

# Rosette Linguistics Platform

## The Leading Text Analysis Platform for Multilingual, Search-Based Applications

The Rosette® Linguistics Platform is the world’s most widely used component library for multilingual text retrieval and analysis. Rosette provides automatic language identification, text normalization, entity extraction, and entity translation from unstructured text, all in a single, unified framework.

“Text analytics is no longer an academic specialty. It has become a necessary component in most search and discovery software — from selling products, tracking terrorists, delivering news, or playing music — to improving communication among people worldwide. Basis Technology’s new Rosette 7 platform ups the ante with its improvements in accuracy, enabling its customers to power a new breed of intelligent workspace applications.”

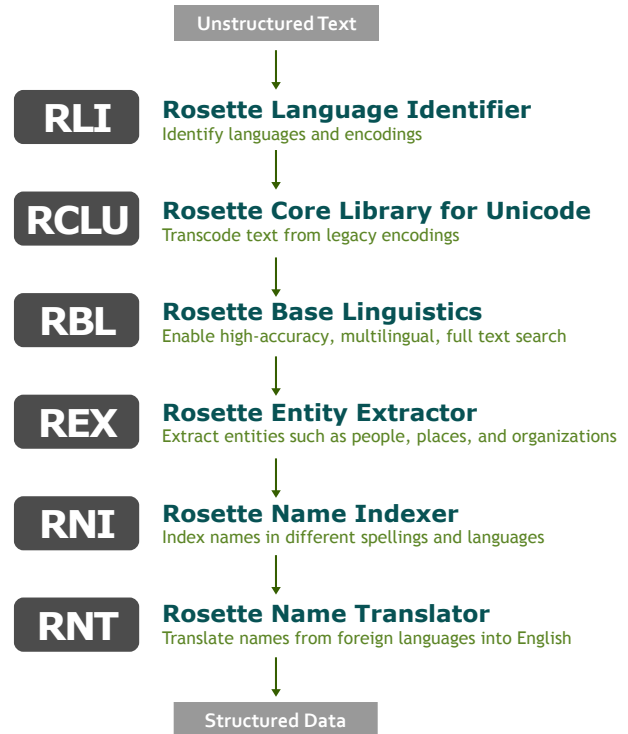
— Susan Feldman, Research Vice President, IDC

### FLEXIBLE BUILDING BLOCKS

Rosette enables your application to examine raw data in multiple languages, process it intelligently, and put it to work. These building blocks can be assembled into flexible solutions that fit your unique requirements, working seamlessly within your existing workflows while enabling new languages, character sets, and data sources. Rosette enhances any application that depends on extracting meaningful intelligence from huge volumes of unstructured text—accurately, quickly, and cost-effectively.

- Automatically identify the language or languages present in a document for correct analysis, filtering, and retrieval of text
- Increase the precision and recall of full-text search in multiple languages
- Extract important concepts from unstructured text, such as names, locations, dates, and identifiers
- Translate names from foreign writing systems into English
- Match names extracted from documents against lists, regardless of language or writing system

Rosette components plug into an application through a single API for Java or web services. Selected components also support C, C++, or .NET. Developers can utilize the modules they need within their application and workflow, and add new capabilities or languages as requirements evolve.



Businesses deploying the popular Apache Lucene and Apache Solr open source search toolkits can now benefit from the same advanced linguistic processing used by high-end web and enterprise search engines. Rosette easily integrates with Lucene to index and search text in many European languages as well as the complex writing systems of the Middle East, Central Asia, and East Asia.

### SELECTED COMMERCIAL CUSTOMERS

Attivio	Exalead/Dassault Sys	NTT Resonant
Autodesk	Fujitsu	Microsoft
Cisco	Google	Oracle
Clearwell Systems	Hitachi	Software AG
EMC	HP	Symantec
Endeca	NEC	Yahoo!

### SELECTED GOVERNMENT CUSTOMERS

Berico Technologies	Northrop Grumman
CACI International	SAIC
In-Q-Tel	U.S. Department of Defense
Lockheed Martin	U.S. Department of Justice
MITRE	U.S. Intelligence Community

Arabic	Chinese (Simp.)	Chinese (Trad.)	Czech	Danish	Dutch	English	Finnish	French	German	Greek	Hebrew	Hungarian	Italian	Japanese	Korean	Norwegian	Pashto	Persian	Polish	Portuguese	Romanian	Russian	Spanish	Swedish	Thai	Turkish	Urdu
--------	-----------------	-----------------	-------	--------	-------	---------	---------	--------	--------	-------	--------	-----------	---------	----------	--------	-----------	--------	---------	--------	------------	----------	---------	---------	---------	------	---------	------

### Rosette Base Linguistics

Divide sentences into words	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Identify parts-of-speech	Y	Y	Y	Y		Y	Y		Y	Y	Y		Y	Y	Y	Y			Y		Y		Y	Y				
Locate sentence boundaries	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Extract noun phrases	Y	Y	Y			Y	Y		Y	Y				Y	Y						Y			Y				
Derive dictionary forms (lemmas)	Y	n/a	n/a	Y	Y	Y	Y		Y	Y	Y		Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y			Y
Split compound words	n/a	Y	Y	n/a	Y	Y	n/a		n/a	Y	n/a		Y	n/a	Y	Y	Y	P	n/a	n/a	n/a	n/a		n/a	Y			n/a

**Additional Base Linguistics Support**  
 Albanian, Bulgarian, Catalan, Croatian, Estonian, Indonesian, Latvian, Malay, Serbian, Slovak, Slovenian, Ukrainian

### Rosette Entity Extractor

Person	S	S	S		S	S			S	S		P	S	S	S		S	S					S	S				S
Location	H	S	S		S	H			S	S		P	S	S	S		S	S					S	S				S
Organization	H	H	H	G	H	H			H	H	G	P	G	H	H	H	H	H	G		G	H	H					H
Title	S	S	S		G	S			S	S			S	S	S		S	G			G	S	S					G
Nationality	G	G	G			G								G	G			G				G						S
Religion	G	G	G			G																						S
Credit Card Number	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
Distance	R	R	R			R	R		R	R				R	R	R		R	R		R		R	R				
Geographic Coordinate	R	R	R	R	P	R	R		R	R	R		R	R	R	R	P	R	R	R	R	P	R	R	P			R
Money	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
Generic Number	R	R	R			R	R		R	R				R	R	R		R	R		R		R	R				
Personal ID Number	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
Phone Number	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
Email Address/URL	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
Date	R	R	R	R		R	R		R	R	R		R	R	R	R		R	R	R	R		R	R				
Time	R	R	R			R	R		R	R				R	R	R		R	R		R		R	R				

### Rosette Name Indexer

Match native script against English	Y	Y	Y			Y								Y	Y		Y	Y					P					Y
Match spelling differences	Y	Y	Y			Y								Y	Y		Y	Y					P					Y
Match missing parts	Y	Y	Y			Y								Y	Y		Y	Y					P					Y
Match missing spaces		P	P			Y								Y	P								P					

### Rosette Name Translator

Native Language Names → English	Y	Y	Y			Y								Y	Y		P	Y					Y					P
Foreign Language Names → English	Y					Y								Y									Y					

<b>Base Linguistics</b>	<b>Entity Extractor</b>	<b>Name Indexer &amp; Translator</b>
<b>Y</b> Fully supported capability	<b>S</b> Statistical model	<b>Y</b> Fully supported capability
<b>n/a</b> Not applicable to this language	<b>G</b> Gazetteer entry (exact match)	<b>P</b> Partially supported capability
	<b>H</b> Hybrid (statistical & gazetteer)	<b>n/a</b> Not applicable to this language
	<b>R</b> Regular expression (pattern match)	
	<b>P</b> Regular expression (partial support)	

VISIT [www.basistech.com](http://www.basistech.com) WRITE [info@basistech.com](mailto:info@basistech.com) CALL 617-386-2090

© 2011 Basis Technology Corporation. "Basis Technology", "Geoscope", "Odyssey Digital Forensics", "Rosette", and "We Put the World in the World Wide Web" are registered trademarks of Basis Technology Corporation. All other trademarks, service marks, and logos used in this document are the property of their respective owners. (2011-12-20)