

ROSETTE

MULTILINGUAL TEXT EXTRACTION



With Rosette, build an enterprise architecture capable of exploiting all-source text intelligence in multiple languages:

- **Triage** incoming documents and message traffic according to language, encoding, and file type
- **Extract** key names, addresses, dates, and concepts from unstructured text
- **Resolve** multiple spelling variants of foreign names.
- **Translate** foreign names consistently to government-specified standards
- **Integrate** a single platform with a single API for both text analytics and entity analytics
- **Achieve** greater accuracy than legacy “machine translation” solutions
- **Support** DOCEX, DOMEX, CELLEX, HUMINT, OSINT, and SIGINT missions

Rosette for Multilingual Text Extraction

Rosette Base Linguistics — RBL

High-performance multilingual text retrieval

Full-text search engines are ubiquitous. We access them daily on the Internet, in the office, and on our home computers. They are used by every branch of government and by every major enterprise. And inside each search engine is sophisticated technology known as *computational linguistics*, the automated analysis of digital text which enables it to be rapidly stored, searched, and retrieved.

For over a decade, the most widely used Internet and enterprise search engines have relied on Rosette Base Linguistics to provide essential linguistic services, including tokenization, lemmatization, decompounding, part-of-speech tagging, sentence boundary detection, and noun phrase extraction.

“Google selected Basis Technology to provide the Asian linguistic technology needed to create the ultimate Chinese, Japanese, and Korean search engine. This marks a key milestone in establishing Google as the preferred search engine for Internet users worldwide.”

— Urs Hölzle, Fellow and Vice President, Google

The same linguistic technology which powers Google, Yahoo!, Bing, and Amazon is now available for search engines of all sizes and budgets. Rosette Base Linguistics is currently offered in 24 languages, with more under development:

Arabic	German	Persian (Farsi/Dari)
Chinese (Simplified)	Greek	Polish
Chinese (Traditional)	Hebrew	Portuguese
Czech	Hungarian	Romanian
Danish	Italian	Russian
Dutch	Japanese	Spanish
English	Korean	Swedish
French	Norwegian	Urdu

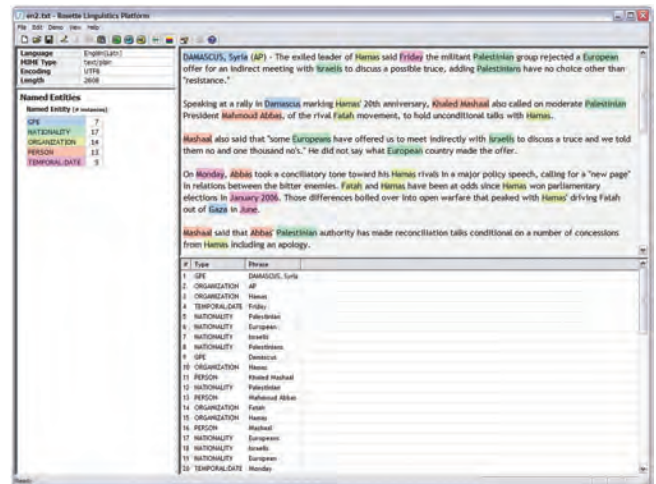
Available Languages

Rosette Base Linguistics is frequently deployed with open source search engines and search applications built with *Apache Lucene* and *Apache Solr*. Both of these technologies are supported by the Apache Software Foundation and are available under the Apache Software License.

Rosette Entity Extractor — REX

Extraction of names, places, dates, and key concepts

Entities — such as names, places, organizations, and dates — are frequently the most critical data in text. The Rosette Entity Extractor automatically scans huge volumes of unstructured documents to find these entities, which enable search and text exploitation systems to perform triage and identify concepts that cannot be found through simple keyword matching.



The Rosette Entity Extractor is pre-trained and tuned for each of 15 supported languages, and it is also customizable to meet the particular entity needs of a specific domain such as healthcare, finance, manufacturing, and intelligence.

Person	URL
Organization	Email address
Location	Phone number
Personal titles	Credit card number
Latitude/Longitude	Currency
UTM coordinate	Date
Personal ID (SSN)	Time
Nationality	Number
Religion	Distance

Predefined Entity Types

REX employs three entity detection technologies — statistical modeling, regular expressions, and gazetteers — to achieve industry-leading accuracy.

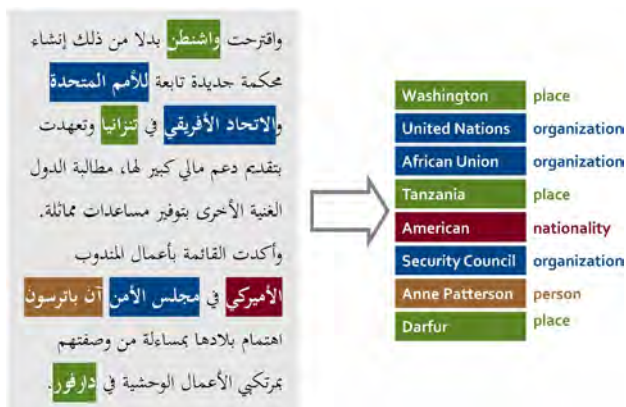
Languages supported include Arabic, Chinese (simplified and traditional), Dutch, English, French, German, Italian, Japanese, Korean, Persian (Farsi/Dari), Pushto, Russian, Spanish, and Urdu — with others under development.



Rosette Name Translator — RNT

Entity translation from foreign languages to English

Legacy “machine translation” systems translate names with widely varying degrees of quality. Combining a high-accuracy entity extractor, such as the Rosette Entity Extractor, with a high-accuracy entity translator, such as the Rosette Name Translator, delivers consistent, high-quality, meaningful results on many types of input.



The Rosette Name Translator provides multilingual name translation through a combination of dictionaries, linguistic algorithms, and statistical inference to derive accurate translations. For frequently appearing names, RNT supplies the “conventional spelling,” and for other names RNT applies the user-selected transliteration standard resulting in consistent translations.



Applications of the Rosette Name Translator include preparing names for indexing, pre-processing text to avoid machine translation errors, and assisting linguists in translating foreign names.

Languages supported include Arabic, Chinese, Dari/Farsi, Japanese, Korean, Pushto, Russian, and Urdu — in both their native scripts and the Latin script.

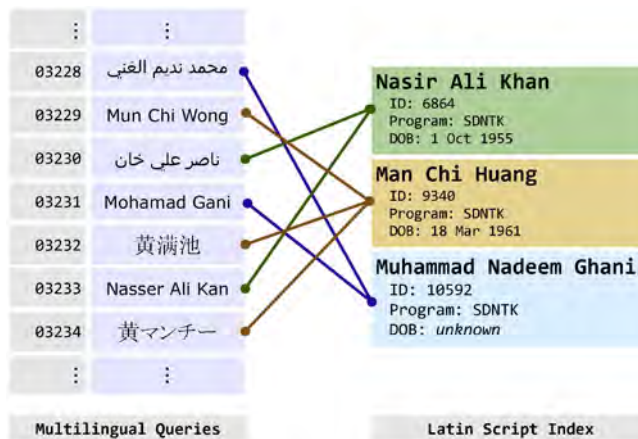
Rosette Name Indexer — RNI

Cross-script name matching and normalization

The Rosette Name Indexer matches names of people, places, and organizations written in different languages — or the same language — against a single, universal index.

Unlike legacy solutions driven by lists containing billions of spelling variants, the Rosette Name Indexer analyzes the intrinsic structure of each name component, and performs an intelligent comparison using advanced linguistic algorithms. This approach is not limited to a particular list of variants and reduces the likelihood of both “false positives” (large numbers of wrong hits) and “false negatives” (zero hits or a failure to uncover relevant matches).

Types of name variations handled by the Rosette Name Indexer include spelling errors, transliteration variations, initials, nicknames, reordered name components, missing name components, missing spaces, truncated name components, and the same name written in different languages or scripts.



The Rosette Name Indexer matches names of people, places, or organizations against entries in a multilingual database, and returns results ranked by a similarity score. When data is incomplete, partial matches may be returned.

This capability is designed to be integrated into such applications as watch list management, fraud detection, money laundering, geospatial analysis, and document triage.

Languages supported are the same as those supported by the Rosette Name Translator.

About Basis Technology

Basis Technology is the leading provider of software solutions for extracting meaningful intelligence from unstructured multilingual text.

Our products and services are used by over 250 major firms including Cisco, EMC, Endeca, HP, Microsoft, Oracle, and Symantec.

Our text analysis products are widely used in the U.S. defense and intelligence industry by such firms as CACI, Lockheed Martin, MITRE, Northrop Grumman, SAIC, and SRI. We are also the top provider of multilingual search technology to web search engines, such as Answers.com, Ask.com, Google, Microsoft Bing, and Yahoo!

Our Rosette linguistics platform is the world's most widely used family of commercial software products for multilingual text retrieval and analysis. Rosette provides services including automatic language identification, Unicode text normalization, entity extraction from structured or unstructured text, and name matching and translation across or within languages.

"Basis Technology" and "Rosette" are registered trademarks of Basis Technology Corporation. All other company and product names mentioned herein are trademarks or registered trademarks of their respective owners.

© 2010 Basis Technology Corporation. All rights reserved. (2010-11-24)

Select Government-Related Customers

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

Select Commercial Customers

Attivio
Autodesk
Cisco
EMC
Endeca
Google
HP
Kroll
Mark Logic
Microsoft
Oracle
Symantec
Yahoo!

Browse

www.basistech.com

Write

info2010@basistech.com

Call

617-386-2090 or 800-697-2062

Boston

One Alewife Center, Cambridge, MA 02140

Washington

2553 Dulles View Drive, Herndon, VA 20171

San Francisco

171 Second Street, San Francisco, CA 94105