What native-language searches can be performed in the Summon service?

The Summon service provides native-language search capabilities in many languages, and employs several strategies in combination to provide accurate search results. We are working to provide enhanced search capabilities in other languages that are as precise and as accurate as an English-search query.

Patrons can search in any language; however, the Summon service provides richer search capabilities and results for the languages listed below. As always, search results will vary depending on your library's database subscriptions and resources.

The Summon service currently supports native-language searches in:

Arabic
Chinese (Simplified and Traditional)
Danish
Dutch
English
French
German
Japanese
Italian
Korean
Malay
Portuguese
Romanian
Spanish
Swedish
Thai
Turkish

The Summon service uses several techniques to provide accurate non-English search capabilities. Some of the most important processes are listed below. These processes are applied to search results based on the language of each Summon record. For example, English search features (tokenization, stemming, etcetera) are applied to English records and German search features are applied to German records.

- Tokenization
- Decompounding
- Stemming/Lemmatization
- Character Normalization
These techniques are described in detail below. Click the links above to jump to the descriptions.

This document also contains information about other Summon native-language search features:

- [Verbatim Match Boost](#)
- [Preferred Language Boost](#)
- [Native Language Boolean Operators](#)

As well as an FAQ (Frequently Asked Questions) about Native-Language Search features:

- [FAQ (Frequently Asked Questions)](#)

Lastly, the Summon service maintains language specific lists of stop words (words such as "a" and "of" in English). For more information about how Summon handles stop words, please see the [Stop Words](#) document.

---

**Tokenization**

Tokenization is the process of breaking a stream of letters or text into words, phrases, or meaningful elements. Tokenization is part of Summon's language-specific text analysis, which is performed at both index time and query time, and resulting tokens constitute the smallest searchable unit in the Summon service.

In most languages, words are separated by white space or punctuation, so tokenization is a simple process for those languages. However, in languages such as Chinese, Japanese and Thai, words are not separated by white space. For these languages, Summon's text analysis uses sophisticated techniques to identify word boundaries, and use that information to perform tokenization.

Examples of tokenization:

black cat  becomes the two searchable units black and cat, becomes the three searchable units and and .

Hyphenated words are tokenized as two separate words without the hyphen. The words must, however, appear as a phrase in order for Summon to consider the words to be a match with the hyphenated word.

For example:

the query health-care will return results for "health-care" and "health care", but not "healthcare" or a non-exact phrase such as "health and care".

Note: We are working to address cases like the above health-care and health care versus healthcare as part of the spelling variation normalization, described in the Synonym Mapping and Spelling Normalization section of this document.
**Decompounding**

Compound words are words that consist of multiple components that can stand as individual words on their own. In languages such as German, Swedish, and Danish, compound words are spelled without white space, and as a result, they can be very long.

Decompounding is the process of finding constituent parts in a compound word, and the Summon service performs this process for languages such as German, Swedish, Danish and Korean. This process allows the patron to search for those constituent parts and get matches on the compound word.

Example:

the German search abwasser anlagen (wastewater plant in English) returns results for the compound word abwasserbehandlungsanlage (wastewater treatment plant)

---

**Stemming/Lemmatization**

Stemming is the process of reducing inflected (or sometimes derived) words to their stems, or the root forms. Lemmatization is the process of converting various forms of a word to its dictionary form. Despite the slight differences, these processes have the same goals, and these terms are often used interchangeably. The Summon service performs language-specific stemming or lemmatization to allow the patron to search for a form of a word and get matches on other forms of the same word.

Examples:

books book (English)
ponies pony (English)
theses thesis (English)
maisons maison (French)
grandes grande (French)
Kinder Kind (German)
(Japanese)

In other words, the first example above (the English search book) will return results for both book and books. A French example search grande maison will return results for both grande maison and grandes maisons. The phrase search "grande maison" (with double quotes), as well, would have stemming and lemmatization applied to it but verbatim matches will appear higher in the results to allow the user to focus on the most relevant matches.

---

**Character Normalization**

Character normalization is the process of normalizing variants of a character to its basic version. Characters with diacritics are, for example, normalized to the characters without diacritics. The Summon service also provides character normalization for variants of Chinese characters.

Character normalization allows the patron to search for a word containing a diacritic and get results on the word without the diacritic, and vice versa. Similarly, it allows the patron to search for a Chinese word using the traditional characters, and get hits on the word spelled with the simplified characters, and vice versa. The character normalization mappings are mostly the same across all languages, but in some cases, language specific character normalization mappings are defined.

Examples:
The Chinese search will return results for 墨西哥, the Spanish search mexico will return results for méxico.

In some cases, the Summon service allows for multiple ways to represent a character with a diacritic. For example, the German umlauts ä, ö, and ü can be spelled without the diacritic as ae, oe and ue, or a, o, and u. The Summon service allows both variations. This allows the patron to search for schoen or schon and get hits on schön. Another example is the Spanish ñ. It can be searched for using ñ, n, or ni, and that allows for query terms Espanol and Espaniol to return Español results.

**Transliteration**

Transliteration is a conversion of one script to another. This process allows for searching in one script and get hits on the same words written in another script.

The Summon service currently provides transliteration search features for Chinese (Hanzi-Pinyin), Japanese (Kanji/Katakana-Hiragana) and Korean (Hanja-Hangul) for titles and author names. Chinese Pinyin transliterations can be written with spaces between words (for example, beijing daxue), or with spaces at the Hanji-character boundaries or syllable boundaries (for example, bei jing da xue).

Examples:

The Chinese query beijingdaxue ("Peking University" in Pinyin transliteration) would return results containing the string ("Peking University" in Hanzi script).

The same Chinese query written as beijing daxue or bei jing da xue (use double quotes for better results) would also return results containing the string ("Peking University" in Hanzi script).

The Japanese query ("Natsume Souseki" in Hiragana script) would return results containing the string ("Natsume Souseki" in Kanji script).

The Korean query ("Kim Dae-jung in Hangul script) would return results containing the string ("Kim Dae-jung" in Hanja script).

NOTE: If a search is performed using transliteration, then transliterated search results are not necessarily the first results to display.

**Elision Handling**

Elision in this case refers to the omission of a final vowel of a word when the following word begins with a vowel, and is observed in languages such as French and Italian.

For example, in French, the word sequence le + arbre becomes "l'arbre". In Italian, lo + amico becomes l'amico.

The Summon service’s elision handling allows the patron to search for amico and get hits on l'amico.
Synonym Mapping and Spelling Normalization

Summon provides language-specific simple synonym mappings and spelling normalization. For example, in English, theater and theatre are two spellings of the same word. These are normalized during Summon’s English text analysis, and as a result, the patron can search using one of these spellings and get hits on both spellings. Language-specific synonyms are also defined for cases where two words have the same meaning.

Examples:

theater vs. theatre (English)
accessorize vs. accessorise (English)
analog vs. analogue (English)
vs. (Chinese)
ordenador vs. computadora (Spanish)

Also, the ampersand (&) is equated with the appropriate word for "and" in each language.

Verbatim Match Boost

This is one of the most important features of Summon’s native language search support. Many of the features described in this Answer allow the patron to get results when the query terms and the indexed terms are equivalent, but not exactly the same word for word (verbatim). These features have an effect of increasing the number of results, or in other words, increasing the search recall. While these features will provide a better user experience to the patrons, there is a risk of including less relevant or non-relevant results, and reducing the search precision.

The Verbatim Match Boost feature addresses this concern by boosting the relevance score of a result when the matching of the query terms and the indexed terms is verbatim or near verbatim.

Example:

For the English query theatres, the results for "theatres" get higher relevance scores (and are ranked higher) than the results for "theaters" or "theatre" if all factors that contribute to the relevance scoring calculation are equal.

The Verbatim Match Boost feature is applied to almost all processes mentioned in this support document. The actual implementation of this feature works by penalizing non-verbatim matches, thus effectively boosting the verbatim matches. Penalties are computed for each term/token, and the penalty amount is defined for each process in each language, such that major differences, such as synonyms, are penalized more than minor differences, such as spelling differences and singular vs. plural forms of nouns.

Preferred Language Boost

The Preferred Language Boost feature is the feature that boosts the results in the user interface language the patron is using. This feature is useful when a given query term can be a word in multiple languages, and when the results include records in multiple languages.

For example, the word (psychology) is the same in both Japanese and Chinese. When the patron issues the query psychology, the results may include both Japanese and Chinese records. The Preferred Language Boost feature boosts the Japanese results for the user in the Japanese user interface and the Chinese results are boosted for the user in the Chinese user interface, as such behavior would be the expectation by most users.
The Settings page in the Summon Administration Console provides the ability to select the library's default language for the Summon user interface, along with additional languages that can be offered to users.

### Native Language Boolean Operators

Major search engines in some countries allow the use of Boolean operators (AND, OR, NOT) in their native languages. Specifically, in German speaking countries, users expect to be able to use "UND", "ODER" and "NIGHT" in place of "AND", "OR", and "NOT". The Summon service allows for the use of these German Boolean operators from its German UI. The English operators will continue to work in the German UI.

### FAQ (Frequently Asked Questions)

Below are questions that we tend to hear from libraries about native-language search features:

**Why does searching for different forms of the same word return different numbers of results?**

For example, searching for "joke" and "jokes" in Summon usually returns close, but different numbers of results. Since "jokes" is the plural form of "joke" in English, one might expect Summon should return the same number of results for both queries. The difference is due to the fact that the two word forms "joke" and "jokes" may appear in non-English Summon records.

Summon uses the language of the record to determine which language's text analysis to apply. If the word form "jokes" appears in an Italian record, for example, it is not stemmed to "joke" and therefore does not match the query "joke". This could happen with stemming/lemmatization, character normalization and other language-specific text analysis processes.

**I'm limiting my search to my language using the Language facet, but the search behavior does not seem to be limited to my language. Why?**

Summon distinguishes two types of language attributes: the language of the record and the language of the work itself (for example: book, article). Summon's Language facet uses the latter, while Summon's native-language search features use the former. So, choosing a language in the Language facet does not mean the search is limited to only using that language's search features.

**I'm searching in language X (for example: English). Why do my search results include stemmed matches in another language?**

This is a side effect of the fact that Summon is a true multilingual search engine, and the fact that your library has the rights to multilingual content. Summon's Verbatim Match Boost and Preferred Language Boost features, as well as our carefully designed stemming/lemmatization algorithms, should alleviate this side effect. Summon's development team continues to study any existing problematic cases in order to further strengthen its algorithms.
• **Old Article Number**: 8813