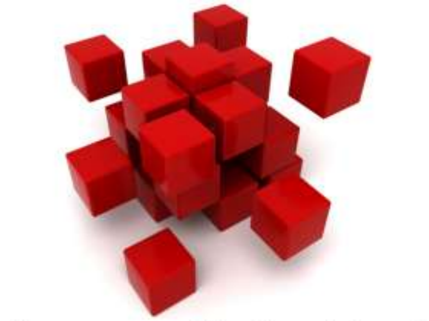


Entity Extraction Visualization (E2.0)

Entity extraction is the foundation for successful search, analysis, and visualization solutions! Rocket AeroText is that foundation. It is Entity Extraction 2.0

Technical Whitepaper

Estimates of 80% or higher of the "Collective Knowledge Pool" within an organization is contained within Unstructured Information (text-based documents). Emails, Analysis Reports, Field Observation Notes, Interviews; for the most part unstructured text information. Making sense of that information requires more than just the ability to search for a word within it. To leverage this information the application of "meaning" is required. The Rocket AeroText product suite provides a fast, agile information extraction system for developing knowledge-based content analysis applications. The technology excels at developing a core understanding of content contained within unstructured text, such as emails and documents, as well as an ability to reconcile automatically information cited across multiple documents. Such a capability makes it suited for a variety of applications, from counter-terrorism and law enforcement to business intelligence and enterprise content management.

Arnold Villeneuve

6/24/2011

6/24/2011

ENTITY EXTRACTION VISUALIZATION (E2.0)

Entity extraction is the foundation for successful search, analysis, and visualization solutions! Rocket AeroText is that foundation. It is Entity Extraction 2.0

Table of Contents

DEVELOPING VISUAL INSIGHT FROM TELEPHONE NUMBERS

CONCEPT OBJECTIVES:

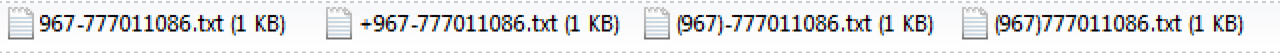
As a test of our solution capabilities we were provided with one million telephone numbers. The objective of the effort was to use the one million telephone numbers as input to search the entire Internet to locate web sites that contained the any of the input telephone numbers. Once a positive website hit was located our task was to download all the publically available files from the web site for post semantic processing. The web site page files are all unstructured text files. Once the website was downloaded we post processed it with AeroText Entity, Identity, and Relationship extraction software. The results of this process were stored in a structured SQL relational database. Now that the web site information was available in structured metadata format we integrated it with Visual Analytics Clarity server for data discovery and insight visualization.

The following is a description of the process and the results of our effort.

Phase 1: Input File of Telephone Numbers Re-Formatting

We were provided with a list of Telephone numbers (1 million of them) and asked to create a solution that would use the input file of telephone numbers to locate web sites that contained evidence of these telephone numbers.

The VAI Search Runner is demonstrated below. It takes the original telephone number in this case and creates several different formats of it.



967-777011086.txt (1 KB)   +967-777011086.txt (1 KB)   (967)-777011086.txt (1 KB)   (967)777011086.txt (1 KB)

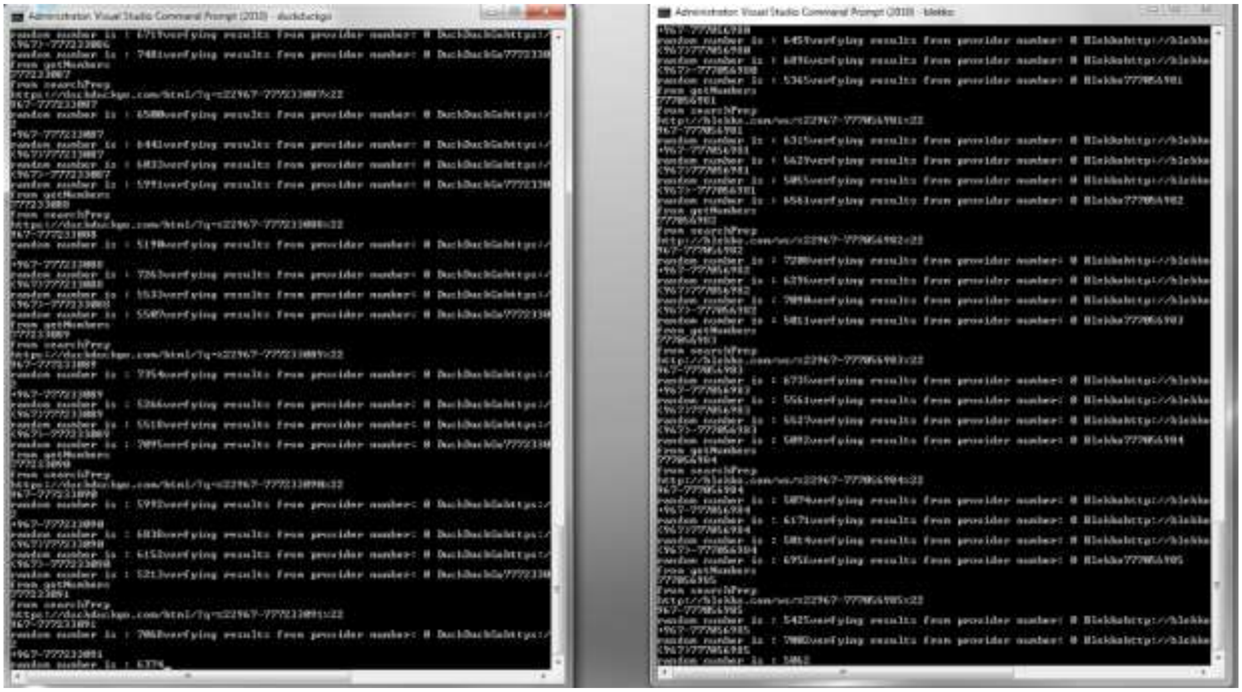
Depending on the person or organization the telephone number displayed on the website may use various formats and we wanted to ensure that we captured them all. When re-formatting is done the total number of numbers to search becomes 4 million!

Phase 2: Query the Search Engines

This is rather difficult to do when the input for the search query is 1 million telephone numbers because the popular search engines limit you to the number of search queries you can perform using their direct API. The other thing we discovered in this process is that the main web search page searches the entire search provider index while the API is very limited. That is to say that one could search from the main search engine landing page and get a hit result whereas using the API no hit result comes back for the same search query. Finally, major search engines limit the number of queries you can do per minute via the API and block you if you exceed it. That's a moot point however given that the API approach does not yield good results in the first place. Thus special software had to be developed to access the entire search engine index while not cut of while searching and then processing the results from a web page.

We ran through 17 different search engines in the process of creating the solution.

Here is an example of two Search Engines we search against from the telephone number input file we were provided with. They have provided the most promising results and have allowed us to have very good access without cutting us off.



Phase 3: Search Hit Result URLs

Here is a sample of the results we received.

Web queried for: 967-777040500
Maximum search results provided is 10, showing results for the following:
Title: Regional Meeting 21 to 23-04-09</a </h3
Subject: a b c d e f g h i j k l m n o p; 1: mena regional meeting - 21-23 april - amman - jordan list of participants : 2: mena regional meeting - 21-23 april 2009 - amman - jordan
Address:
http://www.theglobalfund.org/documents/regionalmeetings/jordan2009/List_of_Participants.xls

In the following example we queried one number and got back two positive web site URL hit results:

Web queried for: 967-777011086
Maximum search results provided is 10, showing results for the following:

Title: Center Manager, Business Heads, Administrators &amp; Sales Consultants ...</a </h3
Subject: Quick Tip: Make Sure the Seller is in Qatar by checking for a local phone number and don&#39;t ever buy anything listed in $. Read Avoiding Scams and Fraud for more advice.
Address: **http://www.qatarliving.com**/node/30714

Title: Ø±Ø¬Ù„ Ø§Ù„Ø£Ø¹Ù…Ø§Ù„ Ù…Ø¹Ù† Ø§Ù„ØµØ§Ù†Ø¹ ÙŠØªØ¨Ø±Ø¹ Ø¨Ø¹Ù„Ø§Ø¬ 60 ...</a </h3
Subject: ØªØ¨Ø±Ø¹ Ø±Ø¬Ù„ Ø§Ù„Ø§Ø¹Ù…Ø§Ù„ Ø§Ù„Ù…Ø¹Ø±ÙˆÙ⍰ Ù…Ø¹Ù† Ø§Ù„ØµØ§Ù†Ø¹ Ù„Ø¹Ù„Ø§Ø¬ 60 ØØ§Ù„Ù‡ Ù…Ù† Ù…Ø±Ø¶Ù‰ Ø§Ù„Ù⍰Ø´Ù„ Ø§Ù„ÙƒÙ„Ù‰ÙŠâ€¦ ØÙŠØ« ØªÙ‚Ø¯Ù… Ø§Ù„ÙŠ Ø§Ù„Ø§Ù† 4 ...
Address: **http://abusaad.ws**/archives/1368

Search Results Hit Table

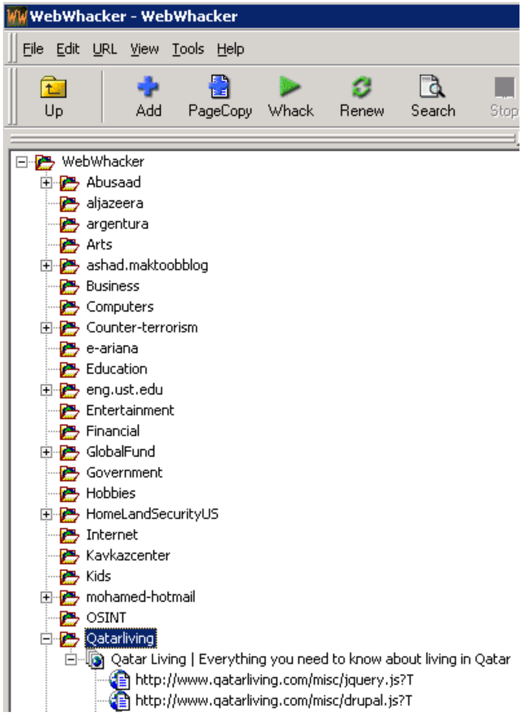
The following is a small sample of actual telephone numbers we were given to search out across the Internet to look for web sites that contained the numbers provided.

Telephone Number	Web Site	Full Web Site URL
967-771201430	http://rashad.maktoobblog.com/	http://rashad.maktoobblog.com/about-g-a-g-h/
967-777005620	http://www.der.gov.ye/	http://www.der.gov.ye/
967-777008060	http://gisbasalama.blogspot.com/	http://gisbasalama.blogspot.com/feeds/posts/default?orderby=updated

967-777009011	http://embassy-finder.com/	http://embassy-finder.com/singapore_in_abu-dhabi_united-arab-emirates?page=2
967-777009416	http://www.iana.org/	http://www.iana.org/domains/root/db/ye.html
967-777009486	http://www.iana.org/	http://www.iana.org/root-whois/ye.htm http://www.101domain.com/whois-ye.php http://whois.smartweb.cz/en/domain/ye/
967-777011086	http://www.qatarliving.com/	http://www.qatarliving.com/node/30714 http://abusaad.ws/archives/1368
967-777011929	http://en.yoxls.net/	http://en.yoxls.net/mohamed-hotmail-com-ceo-yemen-xls.html
967-777040098	http://www.eng.ust.edu.ye/	http://www.eng.ust.edu.ye/Resume/mnasser.pdf
967-777040500	http://www.theglobalfund.org/	http://www.theglobalfund.org/documents/regionalmeetings/jordan2009/List_of_Participants.xls
967-777227855	http://www.trading-mall.com/	http://www.trading-mall.com/aar-98.html
967-777233303	http://en.shac.gov.cn/ http://www.global-buyers.info/	http://en.shac.gov.cn/bs/buy/200905/t20090521_1244547.htm http://www.global-buyers.info/buying-leads/buyers-detail.asp?buyerid=9984

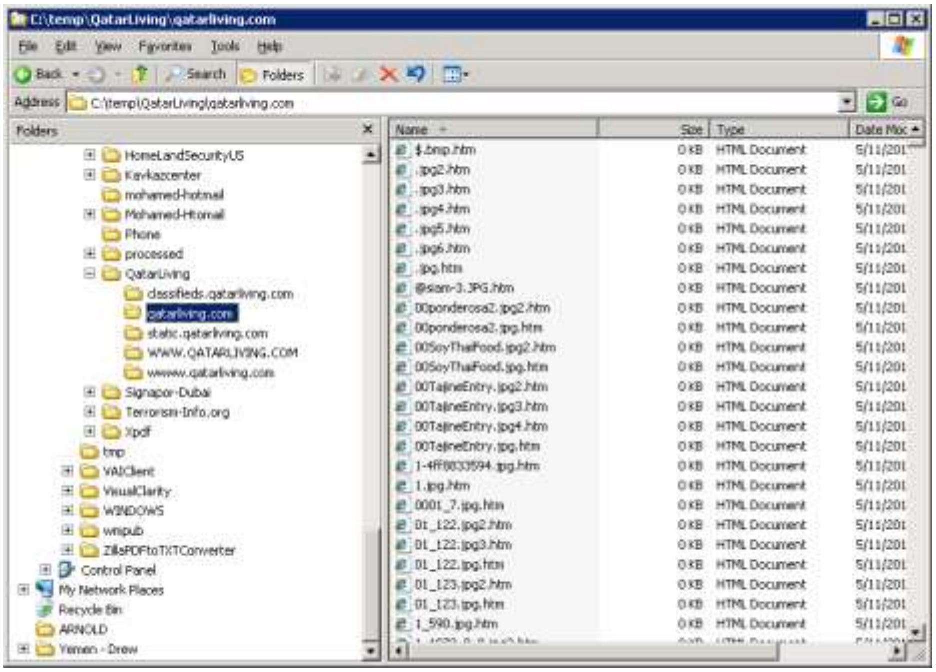
Phase 4: Download the Web Site

We first reviewed the returned websites manually in order to ensure that there was some perceived intelligence value in the returned URL. We used a web whacking software tool to download all publically available files from the discovered web sites.



```
WebWhacker
    Abusaad
    aljazeera
    argentura
    Arts
    ashad.maktoobblog
    Business
    Computers
    Counter-terrorism
    e-ariana
    Education
    eng.ust.edu
    Entertainment
    Financial
    GlobalFund
    Government
    Hobbies
    HomeLandSecurityUS
    Internet
    Kavkazcenter
    Kids
    mohamed-hotmail
    OSINT
    Qatarliving
        Qatar Living | Everything you need to know about living in Qatar
            http://www.qatarliving.com/misc/jquery.js?T
            http://www.qatarliving.com/misc/drupal.js?T
```

We then copied the downloaded web site URL files to a folder for post processing by the AeroText entity extraction software.



C:\temp\QatarLiving\qatarliving.com

Address: C:\temp\QatarLiving\qatarliving.com

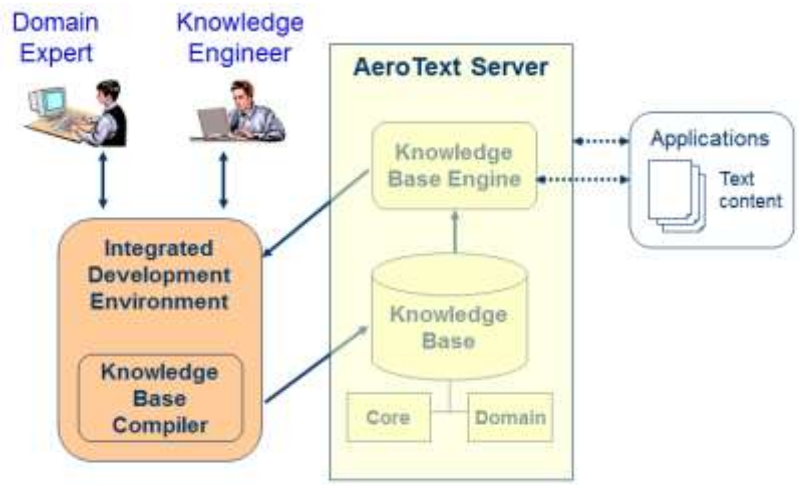
```
        HomeLandSecurityUS
        Kavkazcenter
        mohamed-hotmail
        Mohamed-Htomail
        Phone
        processed
        QatarLiving
            classifieds.qatarliving.com
            qatarliving.com
            static.qatarliving.com
            WWW.QATARLIVING.COM
            wwww.qatarliving.com
        Signapor-Dubai
        Terrorism-Info.org
        Xpdf
      tmp
      VAIClient
      VisualClarity
      WINDOWS
      wnipub
      ZillaPDFtoTXTConverter
    Control Panel
  My Network Places
  Recycle Bin
  ARNOLD
  Yemen - Drew
```

Name	Size	Type	Date Mo
$.bmp.htm	0 KB	HTML Document	5/11/201
.jpg2.htm	0 KB	HTML Document	5/11/201
.jpg3.htm	0 KB	HTML Document	5/11/201
.jpg4.htm	0 KB	HTML Document	5/11/201
.jpg5.htm	0 KB	HTML Document	5/11/201
.jpg6.htm	0 KB	HTML Document	5/11/201
.jpg.htm	0 KB	HTML Document	5/11/201
@siam-3.JPG.htm	0 KB	HTML Document	5/11/201
00ponderosa2.jpg2.htm	0 KB	HTML Document	5/11/201
00ponderosa2.jpg.htm	0 KB	HTML Document	5/11/201
00SoyThaiFood.jpg2.htm	0 KB	HTML Document	5/11/201
00SoyThaiFood.jpg.htm	0 KB	HTML Document	5/11/201
00TajineEntry.jpg2.htm	0 KB	HTML Document	5/11/201
00TajineEntry.jpg3.htm	0 KB	HTML Document	5/11/201
00TajineEntry.jpg4.htm	0 KB	HTML Document	5/11/201
00TajineEntry.jpg.htm	0 KB	HTML Document	5/11/201
1-4ff8833594.jpg.htm	0 KB	HTML Document	5/11/201
1.jpg.htm	0 KB	HTML Document	5/11/201
0001_7.jpg.htm	0 KB	HTML Document	5/11/201
01_122.jpg2.htm	0 KB	HTML Document	5/11/201
01_122.jpg3.htm	0 KB	HTML Document	5/11/201
01_122.jpg.htm	0 KB	HTML Document	5/11/201
01_123.jpg2.htm	0 KB	HTML Document	5/11/201
01_123.jpg.htm	0 KB	HTML Document	5/11/201
1_590.jpg.htm	0 KB	HTML Document	5/11/201

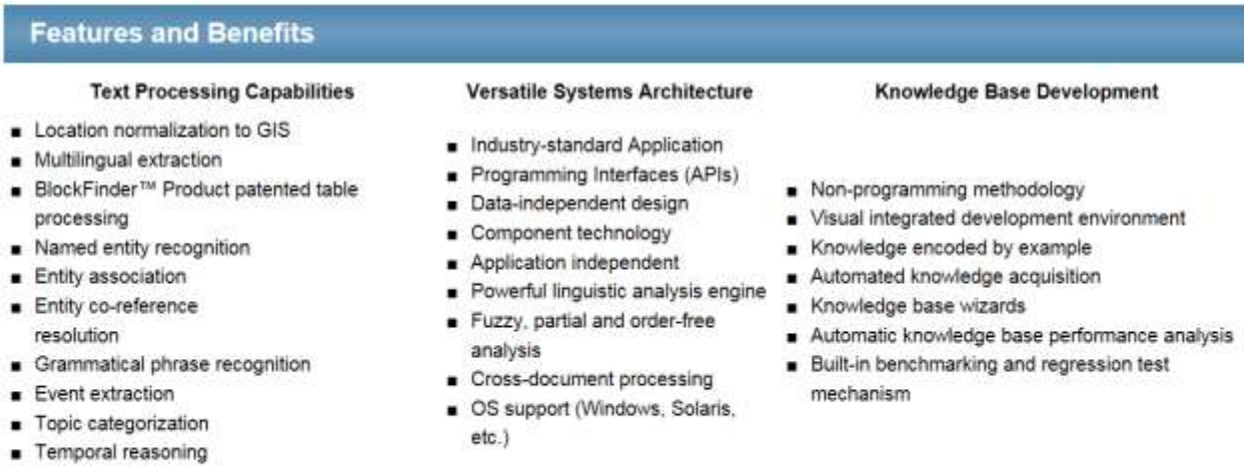
Phase 5: AeroText Based Semantic Post Processing of Unstructured Text

We selected AeroText from Rocket Software as the "Best Of Breed" entity extraction software tool. In addition to standard entity extraction the unique feature that AeroText has is that it will also create Identity and MergedRelation metadata of unstructured text. In our view it is the best software for entity extraction from unstructured text files that facilitates the creation of structured metadata content for further insight analysis. AeroText has a variety of tools to create knowledge specific domain analysis of unstructured text content. AeroText also supports entity extraction of different languages.

Building a Domain Specific Knowledge Base to Process Unstructured Text



AeroText provides the following features:



Features and Benefits

Text Processing Capabilities

- Location normalization to GIS
- Multilingual extraction
- BlockFinder™ Product patented table processing
- Named entity recognition
- Entity association
- Entity co-reference resolution
- Grammatical phrase recognition
- Event extraction
- Topic categorization
- Temporal reasoning

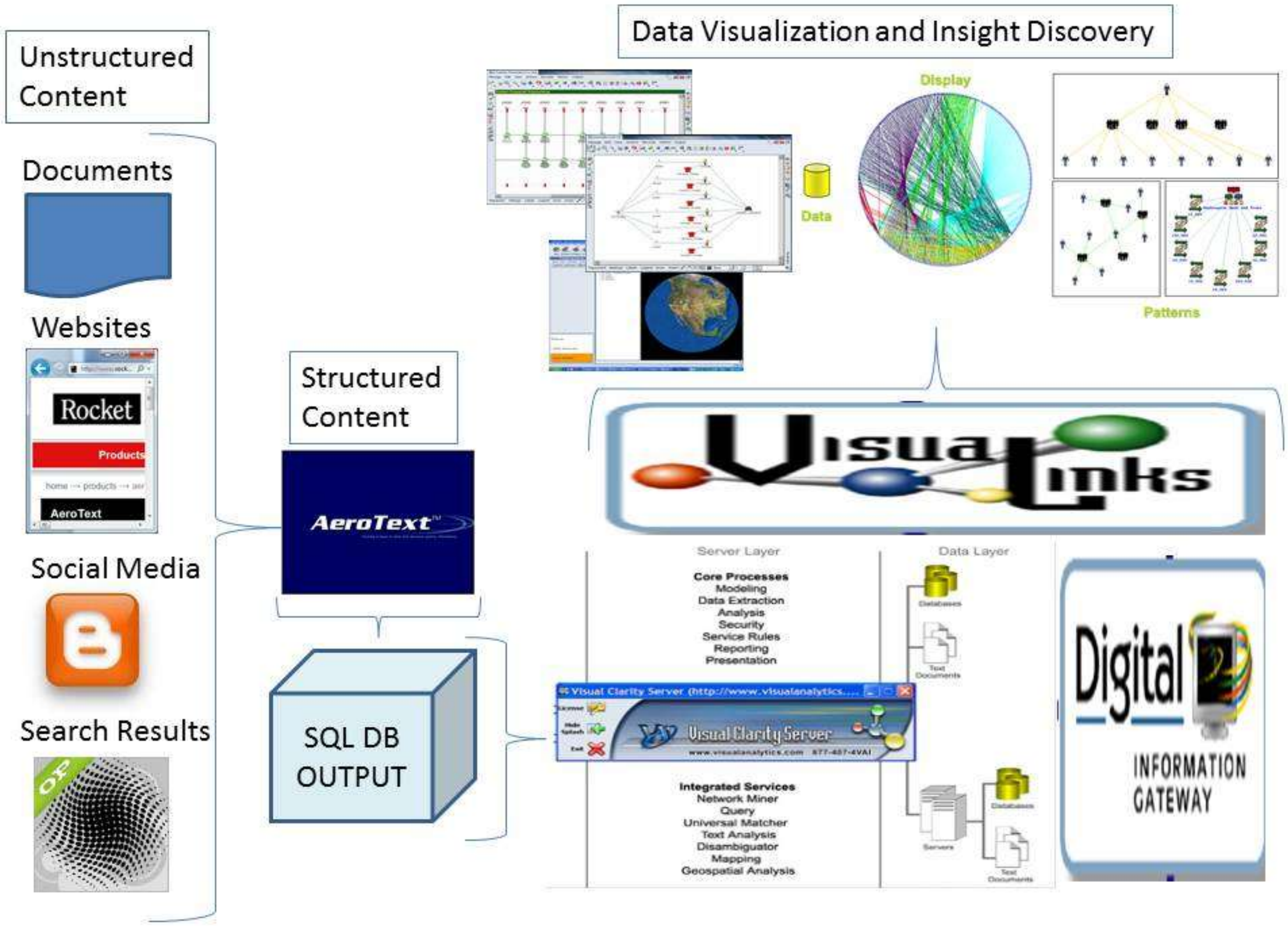
Versatile Systems Architecture

- Industry-standard Application
- Programming Interfaces (APIs)
- Data-independent design
- Component technology
- Application independent
- Powerful linguistic analysis engine
- Fuzzy, partial and order-free analysis
- Cross-document processing
- OS support (Windows, Solaris, etc.)

Knowledge Base Development

- Non-programming methodology
- Visual integrated development environment
- Knowledge encoded by example
- Automated knowledge acquisition
- Knowledge base wizards
- Automatic knowledge base performance analysis
- Built-in benchmarking and regression test mechanism

The following diagram shows the process from entity extraction to visualization.



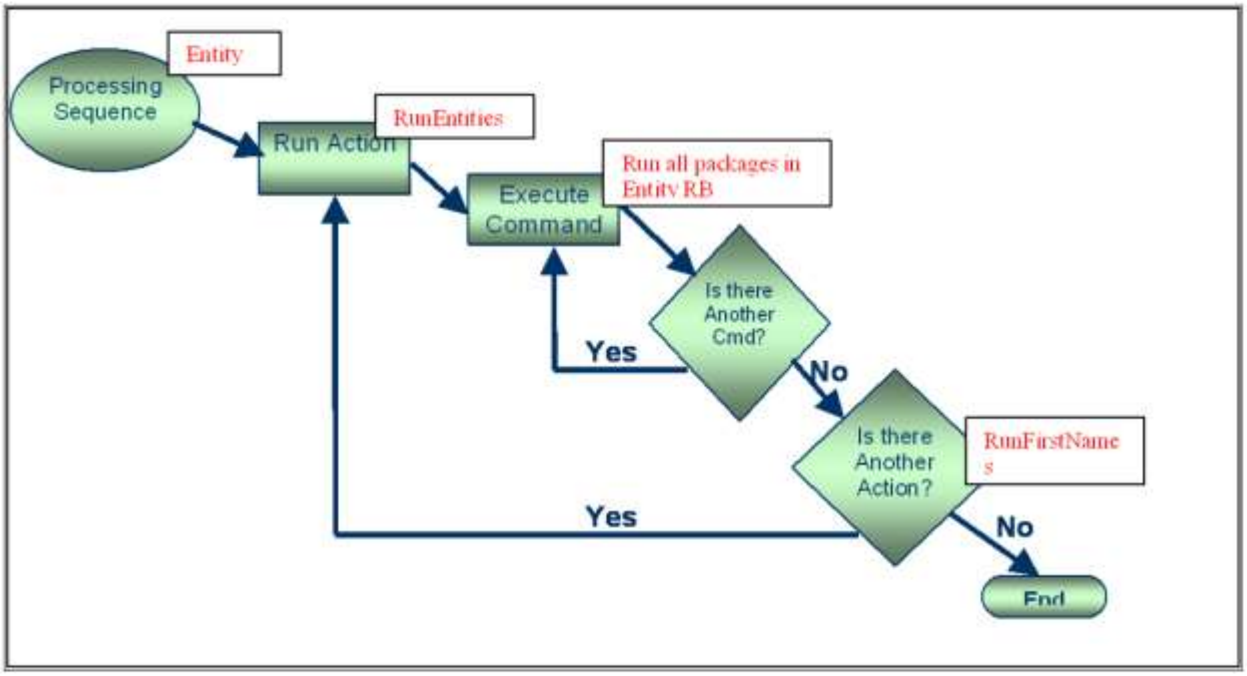
AeroText Entity Extraction Process Sequence

Data Flow for Processing of a Document While the specifics might vary depending upon input and application, the data flow for processing a document, whether singly or one of a batch, will likely consist of the following steps:

1. A document is loaded and broken into paragraphs and sentences by the load sequence (defined in the project properties). See the end of this chapter for more information on load sequences (Segmentation.)

2. The processing sequence of choice is then run on the document. At this level of processing, actions run rule bases. Each action in the processing sequence defines which cache or parts of a cache will be given to the pattern matcher and which rule base will be run against that cache. A cache is a container that can store one or more records. If any caches are to be attached (i.e., pattern matching run against the contents of the cache), they are specified in the action at this level. Information obtained during rule matching is stored in a cache by the rule actions (see below).

3. The activated rule bases consist of one or more packages. Packages are a collection of rules, which are in turn comprised of constituents designed to match regions of text. When all the required constituents successfully match, the rule then triggers an action. This action (at a lower level of complexity than the ones which comprise the processing sequence) then uses the matched region to create a new record in a results cache.



The above process can be accomplished within the AeroText Visual AT Explorer interface or the automated runtime RIT application. The results of the processing can be output to CSV, XML, and SQL database formats for further post processing and application integration.

The Run-time Integration Toolkit (RIT) provides a way for AeroText users to apply knowledge bases to document sets. Knowledge bases are developed using the AeroText IDE and then exported for use in RIT. RIT enables the rapid development of custom input sources and output targets, isolating these elements from the processing tasks.

Once an AeroText Knowledge Base is built within the Visual AT Editor (used by the Knowledge Engineer and Domain Expert) it is compiled and incorporated into the RIT application for automated processing of selected unstructured content.

Building an Application

Assembling an application using the RIT consists of five high-level steps:

1. Using the AeroText IDE, develop a KB to extract the information relevant to the domain.

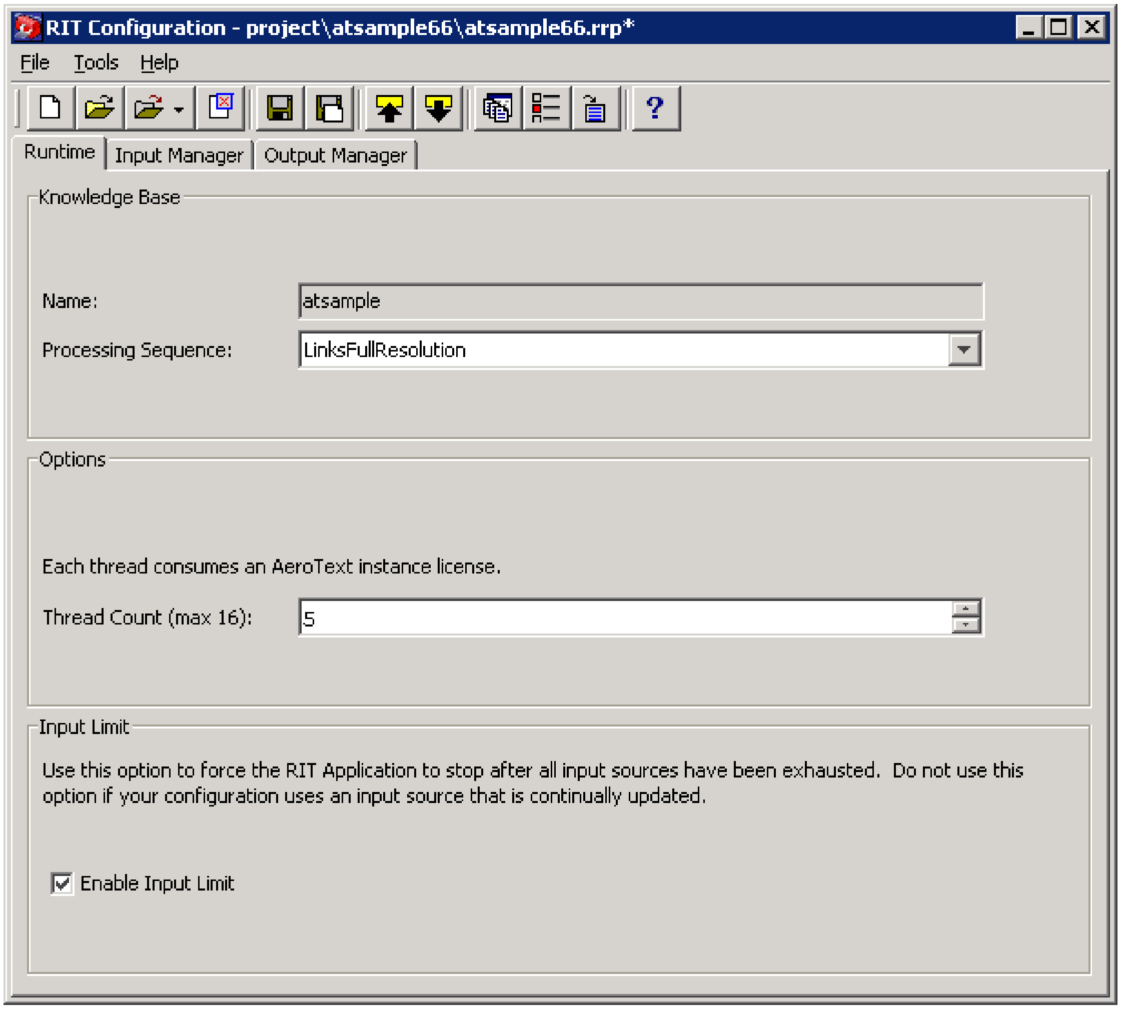
2. Import the KB from the development location using the RIT Configuration tool.

3. Select one or more RIT Input Sources. If no existing Input Source provides the needed functionality, develop a new one.

4. Select one or more RIT Output Targets. If no existing Output Target provides the needed functionality, develop a new one.

5. Using the RIT Configuration tool, select and configure a KB, configure one or more input sources, configure one or more output targets.

6. Run the RIT application using the RIT Project configured in step 4.



Knowledge Base

Name: atsample

Processing Sequence: LinksFullResolution

Options

Each thread consumes an AeroText instance license.

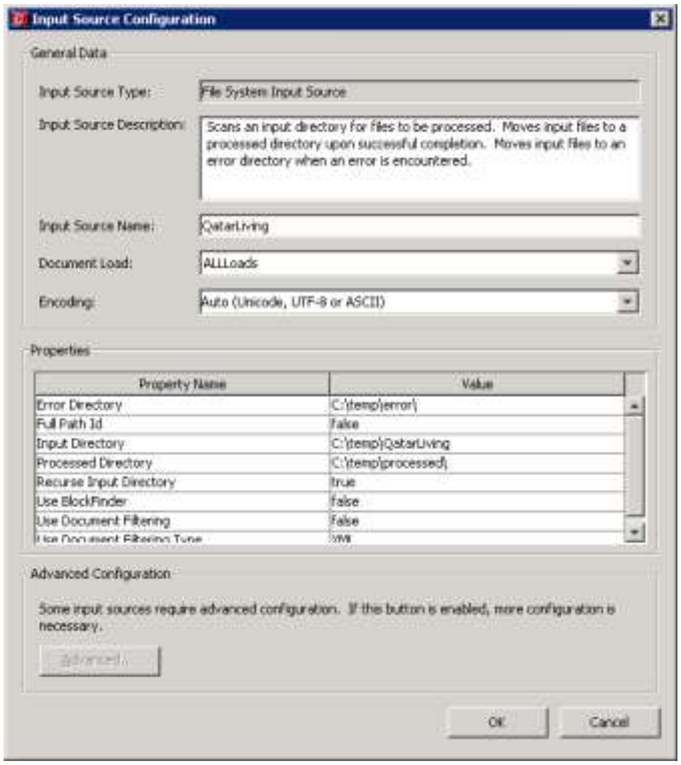
Thread Count (max 16): 5

Input Limit

Use this option to force the RIT Application to stop after all input sources have been exhausted. Do not use this option if your configuration uses an input source that is continually updated.

☑ Enable Input Limit

Input Source (in this case the discovered website that was downloaded for post semantic processing)



Input Source Configuration

General Data

Input Source Type: File System Input Source

Input Source Description: Scans an input directory for files to be processed. Moves input files to a processed directory upon successful completion. Moves input files to an error directory when an error is encountered.

Input Source Name: QatarLiving

Document Load: ALLLoads

Encoding: Auto (Unicode, UTF-8 or ASCII)

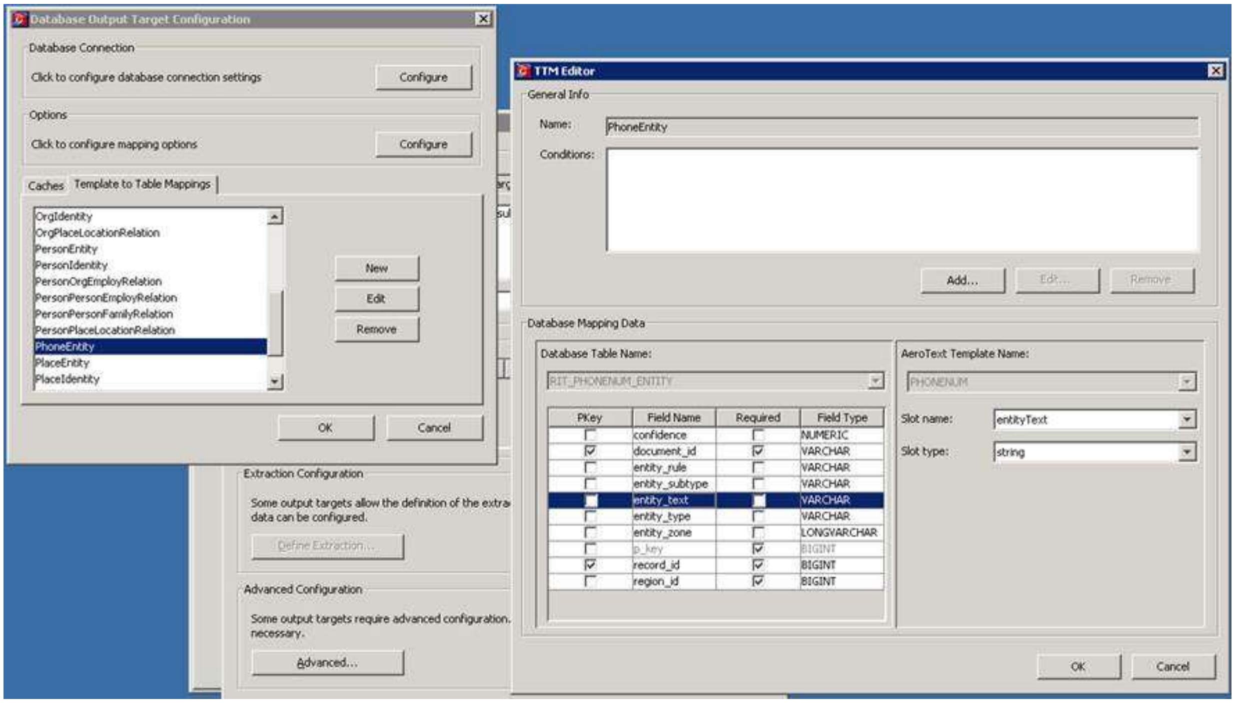
Properties

Property Name	Value
Error Directory	C:\temp\error\
Full Path Id	false
Input Directory	C:\temp\QatarLiving
Processed Directory	C:\temp\processed\
Recurse Input Directory	true
Use BlockFinder	false
Use Document Filtering	false
Use Document Filtering Type	XML

Advanced Configuration

Some input sources require advanced configuration. If this button is enabled, more configuration is necessary.

AeroText Entity Extraction Output to SQL Database



Database Output Target Configuration

Database Connection

Click to configure database connection settings

Options

Click to configure mapping options

- OrgIdentity
- OrgPlaceLocationRelation
- PersonEntity
- PersonIdentity
- PersonOrgEmployRelation
- PersonPersonEmployRelation
- PersonPersonFamilyRelation
- PersonPlaceLocationRelation
- PhoneEntity
- PlaceEntity
- PlaceIdentity

Extraction Configuration

Some output targets allow the definition of the extra data can be configured.

Advanced Configuration

Some output targets require advanced configuration. necessary.

TTM Editor

General Info

Name: PhoneEntity

Conditions:

Database Mapping Data

Database Table Name: RIT_PHONENUM_ENTITY

PKey	Field Name	Required	Field Type
□	confidence	□	NUMERIC
☑	document_id	☑	VARCHAR
□	entity_rule	□	VARCHAR
□	entity_subtype	□	VARCHAR
□	entity_text	□	VARCHAR
□	entity_type	□	VARCHAR
□	entity_zone	□	LONGVARCHAR
□	p_key	☑	BIGINT
☑	record_id	☑	BIGINT
□	region_id	☑	BIGINT

AeroText Template Name: PHONENUM

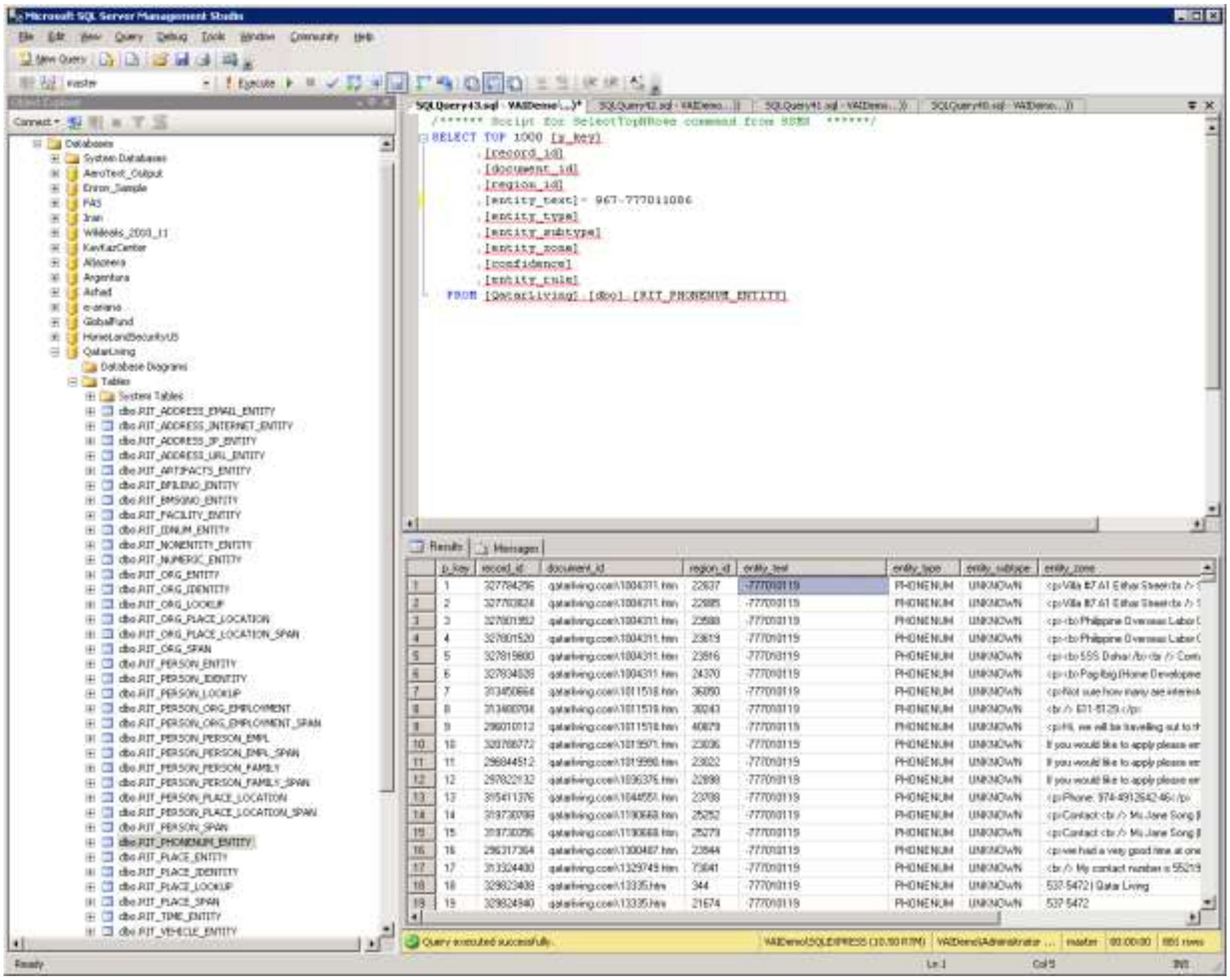
Slot name: entityText

Slot type: string

Phase 6: SQL Database of Entity Meta Data

The following screen shot reveals the level and quality of the entity extraction process. All of the entity meta data discovered within the web site documents, documents that are unstructured content, are now accessible via a structured format within a SQL database.

Web queried for: 967-777011086



```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT TOP 1000 [p_key]
      ,[record_id]
      ,[document_id]
      ,[region_id]
      ,[entity_text]= 967-777011086
      ,[entity_type]
      ,[entity_subtype]
      ,[entity_zone]
      ,[confidence]
      ,[entity_rule]
  FROM [QatarLiving].[dbo].[RIT_PHONENUM_ENTITY]
```

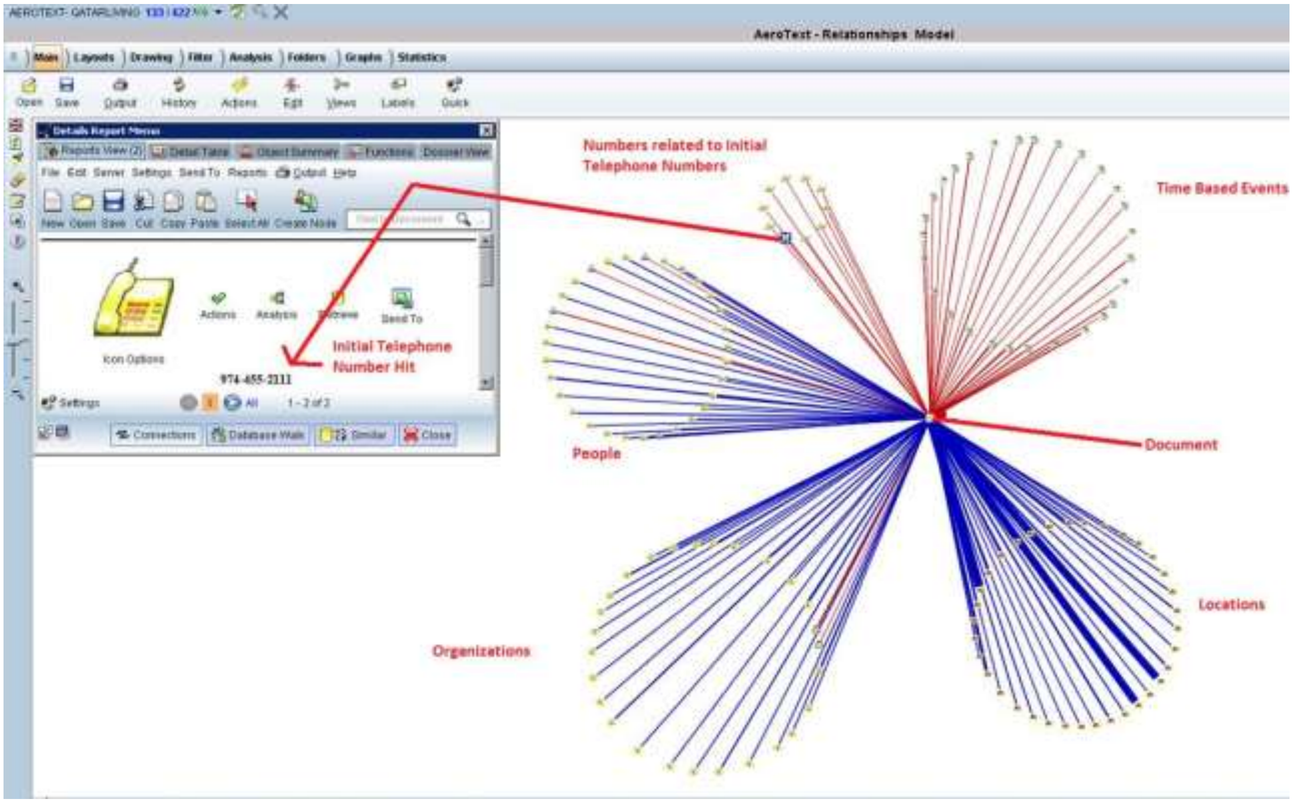
	p_key	record_id	document_id	region_id	entity_text	entity_type	entity_subtype	entity_zone
1	1	327784256	qatarliving.com\1004311.htm	22637	-777010119	PHONENUM	UNKNOWN	<p>Villa #7 61 Eithar Street<br /> (
2	2	327703824	qatarliving.com\1004311.htm	22885	-777010119	PHONENUM	UNKNOWN	<p>Villa #7 61 Eithar Street<br /> (
3	3	327801952	qatarliving.com\1004311.htm	23588	-777010119	PHONENUM	UNKNOWN	<p><b>Philippine Overseas Labor (
4	4	327801520	qatarliving.com\1004311.htm	23619	-777010119	PHONENUM	UNKNOWN	<p><b>Philippine Overseas Labor (
5	5	327819800	qatarliving.com\1004311.htm	23816	-777010119	PHONENUM	UNKNOWN	<p><b>SSS Doha</b><br /> Com
6	6	327834028	qatarliving.com\1004311.htm	24370	-777010118	PHONENUM	UNKNOWN	<p><b>Pag-Ibig (Home Developme
7	7	313450664	qatarliving.com\1011518.htm	36050	-777010119	PHONENUM	UNKNOWN	<p>Not sure how many are interest
8	8	313400704	qatarliving.com\1011518.htm	38243	-777010119	PHONENUM	UNKNOWN	<br /> 671-8129.</p>
9	9	296010012	qatarliving.com\1011518.htm	40879	-777010119	PHONENUM	UNKNOWN	<p>Hi, we will be travelling out to th
10	10	320788772	qatarliving.com\1019971.htm	23036	-777010119	PHONENUM	UNKNOWN	If you would like to apply please em
11	11	296844512	qatarliving.com\1019980.htm	23022	-777010119	PHONENUM	UNKNOWN	If you would like to apply please em
12	12	297822132	qatarliving.com\1036376.htm	22898	-777010118	PHONENUM	UNKNOWN	If you would like to apply please em
13	13	315411376	qatarliving.com\1044551.htm	23708	-777010118	PHONENUM	UNKNOWN	<p>Phone: 974-4912642-46</p>
14	14	319730788	qatarliving.com\1190668.htm	25252	-777010119	PHONENUM	UNKNOWN	<p>Contact:<br /> Ms Jane Song (
15	15	319730056	qatarliving.com\1190668.htm	25279	-777010119	PHONENUM	UNKNOWN	<p>Contact:<br /> Ms Jane Song (
16	16	296317364	qatarliving.com\1300487.htm	23844	-777010119	PHONENUM	UNKNOWN	<p>we had a very good time at one
17	17	313324400	qatarliving.com\1329749.htm	73641	-777010119	PHONENUM	UNKNOWN	<br /> My contact number is 55219
18	18	329823408	qatarliving.com\13335.htm	344	-777010119	PHONENUM	UNKNOWN	537-5472 | Qatar Living
19	19	329824940	qatarliving.com\13335.htm	21674	-777010118	PHONENUM	UNKNOWN	537-5472

Query executed successfully. | VAIDemo\SQLEXPRESS (10.50 RTM) | VAIDemo\Administrator ... | master | 00:00:00 | 881 rows

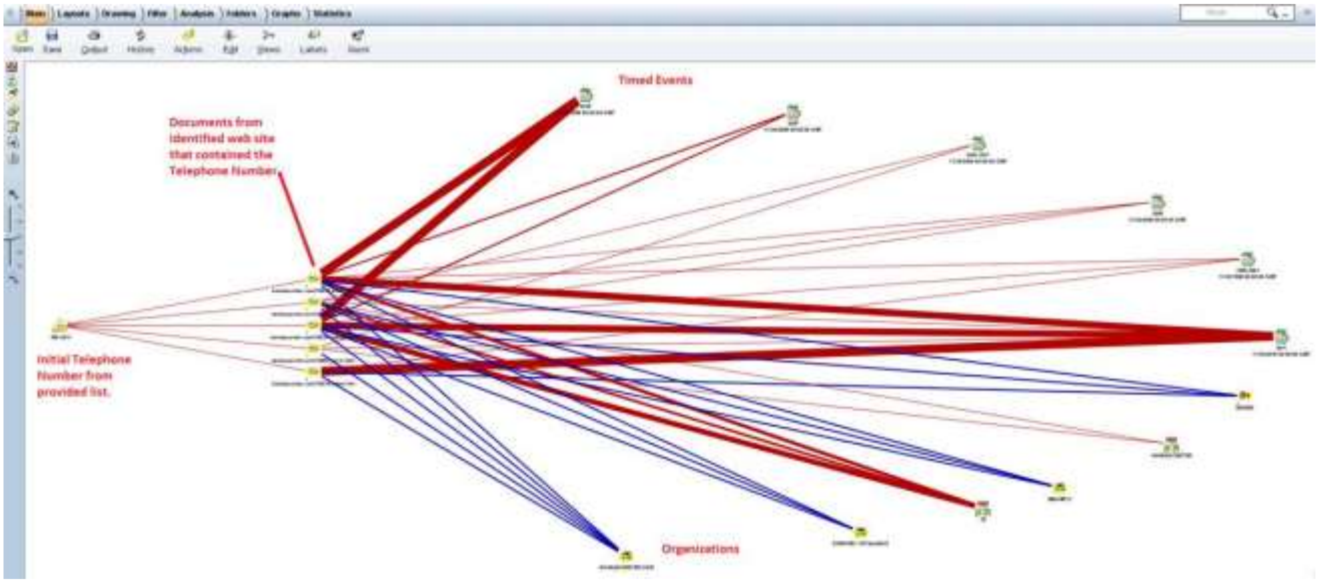
Phase 7: Data Discovery and Insight Visualization with Visual Analytics

The following are some examples of what we discovered along the way as a result of finding web sites that contained telephone numbers of interest.

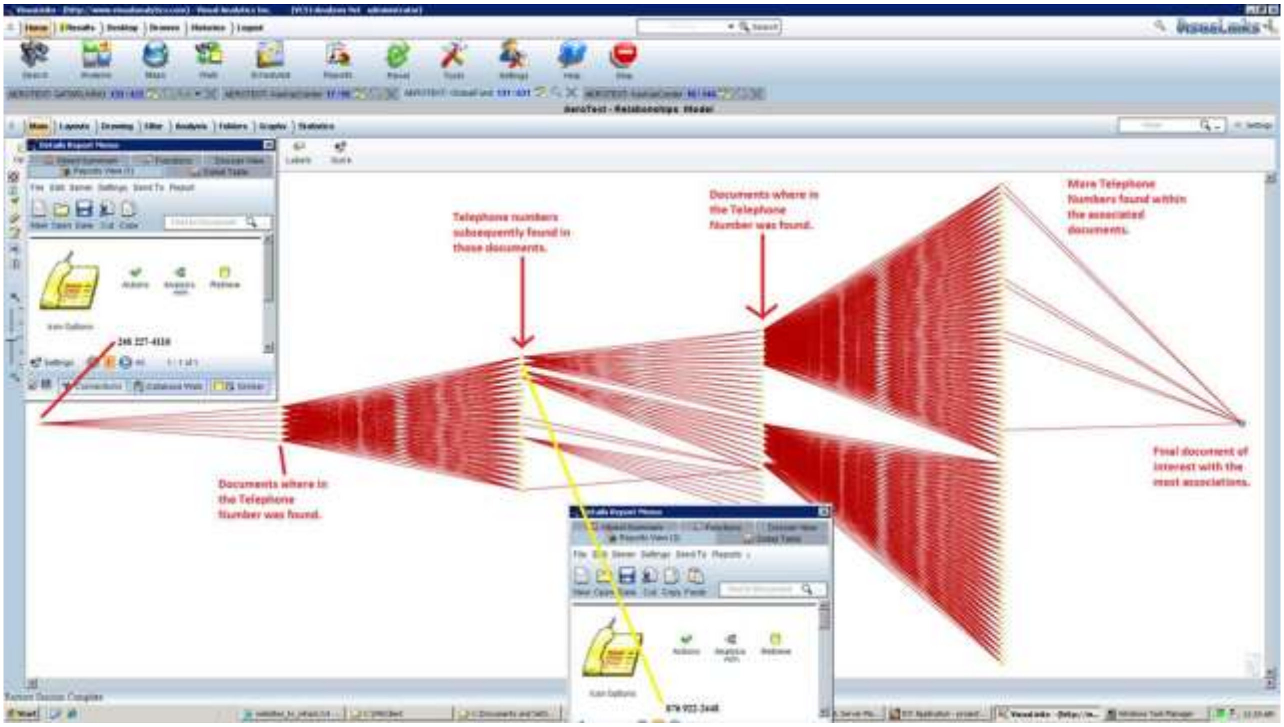
In this example we started with phone number 974-455-2111 and we found the document that contained the number. This document was downloaded from a website identified during our Visual Search Runner process. We then extended the search to find other telephone numbers within the same document. We then extended the search to identify other entities within the document. This is basic Entity Extraction, Data Discovery, and Insight Visualization at a fundamental level.



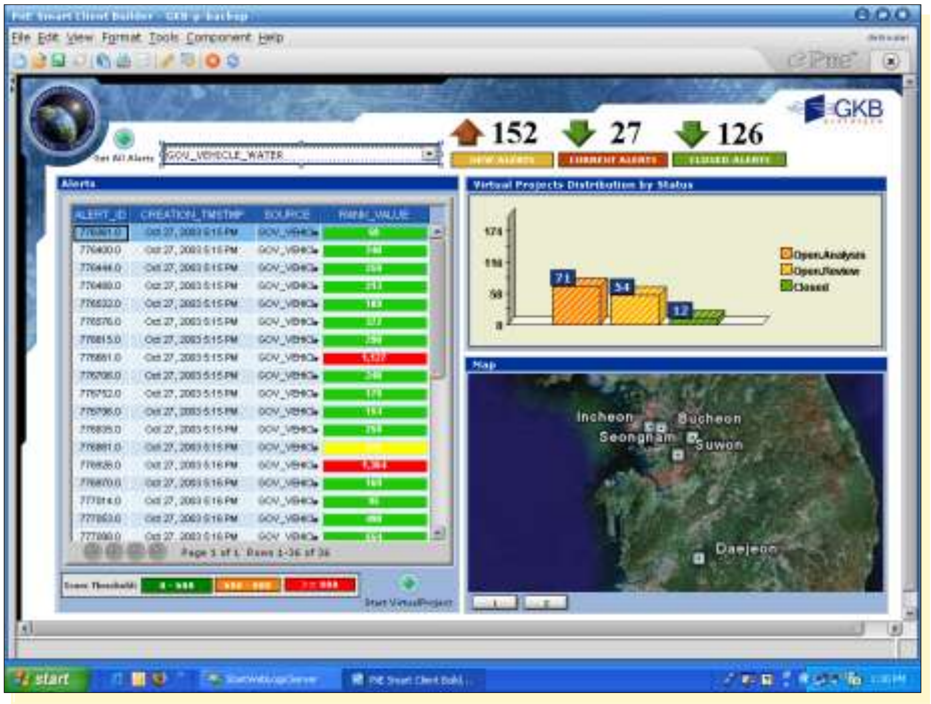
In the following example we started with a telephone number and found all of the documents from the downloaded Kavkaz Center website, a web site that was identified as having a positive search hit on one of the telephone numbers we were interested in. From the initial telephone number we located all of the documents that contained the same telephone number. From there we continued to walk the data from those documents to discovery what other entities existed within them.


Timed Events
Documents from identified web site that contained the Telephone Number.
Initial Telephone Number from provided list.
Organizations

In the following example we start with another initial telephone number of interest and then expand it out related document in order to discover other telephone numbers and eventually more documents of interest.


Telephone numbers subsequently found in those documents.
Documents where in the Telephone Number was found.
More Telephone Numbers found within the associated documents.
Final document of interest with the most associations.
Documents where in the Telephone Number was found.

In the following example we demonstrate how the original unstructured text can be leveraged to find locations information and present it through a geographic based search facility.



ALERT_ID	CREATION_TMSTMP	SOURCE	RANK_VALUE
[illegible]	Oct 27, 2003 5:15 PM	GOV_VEHICL	[illegible]
776400.0	Oct 27, 2003 5:15 PM	GOV_VEHICL	[illegible]
776444.0	Oct 27, 2003 5:15 PM	GOV_VEHICL	[illegible]
776488.0	Oct 27, 2003 5:15 PM	GOV_VEHICL	[illegible]
776532.0	Oct 27, 2003 5:15 PM	GOV_VEHICL	[illegible]
776576.0	Oct 27, 2003 5:15 PM	GOV_VEHICL	[illegible]
776815.0	Oct 27, 2003 5:15 PM	GOV_VEHICL	[illegible]
776661.0	Oct 27, 2003 5:15 PM	GOV_VEHICL	1,127
776708.0	Oct 27, 2003 5:15 PM	GOV_VEHICL	[illegible]
776752.0	Oct 27, 2003 5:15 PM	GOV_VEHICL	[illegible]
776796.0	Oct 27, 2003 5:15 PM	GOV_VEHICL	[illegible]
776835.0	Oct 27, 2003 5:15 PM	GOV_VEHICL	[illegible]
776881.0	Oct 27, 2003 5:15 PM	GOV_VEHICL	[illegible]
776826.0	Oct 27, 2003 5:16 PM	GOV_VEHICL	1,304
776870.0	Oct 27, 2003 5:16 PM	GOV_VEHICL	[illegible]
777014.0	Oct 27, 2003 5:16 PM	GOV_VEHICL	[illegible]
777053.0	Oct 27, 2003 5:16 PM	GOV_VEHICL	[illegible]
777090.0	Oct 27, 2003 5:16 PM	GOV_VEHICL	[illegible]

Page 1 of 1 Rows 1-36 of 36

In the case of Deterministic the Analyst already has an idea of what they are looking for. The above examples demonstrate a Deterministic approach because we start with a Telephone Number we know about and begin the search process with this number and expand it out from there.

Each of the above examples serves to develop insight within a single web site database download. In the next example we will demonstrate the real power of Visual Analytics and that is to provide an Analyst with some insight. In our view there are two ways to perform research:

- Deterministic
- Discovery