

## Find Useful Information Hidden in Unstructured Text

High-quality information is important to success in any venture. Today, this can mean that your search engine must identify and understand documents in many languages, and find people, places, products or organizations in emails, web pages, interview transcripts, reports, field notes and many other materials broadly classified as unstructured text. However, names can have multiple renderings, word context is often ambiguous, and fragments of information are scattered across millions of documents, in many languages and writing systems, structured in radically different ways. How can you retrieve and understand the information locked away in the masses of raw data?

### Introducing Rosette

Since 1998, hundreds of leading enterprises and many governments have trusted the Rosette Platform from Basis Technology for their most critical text analysis requirements. As the world's most advanced linguistics toolkit, Rosette is dedicated to helping users discover documents in many languages and extract meaning from them. The latest edition, Rosette 7, includes innovations that help you rapidly explore your raw data to find the information you need – in more languages than ever before.

### Where Rosette can help

- **Search Engines:** advanced linguistics for online and enterprise search engines, Lucene and Solr based engines, social networking and career planning sites, social monitoring, search-based applications, etc.
- **Federal Government:** text mining for OSINT, DOCEX, DOMEX, HUMINT, and SIGINT missions, information triage, watchlist monitoring.
- **Legal E-Discovery:** identify and enable search in many languages during identification, processing, review and analysis phases of the electronic discovery reference model (EDRM).
- **Financial Compliance:** watchlist screening, anti-money laundering, and fraud detection.

### Capability you can use

Rosette 7 is a suite of components designed to help you examine raw data, process it intelligently, and put it to work. These building blocks can be assembled into flexible solutions that fit the requirements of your application, working seamlessly within your current workflows while handling many different languages, character sets, and data sources. Rosette can enhance any application that depends on extracting meaningful intelligence from huge volumes of unstructured text—accurately, quickly, and cost-effectively.

Whether you need to offer a more powerful search tool, uncover crucial evidence in a legal matter, or catch a money launderer or terrorist, Rosette will make a difference. Now you can find the intelligence you're looking for—or quickly discover new information hidden inside your document collection.

### Dependability You Can Trust

"[Google](#) selected Rosette to provide the Asian linguistic technology needed to create the ultimate Chinese, Japanese and Korean search engine," said Urs Holzle, Google Fellow and Vice President. "This marks a key milestone in establishing Google as the preferred search engine for Internet users worldwide."

"We're really impressed with the Rosette Name Translator's capabilities and how it has improved our OFAC name checking process," said Peter Wilkinson, VP of Application Development at [NCB Capital](#).

Jim King, CEO of [IPRO](#) said, "we partnered with Basis Technology to ensure that all documents demanded in litigation will be identified and analyzed by empowering our applications with the best linguistics-based solutions. This combination lowers the risk of missing critical evidence hiding in documents containing different languages."

### Trusted by Our Customers

Rosette software is used by industry-leading technology companies and government organizations in a wide range of markets. Customers include:

- U.S. defense and intelligence agencies
- Yahoo!
- EMC
- Microsoft Bing & Fast Search
- Clearwell
- Oracle
- Amazon.com
- ...and many others

## Flexible Building Blocks

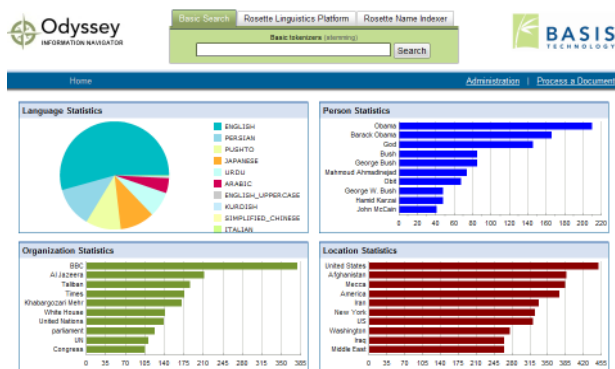
Rosette 7 finds critical intelligence buried in unstructured text, in a single language or across 55 European, Asian, and Middle Eastern languages: By plugging Rosette components into your application, you can:

- **Identify** the language(s) in a document so applications can properly analyze, process and store the data – even when there are multiple languages used in a document or email
- **Analyze** text, taking into account language and context – making your search engine smarter and saving time for your users
- **Extract** important concepts (e.g., names, locations, dates)
- **Translate** names from their native writing system into English
- **Match** names between lists and documents, including variants in multiple languages and writing systems

## Flexible Solutions that Enhance Your Application and Workflow

Rosette 7 is designed for a wide range of large-scale applications that need to identify, classify, analyze, index, and search unstructured text from various sources. It uses sophisticated technology known as *computational linguistics* to establish the true meaning of digital text, in English as well as dozens of major European, Asian, and Middle Eastern languages.

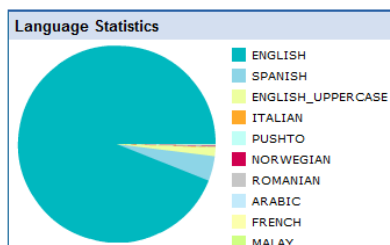
Advanced text analytics is complemented by ease of integration. Rosette 7 components plug into an application through a single API that supports C++, Java, or .NET. Developers can utilize the modules they need where they best within fit their application and workflow, and easily add new capabilities or languages as requirements evolve.



## Rosette Components

### Rosette Language Identifier (RLI)

*Identifies the language(s) and encoding in a document*



RLI identifies the language(s) of a document and its encoding so content can be filtered and processed. RLI can also identify multiple languages

in a single document, recognizing a wide selection of Asian, European, and Middle Eastern languages. Multi-language documents can be segmented into regions that can be routed to separate processes. RLI identifies 55 languages and 35 different legacy encodings with extreme accuracy.

### Rosette Core Library for Unicode (RCLU)

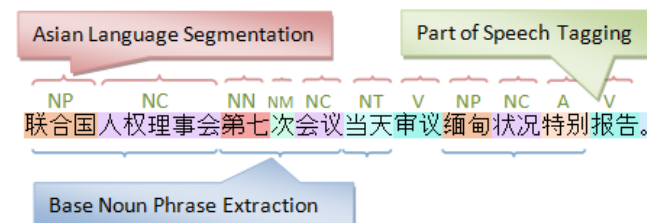
*Display and process information in any language*

Unicode is an international standard that provides a single code base for all the world's languages. Although modern encoding standards mandate the use of Unicode, many existing applications, documents, and data streams utilize legacy encodings such as ASCII, ISO 885901, and countless others. RCLU converts text in over 160 legacy encodings to Unicode to provide a single data source for further processing regardless of language.

### Rosette Base Linguistics (RBL)

*Applies structure to unstructured text*

RBL performs a complete linguistic analysis so search engines can be effective in many languages. RBL processes text the same way a linguist would – using context-sensitive processing of ambiguous terms. For Asian languages, RBL segments the text using a highly accurate combination of language models and customizable dictionaries. In addition, RBL identifies parts of speech, base noun phrases, sentence boundaries, lemmas – or dictionary forms – within a document, in European, Asian, and Middle Eastern languages. A highly accurate morphological approach to analyzing text reflects Basis Technology's deep understanding of the world's languages.



## Rosette Entity Extractor (REX)

*Locates names, places, dates, and other entities*

REX sifts through unstructured text and locates concepts that can be used in many ways. REX takes a hybrid approach, leveraging cutting-edge statistical modeling, customizable rules, and list-based entity extraction to find *entities*—names, places, dates, and other words and phrases that establish the real meaning in text for further analysis. This hybrid model allows REX to use the best engine for each entity type – dates, times, and email addresses are best served with the rules-based engine, while people, organizations and locations rely on the statistical engine. Finally, user taxonomies or simple lists, like weapons, are addressed with the list-based extractor. In Rosette 7, Basis has completely overhauled the statistical extraction engine by taking advantage of the latest research – which has improved both accuracy and speed. The new statistical modeling approach has also dramatically reduced the time it takes to train a new model, which opens the door to customizing the default statistical models.

Named Entity (# instances)	Text
IDENTIFIER:MONEY 1	The fictitious Indian philosopher Abhay Urjavaha Revati was born on January 3, 1624 at 4:00 PM, in Aranaei, a town in the Nagpur province of India. His father, Mr Krupali Setar, a theologian of some repute, together with his mother, Mrs. Sushanti Singh, worked at the Central Bank of India. Abhay was the vice-president of the United Humanist Philosophers Association, which is located in Thiruvananthapuram.
IDENTIFIER:NUMBER 1	
IDENTIFIER:URL 1	
LOCATION 8	
NATIONALITY 1	
ORGANIZATION 2	
PERSON 9	He married his wife Madam Savita Narvas on Dec. 5th, 1657 in Mumbai. They had several children together, all of whom where born in Southeast Asia, with the exception of their youngest daughter Kishana who was born in Boston, Massachusetts.
RELIGION 1	
TEMPORAL:DATE 2	
TEMPORAL:TIME 2	
TITLE 1	Abhay's Hindu background, the influence of his parents, and his favorite teacher, Professor Sungul Kahani all contributed his great work in religious philosophy.
	His daughter Kishana took two years off from work to dedicate a memorial website to her father. The website ( www.revati.com ) raised over \$5 million for children's hospitals across the country. About one million went to Deodhar Children's Hospital.

## Rosette Name Indexer (RNI)

*Match names across writing scripts and languages*

RNI matches names against user-created lists or databases, within a single language or across different languages and scripts. For example a search for “Mao Zedong” locates documents written in the Latin alphabet (“Mao Ze Dong” or “Mao Tse Tung”), in Simplified Chinese (毛泽东), in Traditional Chinese (毛澤東), or in other alphabets such as Arabic (مهاو تسدي تونغ) or Cyrillic (Мао Цзэдун).

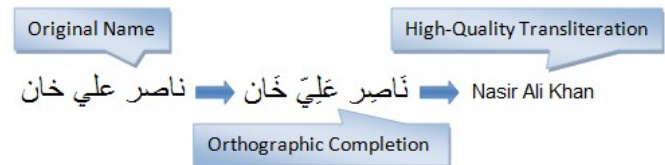
RNI uses knowledge of different cultures and writing systems to handle matching problems and errors like: missing or out-of-order name components, phonetic similarity, nicknames, titles, initials, or combinations of these - and can do this even if the names are written in different scripts.

Nickname	Correct Initial	Search Results
		Search Term: Chuy A. Deaz
		Phonetic Similarity
Matching Name	Match %	Additional Information
Jesus Alfonso LOPEZ DIAZ	53	Program: SDNTK. Nationality: Mexico. Citizenship: Mexico. ID: C.U.R.P. LODJ620930HSLPZS09 Mexico. c/o ESTABLO PUERTO RICO S.A. Sinaloa, Mexico. Date of Birth: 30 Sep 1962.
Missing Name Component		Similarity Score
Results generated by Basis Technology Rosette Name Indexer		

## Rosette Name Translator (RNT)

*Provides the precise, correct English version of a name*

RNT combines name translation and name transliteration to find the most accurate English representation of a name, even when common practice has superseded the “computationally correct” version. It helps analysts and decision-makers rapidly recognize important names in lists, phone books, cell phone address books, and in documents. RNT can also be used to enhance the results of automatic machine translation systems, which commonly mistranslate names when they contain common words or when examining ambiguous sentences and paragraphs.



Together, Rosette 7 building blocks provide end-to-end functionality for analyzing unstructured text, with unmatched accuracy, flexible configurations, and compliance with industry and international standards.

© 2010 Basis Technology Corporation. All rights reserved. “Basis Technology”, “Rosette”, and “We Put the World in the World Wide Web” are registered trademarks. All other company and product names mentioned are trademarks or registered trademarks of their respective owners. (2010-03-08)

### Browse

www.basistech.com

### Write

info@basistech.com

### Call

617-386-2090 or 800-697-2062

### Boston

One Alewife Center, Cambridge, MA 02140

### San Francisco

171 Second Street, San Francisco, CA 94105

### Washington, D.C.

13800 Coppermine Road, Herndon, VA 20171

