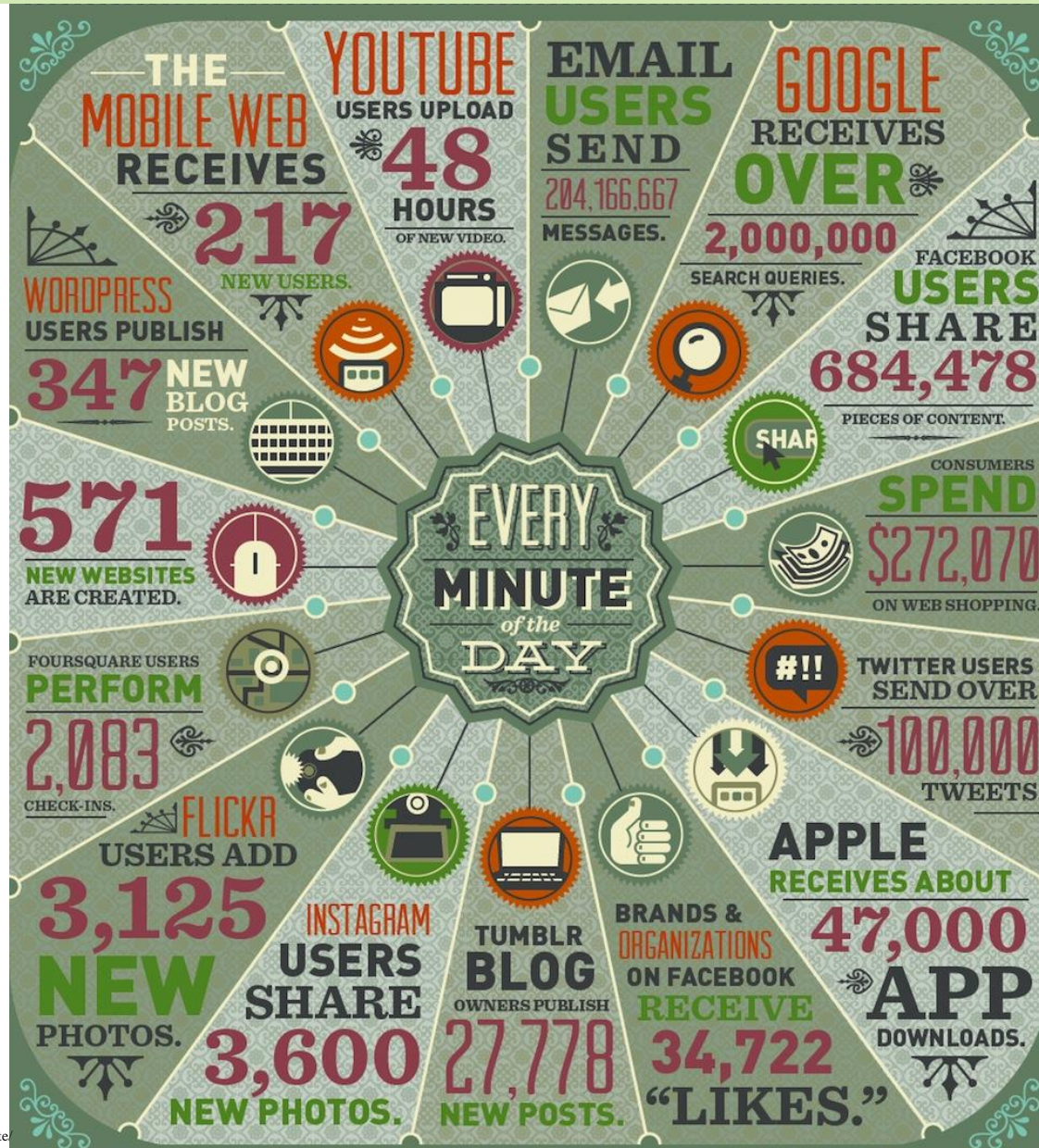


**BASIS**  
TECHNOLOGY

# Big Data

- Volume
- Velocity
- Variety

# Volume



# Velocity

- High-Throughput Sources:
  - ⇒ *Digital Forensics*
    - Rapid Site Exploitation
    - Many Hard Drives
- Rapidly Changing Sources:
  - ⇒ *News*
  - ⇒ *Social Media*
  - ⇒ *Network traffic*
- High Throughput Storage, Analysis, Alerting



# Variety

- Data Types

- ⇒ *DOMEX/DOCEX/MEDEX/OSINT*

- ⇒ *Finished Intel*

- ⇒ *Cables*

- ⇒ *Harmony*

- ⇒ *Biometrics*

- ⇒ *Watch Lists*

- ⇒ *Hard Drive -> File(s) -> Unstructured and Structured Content*

- ⇒ *Sensor Data*

- Structured / Unstructured

- Textual / Visual / Numeric

# The Challenge: Finding Value



# Big Data Problems - Volume

- Where/How do you store it?
  - ⇒ *Single database -> database cluster -> Hadoop/HDFS?*
- Data quality?
  - ⇒ *Manual review or annotation?*
  - ⇒ *People don't scale*
- Query
  - ⇒ *If you can, how fast, how complex and on what can you query?*
  - ⇒ *User Interface? SQL? Programming?*
  - ⇒ *How do you view results?*
  - ⇒ *Can you filter the results to refine your query?*
  - ⇒ *Thematic exploration, where the results of one query inform the next*
  - ⇒ *Security?*

# Big Data Problems - Velocity

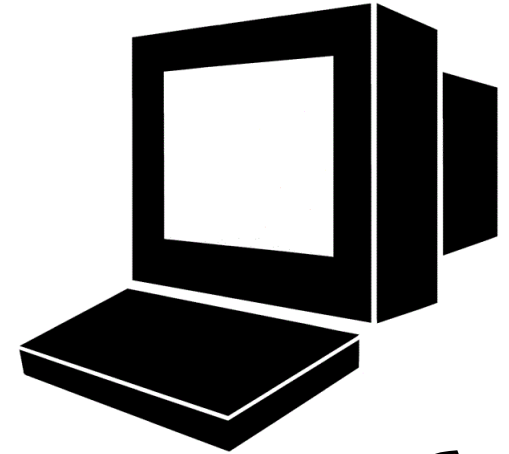
- Time sensitive
  - ⇒ *Value of information decreases over time*
  - ⇒ *How long from “publish” to “discoverable”?*
- Rapid changes/updates
  - ⇒ *Which updates are important?*
  - ⇒ *Which sources/users are important? Which may become important?*
  - ⇒ *Individual pieces of data may be meaningless, but what about in aggregate?*
  - ⇒ *Quality/Verification?*
  - ⇒ *Manual Review?*

# Big Data Problems - Variety

- Many Sources
  - ⇒ *Often stored, formatted, and accessed differently*
  - ⇒ *Access, security?*
  - ⇒ *Many languages*
  - ⇒ *How reliable is each source?*
- Few, if any, links
  - ⇒ *Between sources*
  - ⇒ *Between documents*
  - ⇒ *Between information within documents*

# General Problem

- Computers are great at some things
- Humans are great at others



2 + 2



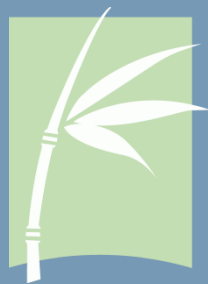
Scale



Human  
Language



# Text Analytics



**BASIS**  
TECHNOLOGY

# Text Analytics

Automated analytical methods operating on the written word to surface insights about the data.

It's purpose is to assist the human in finding things of relevance and interest.



# Text Analytics techniques

LEMMATIZATION  
PART-OF-SPEECH-TAGGING NAME-MATCHING  
PRONOUN-RESOLUTION  
STEMMING SENTIMENT-DETECTION  
ENTITY-EXTRACTION  
WORD-DECOMPOUNDING DOCUMENT-CLUSTERING  
LANGUAGE-IDENTIFICATION  
CO-REFERENCE WORD-SEGMENTATION  
RESOLUTION FACT-EXTRACTION  
RELATIONSHIP-EXTRACTION



# Triage Example

ARTICLE C  
boston

HOME TODAY  
Local National

HOME / COLLECTION

**Al Qaeda**

March 21, 2012 | A

Recommend

BAGHDAD (AP) — A series of bloody attacks that killed 46 people and injured more than 200 will be for next several days.

The attacks on police headquarters and a government building in Baghdad. In all, insurgents

Tanzim Qaidat al-Jihad fi Bilad al-Rafidayn  
(Organization of Jihad's Base in Mesopotamia)  
"Al-Qaeda in Iraq"

Participant in the Iraq War



Jihadist black flag

Active

Leaders [Abu Musab al-Zarqawi](#) †(2004–2006)  
[Abu Ayyub al-Masri](#) †(2006–2010)  
[Huthaifa al-Batawi](#) †(2010–2011)  
[Abu Dua](#) 2011–

Headquarters Formerly [Fallujah](#), central Iraq  
Later [Mosul](#), northern Iraq

Area of operations Iraq, Syria,<sup>[1]</sup> limited activity in the broader Middle East

Strength More than 1,000 "core" members (2005)<sup>[2]</sup>  
Close to 10,000 (including part-time fighters) at its height (2010)<sup>[3]</sup>  
~1,000 (2011)<sup>[4]</sup>

Part of [Al-Qaeda](#) (since 2004)  
[Mujahideen Shura Council](#) (2006)  
[Islamic State of Iraq](#) (since 2006)

Originated as [Jama'at al-Tawhid wal-Jihad](#) (Group of Monotheism and Jihad, 2003–2004)

es  
A&E THING  
Obituaries Spec

reb

Print

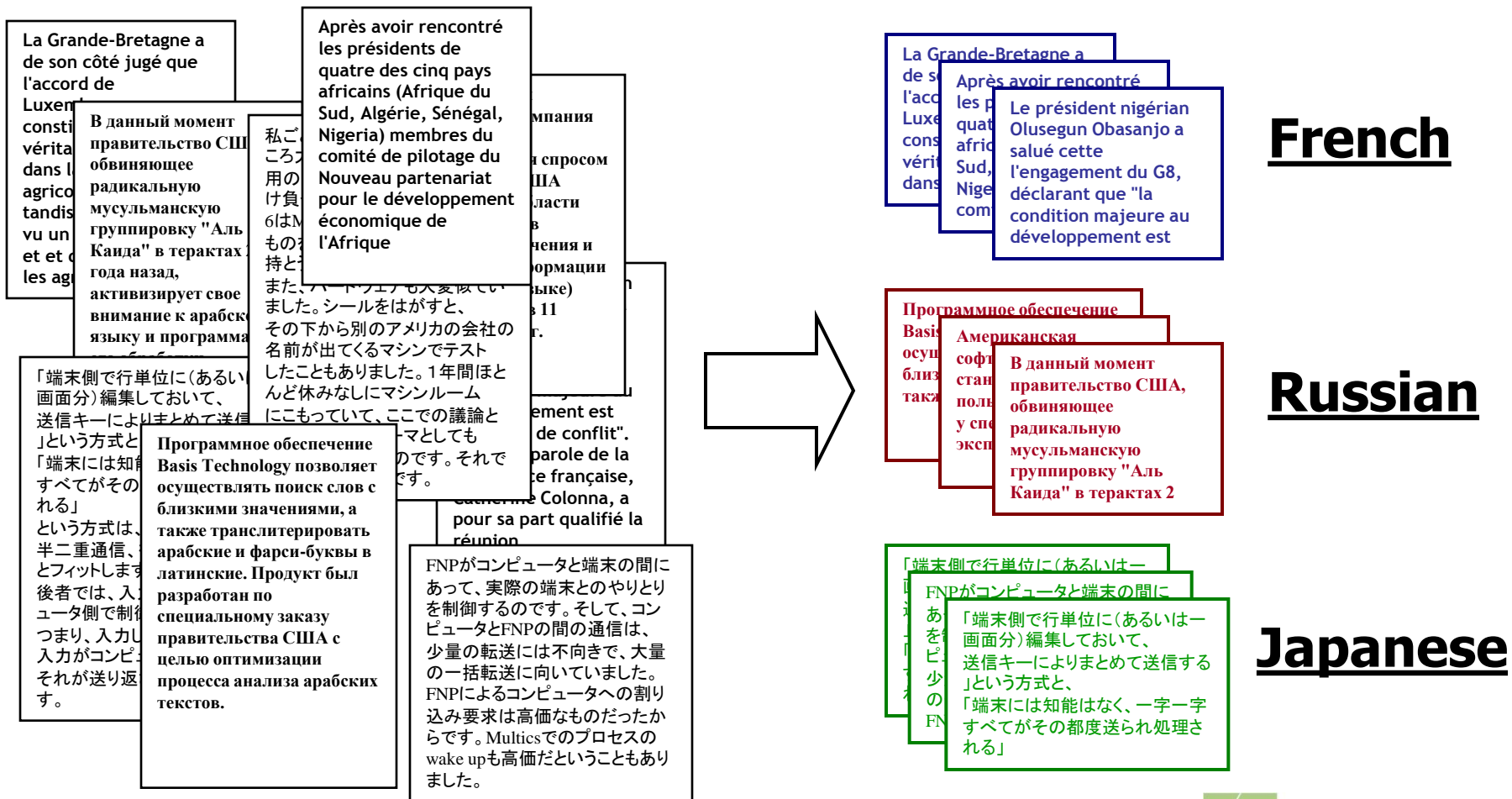
of bloody attacks  
ment security

China

rs on fire near a  
icials in Baghdad.  
nding 200 people.

IS TECHNOLOGY

# Text Analytics : Language ID



# Text Analytics: Lemmatization

Search

## Results

---

fly	132 hits
flying	97 hits
flew	78 hits
flown	61 hits

# Text Analytics: Lemmatization (Arabic)

فجر (Detonated)

Search

## Results

---

وتفجيرها	132 hits
متفجرات	77 hits
تفجيرات	32 hits
فجرها	22 hits
تفجرت	2 hits

# Text Analytics: Entity Extraction

### Document Information

Encoding:	US-ASCII
Primary Language:	English
Primary Script:	Latin

### Language Identification

ENGLISH (1)

### Base Linguistics

### Entity Extraction

- LOCATION (3)**
  - Baghdad
  - Haifa Street
  - Iraq
- ORGANIZATION (3)**
  - AFP
  - Al Qaida
  - interior ministry
- PERSON (1)**
  - Majid Hamed Ameen
- RELIGION (1)**
- TEMPORAL-DATE (4)**
- TEMPORAL-TIME (1)**
- TITLE (2)**

**Iraq's** deadliest day since **March**.

Multiple attacks in six provinces.

AFP Published: **00:00 April 20, 2012**.

**Baghdad** A wave of bombing and shooting attacks in six different provinces across **Iraq** killed at least 34 people and wounded more than 100 **Thursday**, security officials said.

It was the deadliest day in **Iraq** since **March 20**, when shootings and bombings claimed by **Al Qaida** front group the **Islamic** State of **Iraq** killed 50 people and wounded 255 nationwide.

Bombings in and around **Baghdad** killed at least 17 people and wounded 63, an **interior ministry** official said.

A car bomb targeting **Health Minister Majid Hamed Ameen's** convoy in **Haifa Street** in the heart of the capital, killed two civilians and wounded nine people, including four of the **minister's** guards.

# Text Analytics: Relationship Extraction

Relationships Add Relationships for:  **Add**

**Current Entity Relationships**

**Obama**

- communicatedWith**
  - [John O Brennan](#)
  - [82nd Airborne Division](#)
  - [Bashar al-Assad](#)
  - [Yoshihiko Noda](#)
  - [Barbara Walters](#)
  - [Lee \(Myung-Bak\)](#)
  - [John Boehner](#)
  - [Human Rights Watch](#)
  - [Kim Jong III](#)
- traveledTo**
  - [Fort Bragg](#)
- Also mentioned as:**
  - [Pres. Obama](#)
  - [Senator Obama](#)
  - [Mr Obama](#)
  - [president Obama](#)
  - [President Obama](#)
  - [Mr. Obama](#)

To display the relationship details, click on an Entity icon.

```
graph TD; Obama((Obama)) --- JohnO[John O Brennan]; Obama --- BarbaraW[Barbara Walters]; Obama --- LeeM[Lee (Myung-Bak)]; Obama --- NouriM[Nouri al-Maliki]; Obama --- 82nd[82nd Airborne Division]; Obama --- HRW[Human Rights Watch]; Obama --- YoshihikoN[Yoshihiko Noda]; Obama --- JohnB[John Boehner]; Obama --- FortBragg[Fort Bragg]; Obama --- KimJongIII[Kim Jong III]; Obama --- BasharA[Bashar al-Assad];
```

# Text Analytics: Entity Search

Search for:	Vincent Kaminski	
Score	Matching Name	Writing System
1.000	Vincent Kaminski	Latin
0.990	ヴィンセントカミンスキー	Katakana
0.990	ВИНСЕНТ КАМИНСКИЙ	Cyrillic
0.945	Vincent J. Kaminski	Latin
0.942	Vince Kaminski	Latin
0.921	V. Kaminski	Latin
0.910	Vincent Kaminsik	Latin
0.908	Vince Kamainski	Latin
0.903	Vince Kaminiski	Latin
0.903	Vince Kamnski	Latin
0.899	Vince Kaminsky	Latin
0.897	Vince Kaminki	Latin
0.893	文森特 卡明斯基	Hanzi



# Text Analytics: Document Clustering



New York T...

## Anxious Egyptians Jam Tahrir Square in Protest



New York Times - 1 hour ago

CAIRO - Tens of thousands of Egyptians packed into central Cairo's Tahrir Square on Friday in a collective spasm of last minute anxiety that Egypt's ruling generals might be trying to sabotage the transition to civilian democracy just as the ...

[Protests flare in Egypt, Bahrain](#) The Hindu  
[Peaceful Rally Marks Shift in Egypt](#)  
Wall Street Journal

Related  
[Egypt »](#)  
[Hosni Mubarak »](#)  
[Cairo »](#)

[From Egypt: Tahrir sees large turnout for Self-Determination Friday protest](#) Ahram Online

Opinion: [General Rules](#) Daily News Egypt

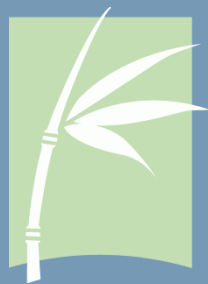
In Depth: [Thousands protest military's rule in Egypt](#) USA TODAY

[See all 329 sources »](#)



# Big Data Triage

## Text Analytics



**BASIS**  
TECHNOLOGY

# Big Data Processing

## Collect

- Identify data sources
- Data cleansing
- Move data into analysis repository

## Analyze

- Identify Entities, Facts, Relationships
- Link between Documents
- Link fact/entity between documents

## Index

- Keyword search + metadata filters
- Thematic exploration - using metadata
- Cross-document links

# Big Data Processing - Technology

## Collect

- Source: News, Twitter, Database, file system, digital forensics, etc.
- Storage: HDFS, MongoDB, SQL, etc.

## Analyze

- Platform: Hadoop, UIMA, Odyssey, Custom
- Analysis type: Language ID, Entity Extraction, Relationship Extraction, Document Clustering, Entity Linking

## Index

- Fulltext Search: Solr, Accumulo, Lucene
- Structured Data: RDF, SQL, OrientDB, Neo4j, Cassandra, HDFS, etc.

# Big Data Triage Requirements

- View results while still processing
  - ⇒ *Incremental collection/analysis/indexing*
- User Interface that allows exploration
  - ⇒ *Dashboard*
  - ⇒ *Keyword Search*
  - ⇒ *Geo Search*
  - ⇒ *Entity Search*
    - Enables thematic exploration
  - ⇒ *Metadata produced by Analysis makes this easier*

# Dashboard



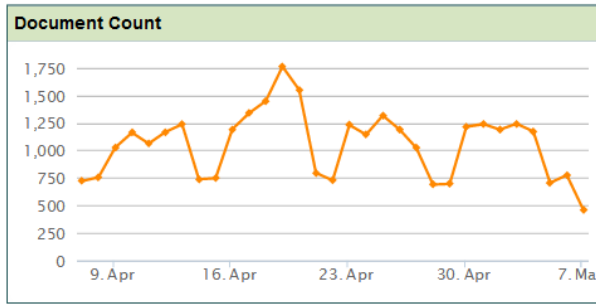
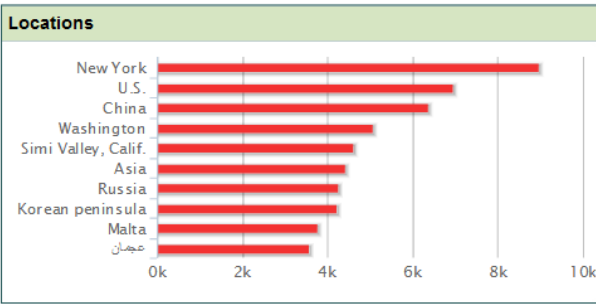
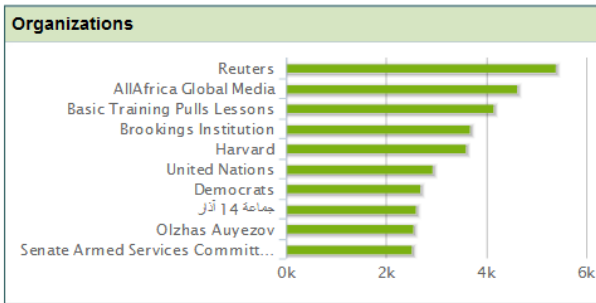
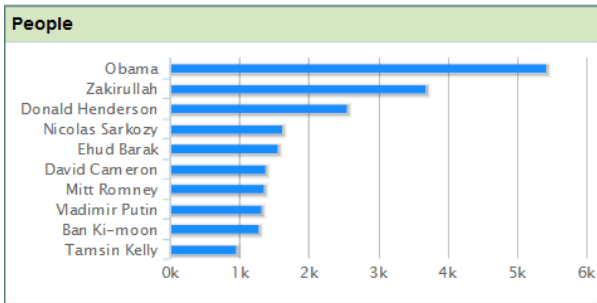
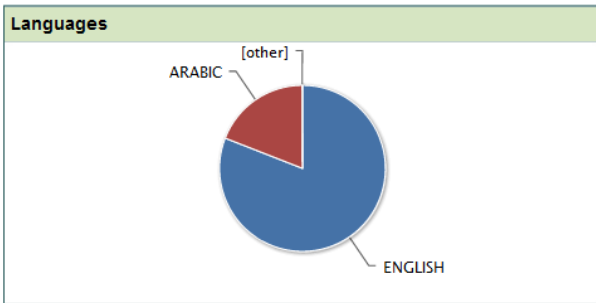
Solr Search Rosette Search Rosette Name Search

( Rosette language ID, morphological analysis, lemmatization, entity facets )

Search



Home



# Search and Filter

Search results | Odyssey Infr x Rosette Name Translator (a) x World x Dashboard | Odyssey Inform x

10.1.5.210:8090/ui/results.jsp?query\_mode=RLP&page=0&facet\_filter=&query\_string=Sudan+war

Solr Search Rosette Search Rosette Name Search

(Rosette language ID, morphological analysis, lemmatization, entity facets)

Sudan war Search

Odyssey INFORMATION NAVIGATOR BASIS TECHNOLOGY

Home > Search for [ Sudan war ]

**Filter Results By**

**Languages**

ENGLISH	1137
---------	------

**People**

Hryhory Sunda	466
محمد زعيبة	253
Obama	223
Penguin	166
Ban Ki-moon	165
Dries van Noten	161

**Organizations**

African Union	300
United Nations	296
Reuters	191
Democrats	180
AllAfrica Global Media	171
International Criminal Court	170

**Result groups 1 - 10 of 193** Page: 1 of 20

[World Briefing | Africa: Sudan: A Vow to 'Liberate' South Sudan](#)  
settlement ending a civil war. South Sudan last week seized Heglig, a contested oil-producing region  
Wed Apr 18 20:28:46 EDT 2012 - ENGLISH  
[590 similar articles](#)

[South Sudan quits captured oilfield](#)  
South Sudan quits captured oilfield South Sudan announced the immediate withdrawal on Friday  
Fri Apr 20 12:15:38 EDT 2012 - ENGLISH

[South Sudan: Presidential Statement On the Current Crisis](#)  
the border, we ask what the implications of war would be for both countries. From South Sudan is Atem Yah  
Thu Apr 12 09:37:54 EDT 2012 - ENGLISH  
[35 similar articles](#)

# Foreign Language Search



Keyword Search

Name Search

( Rosette language ID, morphological analysis, lemmatization, entity facets )

فجر

Search



Home > Search for [ فجر ]

## Filter Results By

### Languages

Arabic 154 🔍

### People

تويتر (Twitter) 25 🔍

القذافي (al-Qadhafi) 16 🔍

أ ف ب (A. F B) 14 🔍

لعمر (L'amir) 11 🔍

مبارك (Mubarak) 9 🔍

العريفي (al-'Arifi) 8 🔍

### Organizations

مجلس الأمة (Majlis al-Amah) 44 🔍

جريدة الوطن (Jaridah al-Watan) 34 🔍

Alwatan 26 🔍

القاعدة (Al-Qaeda) 10 🔍

الناتو (NATO) 10 🔍

طالبان (Taliban) 10 🔍

### Locations

## وفاة عربي اصطدمت سيارته بشجرة

وفاة عربي اصطدمت سيارته بشجرة لقي عربي مصرعه فجر أمس الأول، في حادث مروري إثر انحراف مركبته

Tue May 10 18:04:38 EDT 2011 - ARABIC

## شباب ثورة مصر يؤبنون ناصر الخرافي

م الساعة 11:03 . توقيت الصلاة . الفجر 3:51 . شباب ثورة مصر يؤبنون ناصر الخرافي .

Thu Apr 21 20:51:38 EDT 2011 - ARABIC

## شادي الخليج حزينا لفقد شقيقته

الساعة 11:48 . توقيت الصلاة . الفجر 3:47 . شادي الخليج حزينا لفقد شقيقته . شكرا لتصويت

Mon Apr 25 21:59:37 EDT 2011 - ARABIC

## انقطاع التيار الكهربائي في خورفكان 3 ساعات

الساعات الأولى من فجر أمس لأكثر من 3 ساعات في بعض المناطق. وبدأ الانقطاع في حوالي الساعة الواحدة والربع

Fri May 20 20:19:26 EDT 2011 - ARABIC

## عاشقان يمارسان أعمالاً (مؤجلة) بالشويخ

اسبويان (رجل وامرأة) فجر أمس في منطقة الشويخ الصناعية فتم استيقافها وبعد التدقيق على اثباتهما تبين ان

Mon May 23 20:00:20 EDT 2011 - ARABIC



# Detailed Document View



Home > Search for [ Osama Bin Ladin ] > Document: 07nordland.html.txt

Document Information	
Mime Type:	null
Encoding:	UTF-8
Primary Language:	English
Primary Script:	Latin

Language Identification
<input type="checkbox"/> ENGLISH (1)

Base Linguistics
<input type="checkbox"/> IDENTIFIER-LATITUDE_LONGITUDE (1)
<input type="checkbox"/> IDENTIFIER-NUMBER (10)
<input type="checkbox"/> IDENTIFIER-URL (1)
<input type="checkbox"/> LOCATION (13)
<input type="checkbox"/> NATIONALITY (5)
<input type="checkbox"/> ORGANIZATION (11)
<input type="checkbox"/> PERSON (21)
<input type="checkbox"/> RELIGION (4)
<input type="checkbox"/> TEMPORAL-DATE (4)
<input type="checkbox"/> TEMPORAL-TIME (1)
<input type="checkbox"/> TITLE (1)

## Simulcast - Obama's Cairo Speech - Forceful Words and Fateful Realities - NYTimes.com

Simulcast | Obama's Cairo Speech

Forceful Words and Fateful Realities

Published: June 6, 2009

BAGHDAD — Barack Obama's speech in Cairo last Thursday was "soft spoken and eloquent," said Moqtada al-Sadr, the radical Iraqi cleric, grudgingly, since he also said he despised it.

It was a speech that meant different things to different people, a quality that has been much noted in this president.

He supported Israel, but reached out to the Muslim world in an unprecedented way.

Some friends were troubled, others reassured.

Some of America's enemies denounced it, but none dismissed it.

Not even the arch-enemies at whom, in some important way, the speech was directed.

Multimedia

Related

(June 5, 2009)

(June 4, 2009)

Just the day before, in fact, a pre-emptive audio tape attributed to Osama bin Laden warned his followers not to trust whatever Mr. Obama would say.

And as it turned out, his fear was justified.



# Entity Search – Cross Language



Keyword Search

Name Search

( Rosette language ID, fuzzy name search, entity facets )

Mohammed Abdulla Saleh

Search



[Home](#) > Search for [ Mohammed Abdulla Saleh ]

## Filter Results By

### Languages

English	413	🔍
Arabic	32	🔍

### People

Ali Abdullah Saleh	387	🔍
Obama	49	🔍
Sadiq al-Ahmar	46	🔍
Nasser Arrabyee	43	🔍
Osama bin Laden	38	🔍
Sadeq al-Ahmar	35	🔍

### Organizations

Reuters	126	🔍
Gulf Cooperation Council	115	🔍
GCC	114	🔍
Al Jazeera	88	🔍
Al Qaeda	71	🔍
al-Qaeda	69	🔍

### Locations

Yemen	388	🔍
Sanaa	182	🔍
Gulf	172	🔍
Saudi Arabia	153	🔍
Taiz	149	🔍
US	148	🔍

## U.S. backs Gulf initiative to end Yemen crisis

from office by [Ammar Mohammed Abdullah Saleh](#), head of the National Security, and [Yahya Mohammed](#)

Sat Apr 23 15:34:19 EDT 2011 - ENGLISH

[23 similar articles](#)

## Al-Qaeda will be stronger in Yemen after President Saleh, e ...

strengthen in Yemen after collapse of the regime of President [Ali Abdullah Saleh](#), said an expert on Al Qaeda

Wed Apr 20 22:29:26 EDT 2011 - ENGLISH

[28 similar articles](#)

## Saleh resignation proposal agreed

soon under which President [Ali Abdullah Saleh](#) will step down after 30 days, officials say. Mr Saleh's

Tue Apr 26 13:16:05 EDT 2011 - ENGLISH

[75 similar articles](#)

## Saleh apologises to UAE for embassy blockade

Saleh apologises to UAE for embassy blockade DUBAI - Yemeni President [Ali Abdullah Saleh](#)

Tue May 24 03:03:33 EDT 2011 - ENGLISH

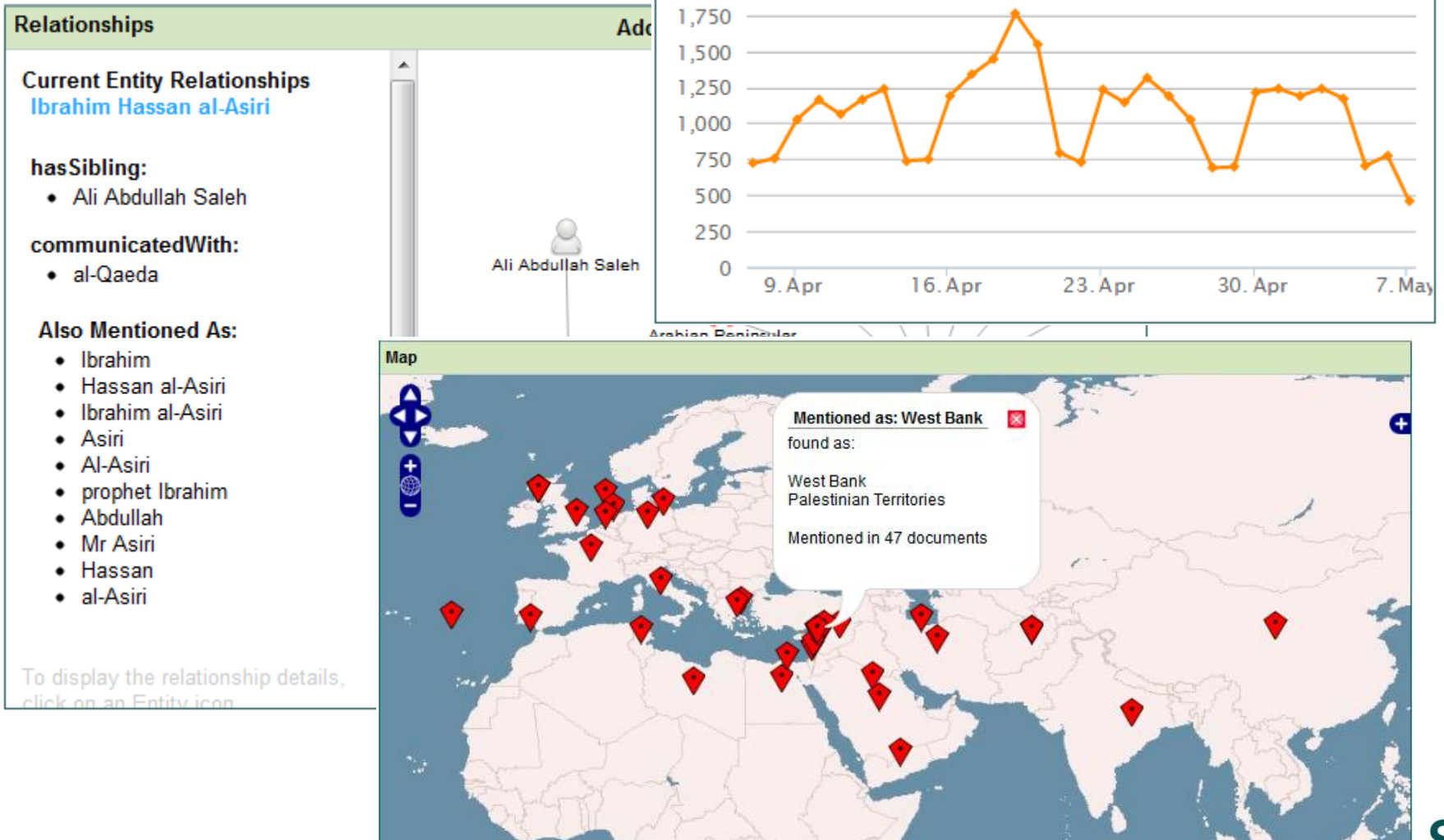
## Related Entities

Mohammed Abdulla Saleh	1.0	🔍
محمد صالح (Muhammad Salih)	0.79	🔍
Mohammed Abdullah	0.72	🔍
Ali Abdulla Saleh	0.70	🔍
Ammar Mohammed Abdullah Saleh	0.68	🔍
Yahya Mohammed Abdullah Saleh	0.68	🔍
Mohammad Saleh	0.68	🔍
Mohammad Abdullah	0.63	🔍
محمد صالح السبتي (Muhammad Salih al-Sabti)	0.63	🔍
Abdullah Mohammed	0.62	🔍
Abdullah Saleh	0.62	🔍
Ali Abdullah Saleh	0.61	🔍
Mohammed Salem	0.60	🔍

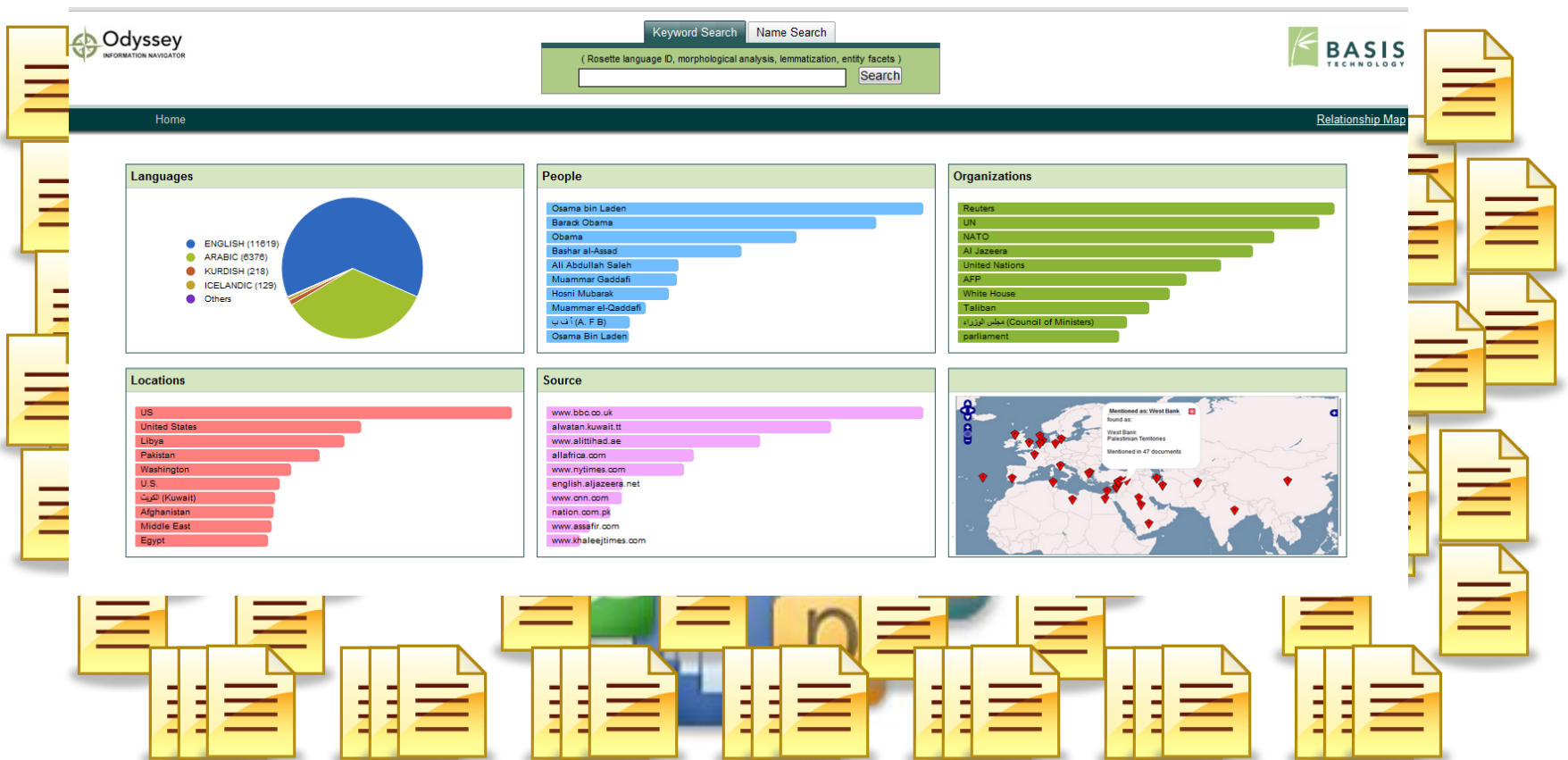
«عورة وينت عيد ودخلتها..»

سويغات من أجل قراءته وهو من اعداد وتجميع المحامي «محمد صالح السبتي» وبعنوان .. «ألف مثل من عيون الأمثال»

# Search/Filter/Explore



# Summary



Text Analytics enables Big Data Triage

# Thank you!

- For more information:
- Visit [www.basistech.com](http://www.basistech.com)

