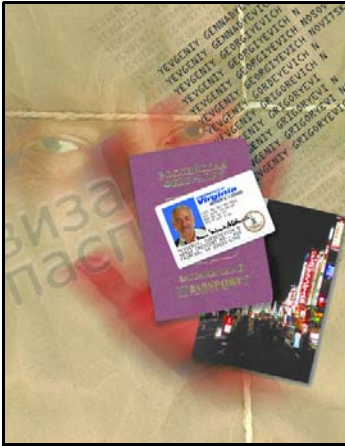


Celatro™

pattern matching library

Russian Name Search Plug-In



Abstract

Finding Russian name variations in transliterated text is a notoriously difficult problem. Since Russian is written in the Cyrillic script, transliteration into the Roman script routinely produces numerous variations of the same name, any number of which may not be known ahead of time. Finding all these variations in a search corpus can be next to impossible.

For example, the name *Josif* can be variously rendered as *Iosiff*, *Yosiph*, and *Jozeph* (to list a few variants). Obviously, if exact-matching is used to retrieve all mentions of *Josif*, these and other alternative spellings will be missed. But standard fuzzy logic (which is based on a character-by-character comparison of two strings) will also miss some alternate spellings, such as those which have very few characters in common (e.g. *Iosiff* and *Jozeph* only share one character, *o*).

The Celatro™ Russian Name Search™ plug-in was designed specifically to address this problem. It contains both algorithmic and language- and culture-specific approaches that increase both the *precision* and the *recall* of transliterated Russian names when compared to other approaches.

Russian Name Taxonomy

Every Russian name has three specific segments:

- The first, or given, name;
- The patronymic; and
- The surname.

e.g., *Pyotr Ilich Tchaikovsky*

First Name: Given

The first name of a Russian is his/her *given name* (e.g., *Anton* or *Elena*).

Russian given names can take very different forms depending on the namer's relationship (emotional, familial, etc.) with the person named. A well-known example of this is the name *Maria* and the nicknames derived from it: *Masha*, *Mashka*, *Mashenka*, *Mashunechka*, *Mashunya*, and so on.

Second Name: Patronymic

The second name of a Russian is his/her *patronymic name*, which is a name based on the person's father.

Russian patronymic names are formed by adding the suffixes — -ovich/-evich (for sons) and -ovna/-evna (for daughters) — to the father's first name. For example:

Men: *Yuri Mikhailovich Zolotov*

Women: *Elena Sergeievna Petrova*

Third Name: Surname

Russian *surnames* are family names that differ depending on the gender of the bearer. Women add the suffix -a to their maiden or married name. For example:

Men	Women
<i>Petrov</i>	<i>Petrova</i>
<i>Ivanov</i>	<i>Ivanova</i>
<i>Zolotov</i>	<i>Zolotova</i>

Celatro Russian Name Search Details

RNS achieves exceptional precision and recall through the combined use of proprietary, high-performance fuzzy search logic and specialized algorithms developed especially for the retrieval of transliterated Russian names.

The Framework

RNS is a Celatro plug-in and, as with all Celatro plug-ins, requires the Celatro Core Library (CCL) to run.

In addition to rapidly matching transliterated Russian names in the given search corpus, RNS also produces rank-ordered results (i.e., from best-to-worst match or vice versa) — and the number of matches is user-specified.

RNS does not restrict database queries in any manner. The user need only provide a mapping between *the name* and *its location* in the search data, (e.g., a column name). The number of columns to be queried is arbitrary, which means that querying more than one column at a time is possible. In practical terms, this means that RNS has no restrictions on the number of names to search (middle names, multiple last names, and so on).

In case of multiple columns, users can choose to assign more weight to a particular column, or to specify that only exact matches for a given column should be included in the search.

The RNS Algorithm

RNS combines several proprietary fuzzy logic routines in order to compare and match transliterated Russian names in a search corpus.

Alternative spellings are compared for phonetic similarity based on the way Russian sounds are transliterated into English. For example, *f*, *ph*, or *ff* represent equivalent transliterations of the same sound, and may be substituted for one another in alternative spelling variants. This approach significantly increases recall since different transliteration variations can have very few — even zero — characters in common.

In addition, RNS also recognizes *alternate* names, which allows for the identification of names that are spelled and pronounced differently in English, but are actually the same. For example, if the user is looking for *Dmitry Petrov*, the entry *Dima Petrov* will be retrieved as well, since *Dima* is a nickname of *Dmitry*.

RNS produces rank-ordered results based on the similarity to a given query. For example, given the input *Michael Gorbacev* and a list of candidate names, RNS would produce results as follows:

Measure	First Name	Last Name
1.00000	Michael	Gorbacev
0.90625	Mihhail	Gorbatsev
0.890625	Mikhail	Gorbachov
0.875	Mihail	Gorbatsov
0.734375	Misha	Gorbachev

RNS Advantages

RNS features several key advantages over other name-matching technologies:

- RNS compares two names and generates a similarity score.
- Rather than simply generating a list of hits, RNS produces ranked results that can include the combined scores of more than one algorithm. This allows for:
 - Ordering of results through the use of a *similarity measure*.
 - Application of sophisticated combinatorics using *fuzzy* or *Bayesian* logic.
 - Application of user-specified thresholds.
- As a Celatro plug-in, RNS leverages all the capabilities of the component-based Celatro technology line, providing highly tuned algorithms that support:
 - Handling of structured and unstructured text.
 - Handling of ASCII and Unicode data.
 - Unlimited algorithm aggregation — including your choice of algorithms developed specifically for Russian name retrieval and/or any other algorithm included in the Celatro class library.