MarkLogic®

Language Support

MarkLogic® was designed from the start to handle massive amounts of content, and make that content immediately searchable. MarkLogic makes managing multi-lingual data easy, being language-aware so that you get full comprehension and meaning in all of your data.

Load and Query in Multiple Languages

- **Indexing, tokenizing, and stemming** – MarkLogic will recognize the language of content during load, and then will index, tokenize, and stem the content for querying
- **Confidence scores** – Language detection is difficult, which is why MarkLogic returns a list of possible languages with confidence scores so you can further assess and determine the most appropriate language to ingest the document
- **Tagging elements** – Before loading, you can specify languages in XML documents at the element level by using the xml:lang attribute
- **Custom dictionaries** – You can create custom dictionaries that include stemming and tokenization for proper nouns, technical vocabularies, and spelling variations
- **Language-aware search** – All searches in MarkLogic Server are language-aware. All queries search the text, tokenize the search terms, and stem in a particular language
- **Language encoding** – MarkLogic translates many different document encodings into UTF-8 upon ingestion, though you can also specify the encoding. Supported encodings include ISO-8859-1, ISO-8859-5, ISO-8859-6, cp1251, cp1252, cp1256, Shift JIS, ISO-2022-JP, ISO 2022-KR, GB18030, Big5, and many others

Advanced Language Support

Advanced language support includes language-specific tokenization, stemming, and collation for custom ordering of results. MarkLogic includes advanced language support for English, plus 15 other languages, which are available as options. If the language you are interested in is not on our current list please contact us.

English	German	Arabic	Persian (Farsi)	Portuguese
French	Russian	Chinese (Simplified & Traditional)	Dutch	Norwegian (Nynorsk & Bokmål)
Italian	Spanish	Korean	Japanese	

Basic Language Support

Basic language support enables basic full-text search and includes tokenization using whitespace-delimiters and punctuation. Our basic language support is available for most languages that use Latin, Cyrillic, Arabic, Greek, or Hebrew script, which includes over 200 languages that each have a million or more native speakers.

- Abaza
- Abkhazian
- Achinese
- Adyghe
- Afar
- Afrikaans
- Akan
- Albanian
- Altaic Language
- Amo
- Assyrian Neo-Aramaic
- Asturian
- Atsam
- Avaric
- Aymara
- Balinese
- Bambara
- Bashkir
- Basque
- Batak
- Batak Toba
- Belarusian
- Bemba
- Bini
- Bislama
- Bosnian
- Breton
- Buginese
- Bulgarian
- Bushi
- Catalan
- Cebuano
- Chamorro
- Chechen
- Chukot
- Chuukese
- Chuvash
- Comorian
- Cornish
- Corsican
- Croatian
- Czech
- Danish
- Dargwa
- Dogri
- Dogrib
- Dungan
- Dyula
- Eastern Cham
- Eastern Frisian
- Efik
- English-based Creole or Pidgin
- Erzya
- Esperanto
- Estonian
- Evenki
- Ewe
- Fang
- Faroese
- Fijian
- Filipino
- Finnish
- Finno-Ugrian Language
- Fon
- Friulian
- Ga
- Gagauz
- Galician
- Ganda
- Gilbertesse
- Gorontalo
- Greek
- Guarani
- Guianese Creole French
- Gwich'in
- Haida
- Haitian
- Hanunoo
- Hawaiian
- Hebrew
- Hiligaynon
- Hiri Motu
- Hmong
- Hopi
- Hungarian
- Ibibio
- Icelandic
- Igbo
- Iloko
- Inari Sami
- Indonesian
- Ingush
- Interlingua
- Inupiaq
- Irish
- Javanese
- Jju
- Kabardian
- Kabyle
- Kalaallisut
- Kalmyk
- Kamba
- Kanuri
- Kapampangan
- Kara-Kalpak
- Karachay-Balkar
- Kazakh
- Khakas
- Khanty
- Khasi
- Kikuyu
- Kimbundu
- Kinyarwanda
- Kirghiz
- Komi-Permyak
- Komi-Zyrian
- Koro
- Koryak
- Kosraean
- Kpelle
- Kuanyama
- Kumyk
- Ladino
- Lahnda
- Lak
- Latin
- Latvian
- Lezghian
- Limburgish
- Lingala
- Lisu
- Lithuanian
- Low German
- Lower Sorbian
- Luba-Katanga
- Luba-Lulua
- Lule Sami
- Luo
- Lushootseed
- Luxembourgish
- Macedonian
- Madurese
- Maguindanao
- Makasar
- Malagasy
- Malay
- Maltese
- Mansi
- Manx
- Maori
- Mende
- Minangkabau
- Moksha
- Moldavian
- Mongo
- Morisyen
- Mossi
- Muslim Tat
- Nanai
- Nauru
- Navajo
- Naxi
- Ndonga
- Neapolitan
- Nenets
- Niuean
- Nogai
- North Ndebele
- Norwegian Nynorsk
- Nyamwezi
- Nyanja
- Nyankole
- Occitan
- Norwegian Nynorsk
- Nyamwezi
- Nyanja
- Nyankole
- Occitan
- Oromo
- Palauan
- Marshallese
- Pangasinan
- Papiamento
- Parsi-Dari
- Pashto
- Pohnpeian
- Polish
- Prussian
- Quechua
- Reunion Creole French
- Rhaeto-Romance
- Romanian
- Rundi
- Russia Buriat
- Samaritan Aramaic
- Sami Language
- Samoan
- Sango
- Sasak
- Scots
- Scottish Gaelic
- Selkup
- Serer
- Shona
- Shor
- Sicilian
- Sidamo
- Skolt Sami
- Slovak
- Slovenian
- Somali
- Songhai
- Soninke
- Sundanese
- Susu
- Swahili
- Swati
- Swedish
- Sundanese
- Susu
- Swahili
- Swati
- Swedish
- Swiss German
- Tabassaran
- Tagalog
- Tagbanwa
- Tahitian
- Tamashek
- Tatar
- Tausug
- Tetum
- Timne
- Tiv
- Tok Pisin
- Tokelau
- Tonga
- Tsonga
- Tswana
- Tumbuka
- Turkish
- Turoyo
- Tuvalu
- Tuvinian
- Twi
- Tyap
- Udihe
- Udmurt
- Uighur
- Ukrainian
- Ulithian
- Umbundu
- Upper Sorbian
- Urdu
- Venda
- Vietnamese
- Volapuk
- Walloon
- Yakut
- Yao
- Yapese
- Yiddish
- Yoruba
- Yakut
- Yao
- Yapese
- Yiddish
- Yoruba
- Zulu

About MarkLogic

For more than a decade, MarkLogic has delivered a powerful, agile, and trusted Enterprise NoSQL database platform that enables organizations to turn all data into valuable and actionable information. For more information, please visit www.marklogic.com.

© 2014 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED. This technology is protected by U.S. Patent No. 7,127,469B2, U.S. Patent No. 7,171,404B2, U.S. Patent No. 7,756,858 B2, and U.S. Patent No 7,962,474 B2. MarkLogic is a trademark or registered trademark of MarkLogic Corporation in the United States and/or other countries. All other trademarks mentioned are the property of their respective owners.

MARKLOGIC CORPORATION
999 Skyway Road, Suite 200 San Carlos, CA 94070
+1 650 655 2300 | +1 877 992 8885 | www.marklogic.com | sales@marklogic.com