Big Data Triage with Text Analytics

Steve Kearns
Director of Product Management
www.basistech.com

Agenda

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

BASIS TECHNOLOGY

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

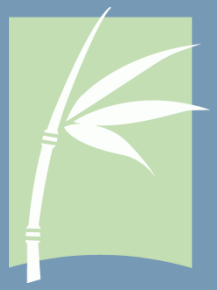
BASIS TECHNOLOGY

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

In-Q-Tel MITRE BAE SYSTEMS BASIS TECHNOLOGY

What is Big Data?



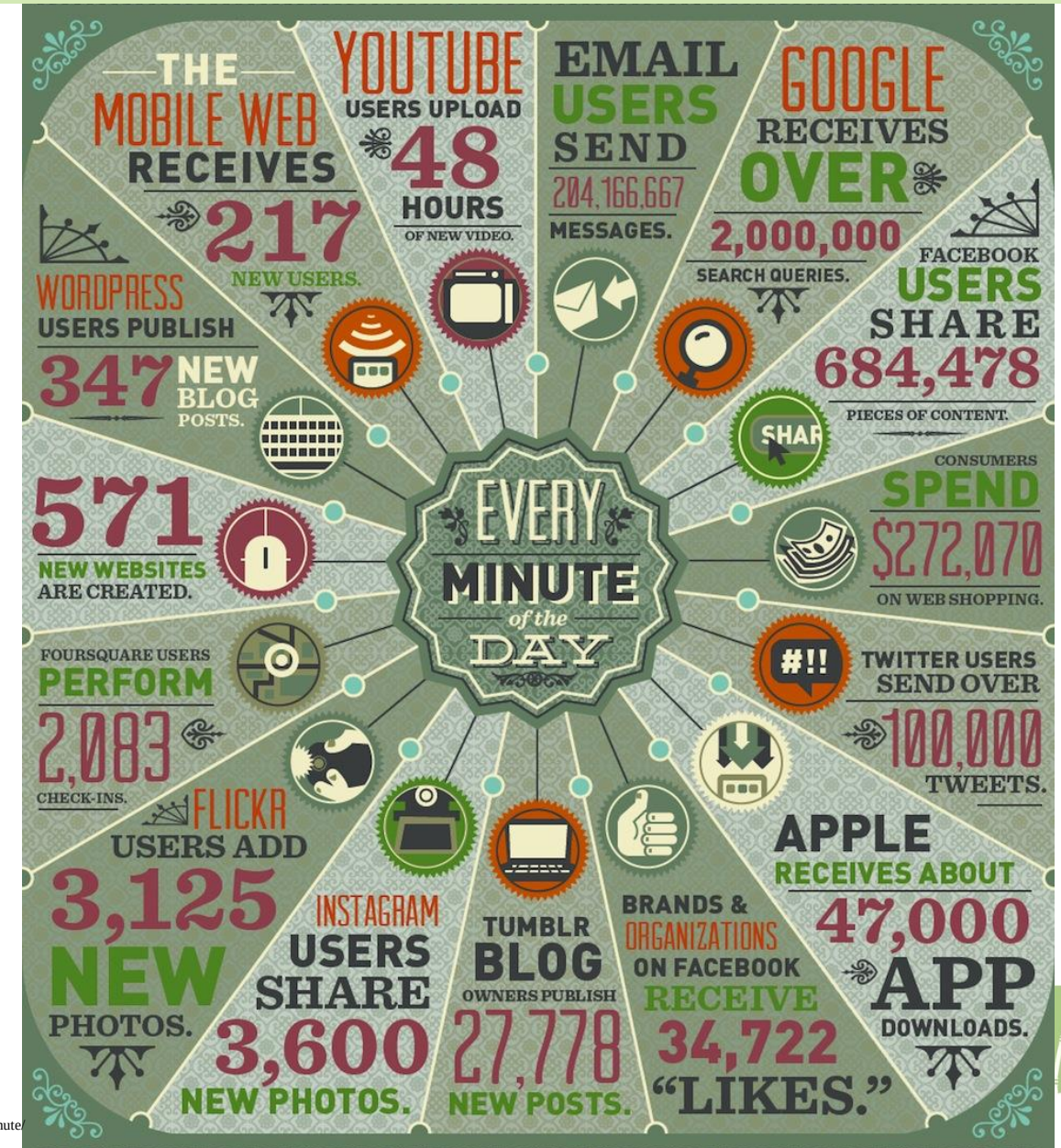
BASIS
TECHNOLOGY

Big Data

- Volume
- Velocity
- Variety

BASIS TECHNOLOGY

Volume



http://mashable.com/2012/06/22/data-created-every-minute/

BASIS TECHNOLOGY

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

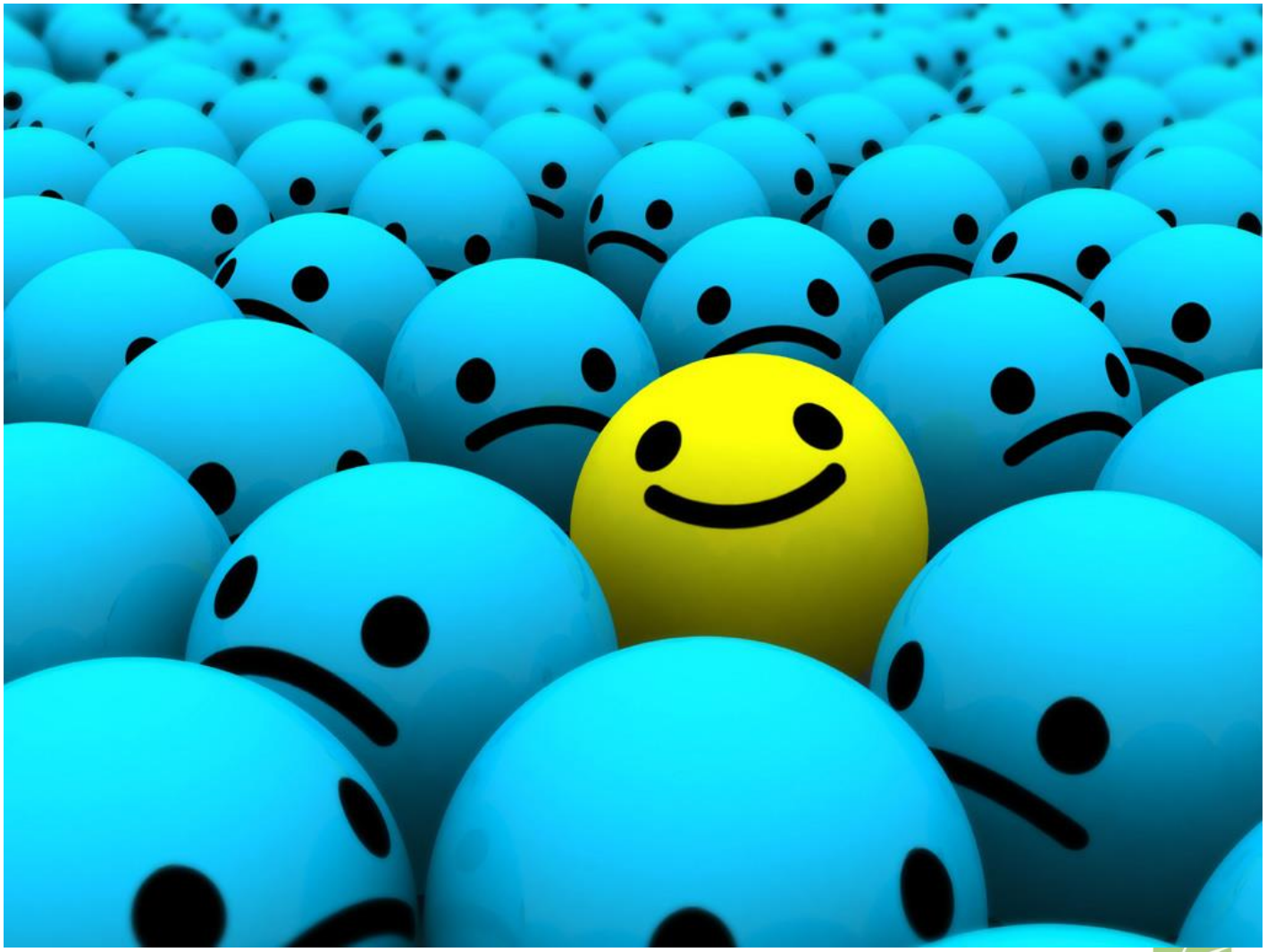
BASIS TECHNOLOGY

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

BASIS TECHNOLOGY

The Challenge: Finding Value



BASIS TECHNOLOGY

http://learn-how-to-be-happy.com/wp-content/uploads/2011/08/happy_face.jpg

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?*

BASIS TECHNOLOGY

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?*

BASIS TECHNOLOGY

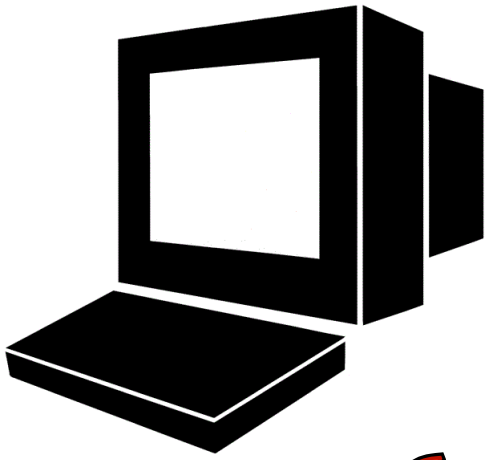
Big Data Problems - Variety

- 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*

BASIS TECHNOLOGY

General Problem

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



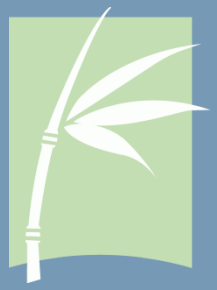
2 + 2 ✓

Scale ✓

Human Language

BASIS TECHNOLOGY

Text Analytics



BASIS
TECHNOLOGY

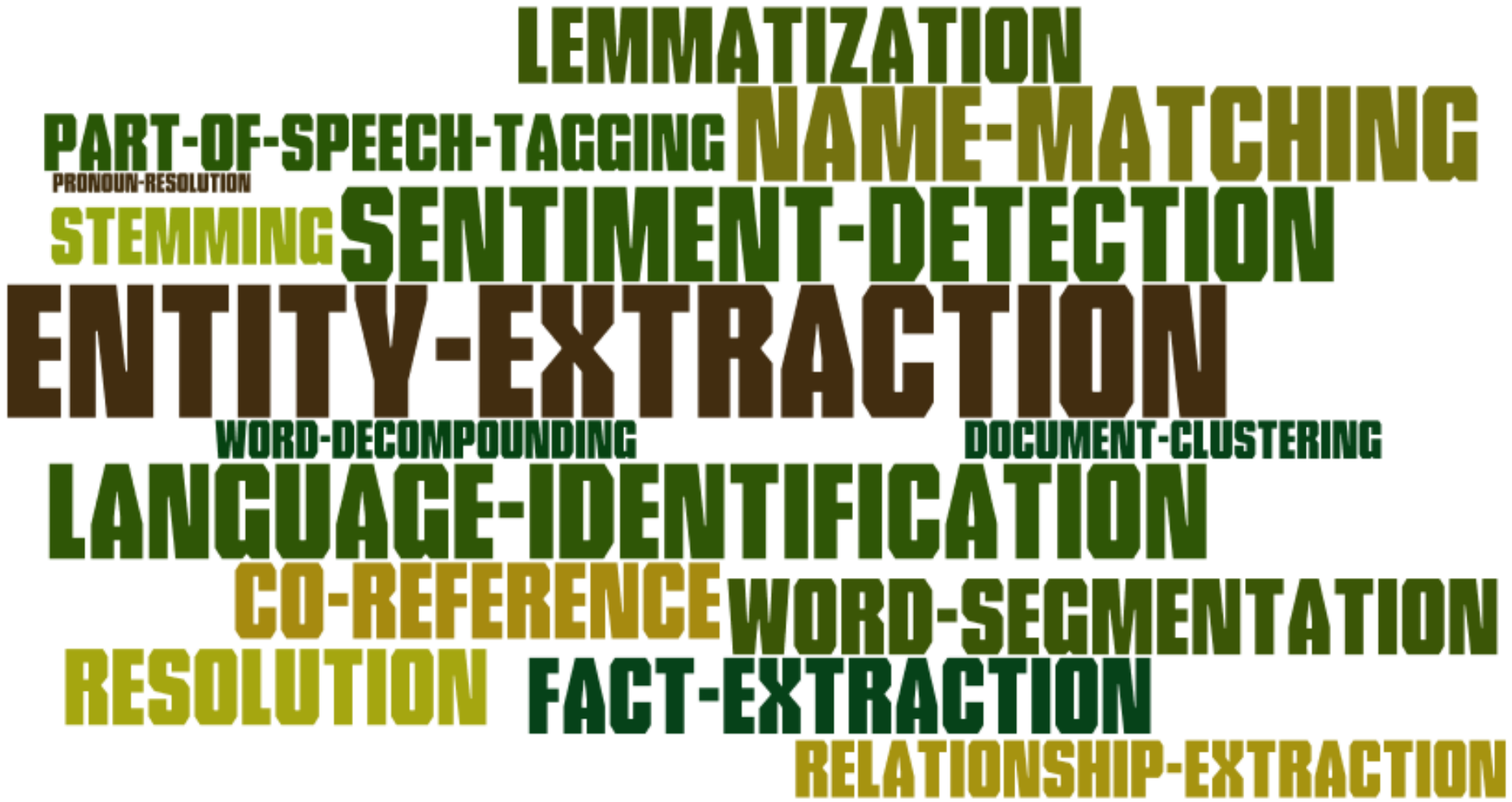
Text Analytics

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.

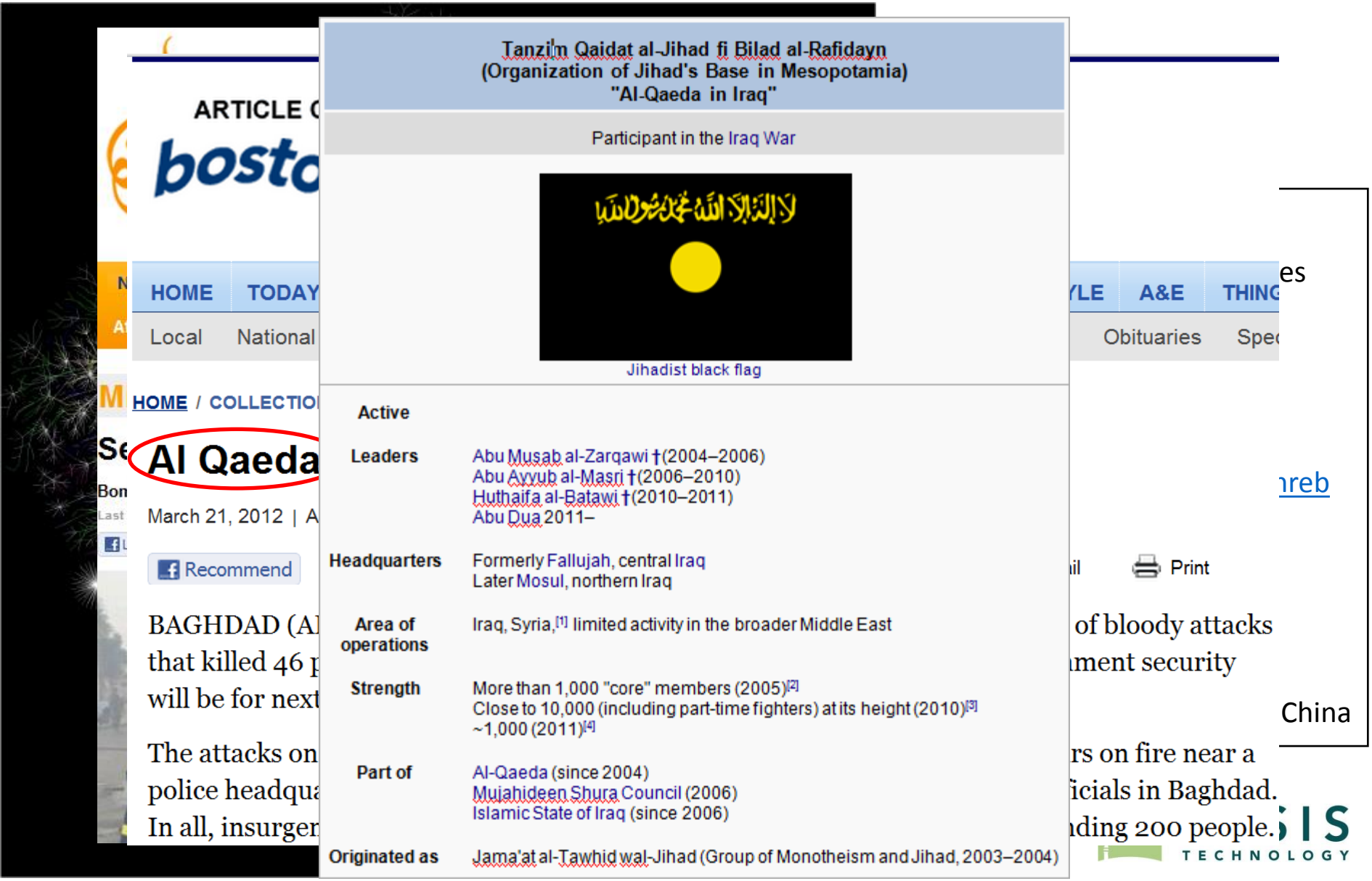
BASIS TECHNOLOGY

Text Analytics techniques

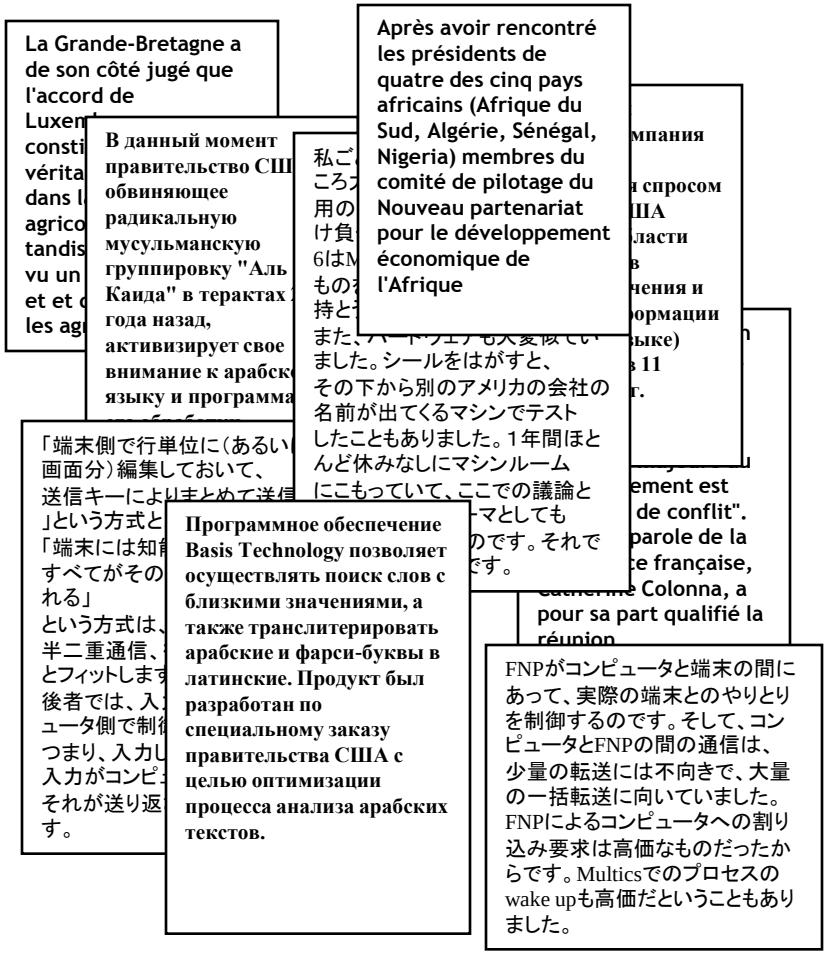


BASIS TECHNOLOGY

Triage Example



Text Analytics : Language ID



French

Russian

Japanese

BASIS TECHNOLOGY

Text Analytics: Lemmatization

flying

Search

Results

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

BASIS TECHNOLOGY

Text Analytics: Lemmatization (Arabic)

فجر (Detonated)

Search

Results

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

BASIS TECHNOLOGY

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.

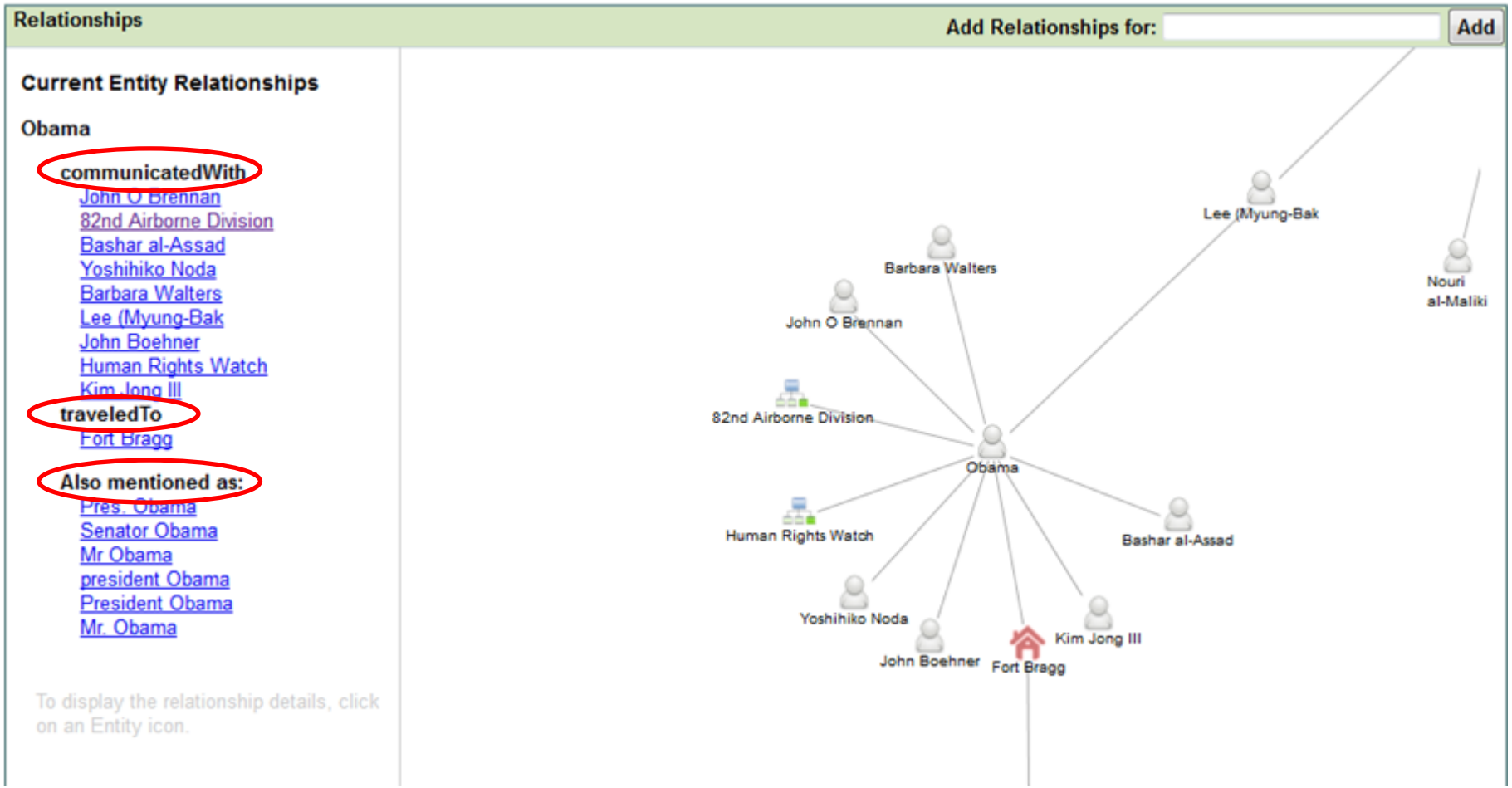
Baghdad A wave of bombing and shooting attacks in six different provinces across **Iraq** killed at least 34 people and wounded more than 100 Thursday, security officials said.

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



BASIS TECHNOLOGY

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

BASIS TECHNOLOGY

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

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 »

Related
Egypt »
Hosni Mubarak »
Cairo »


France 24


BBC News

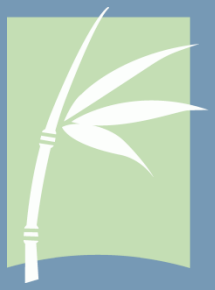

Boston.com


euronews


Telegra

BASIS TECHNOLOGY

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

BASIS TECHNOLOGY

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.

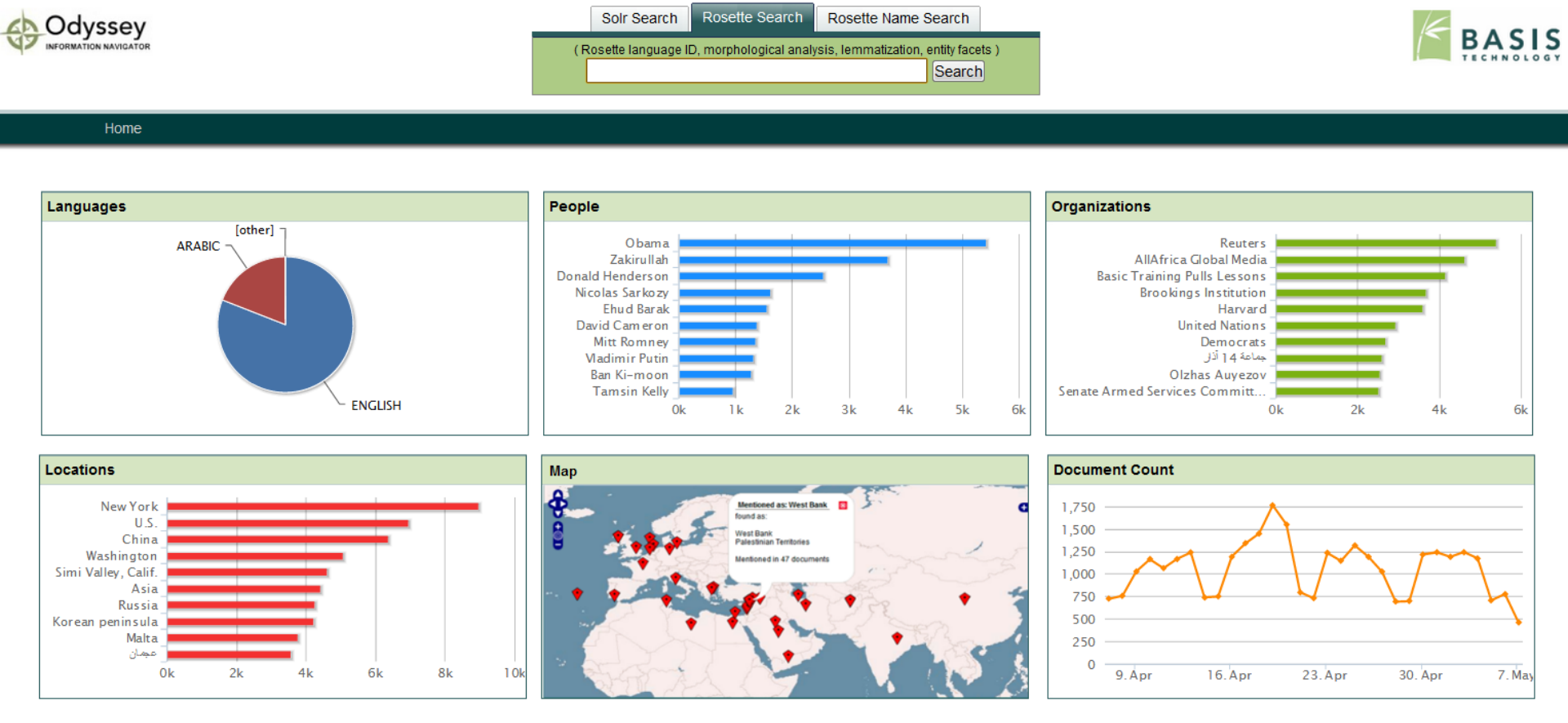
BASIS TECHNOLOGY

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*

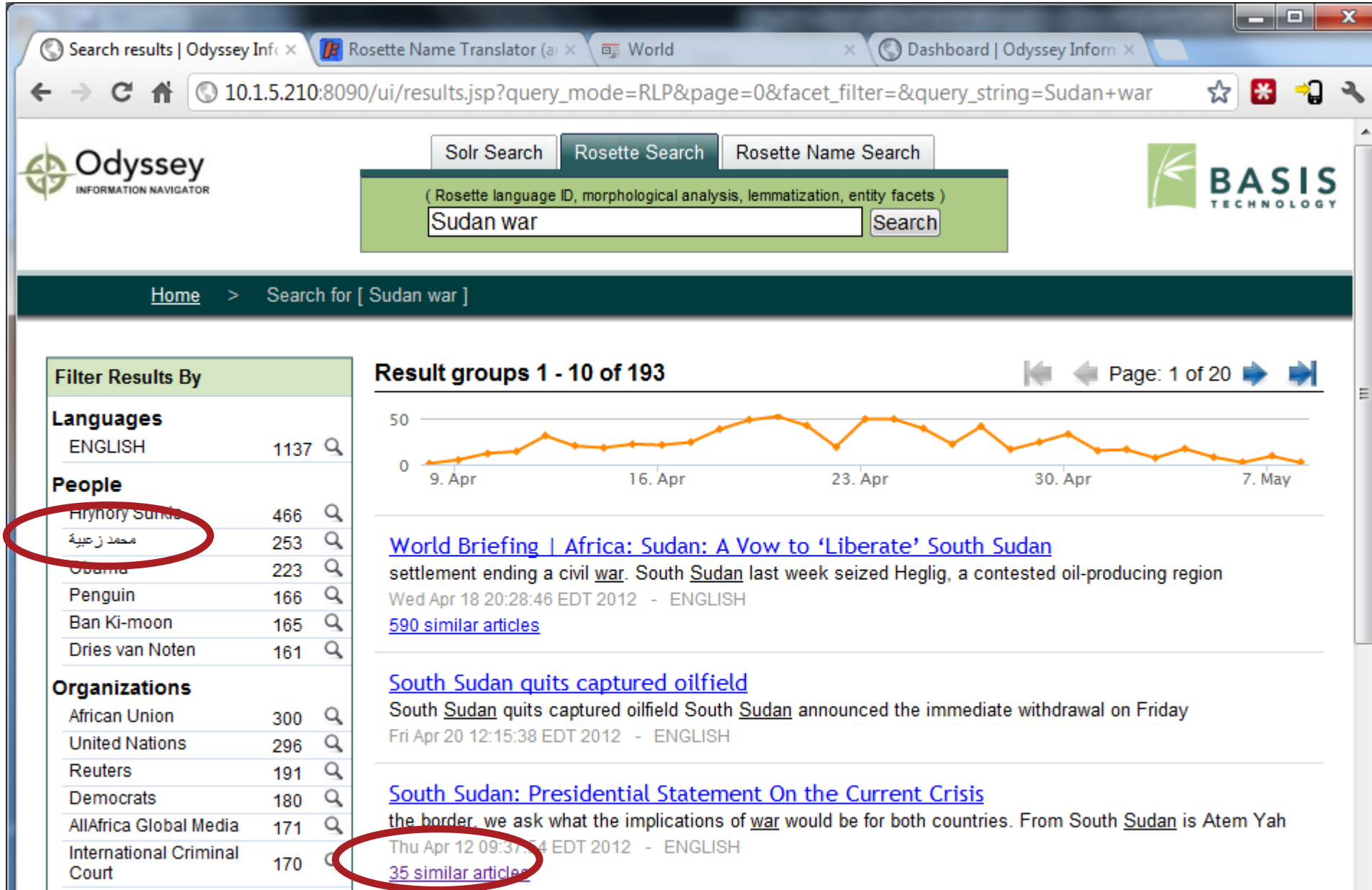
BASIS TECHNOLOGY

Dashboard



BASIS TECHNOLOGY

Search and Filter



Foreign Language Search

Detailed Document View

Odyssey
INFORMATION NAVIGATOR

BASIS TECHNOLOGY

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

☐ENGLISH **(1)**

Base Linguistics

Entity Extraction

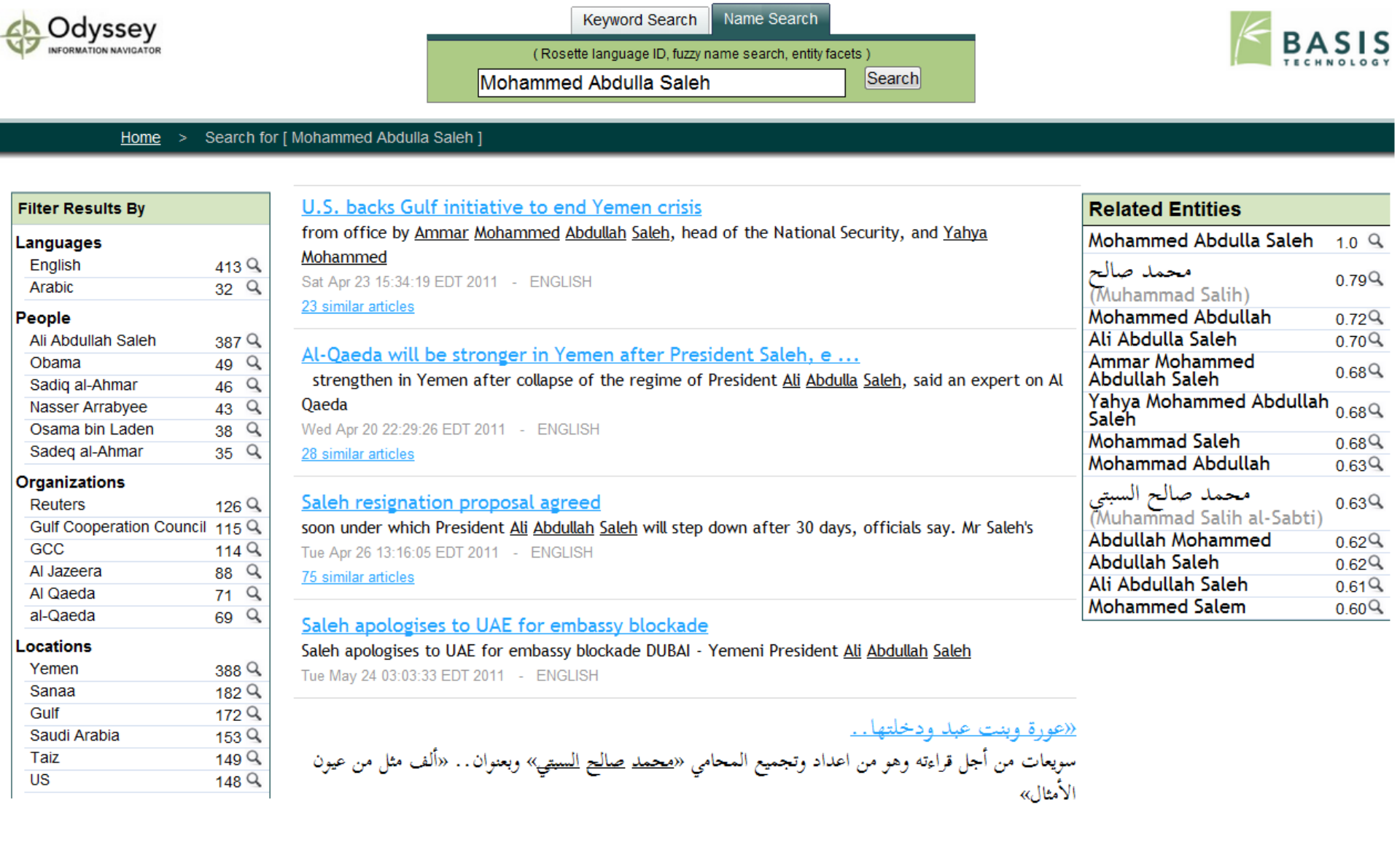
- ☑ IDENTIFIER-LATITUDE_LONGITUDE (1)
- ☑ IDENTIFIER-NUMBER (10)
- ☑ IDENTIFIER-URL (1)
- ☑ LOCATION (13)
- ☑ NATIONALITY (5)
- ☑ ORGANIZATION (11)
- ☑ PERSON (21)
- ☑ RELIGION (4)
- ☑ TEMPORAL-DATE (4)
- ☑ TEMPORAL-TIME (1)
- ☑ TITLE (8)

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

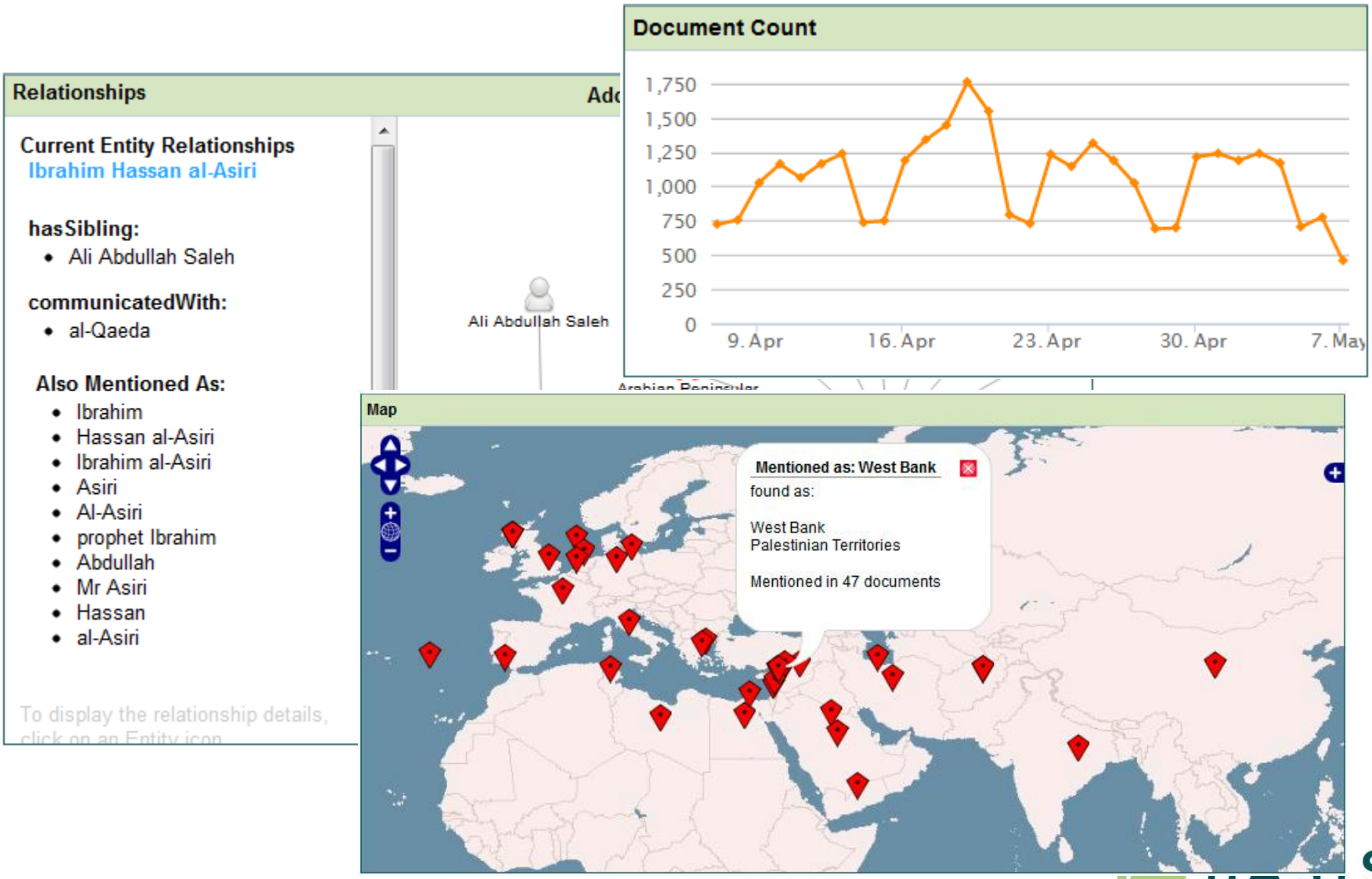
Simulcast | **Obama**'s **Cairo** Speech
Forceful Words and Fateful Realities
Published: June 6, 2009
BAGHDAD — Barack **Obama**'s speech in **Cairo** last Thursday was "soft spoken and eloquent," said **Moqtada al-Sadr**, the radical Iraqi **cleric**, grudgingly, since he also said he despised it.
It 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.

BASIS TECHNOLOGY

Entity Search – Cross Language



Search/Filter/Explore



BASIS TECHNOLOGY

http://www.silobreaker.com/FlashNetwork.aspx?DrillDownItems=11_237360

Summary

Text Analytics enables Big Data Triage

BASIS TECHNOLOGY

Thank you!

- For more information:
- Visit www.basistech.com