Rosette Name Indexer

Comparison to Common Alternatives

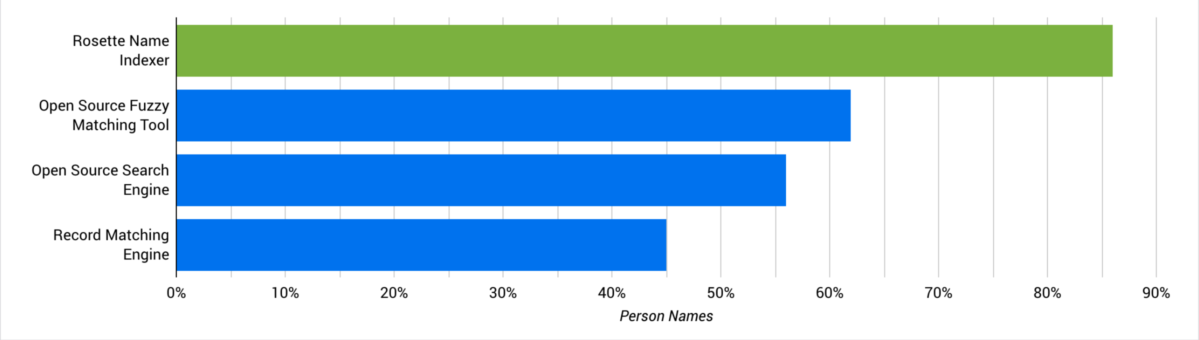
Rosette Name Indexer was evaluated in December 2019 against three common alternatives using a dataset with 7,571 names, with at least 10 variants for each name. These alternatives included:

- An open source fuzzy matching tool
- An open source search engine
- A record matching engine.

Testing and analysis show that these alternatives fall short of Rosette because they lack script/language support, lack essential name phenomena support, and use rigid or overly simplified methods to calculate match scores

Superior Accuracy

Where a correct match is defined as matching a "gold standard" version of a name to one of its variants, Rosette outperforms the alternatives by 24% or more for person name matching.



Coverage of Match Phenomena

	OS Fuzzy Matching	OS Search Engine	Record Engine	Rosette
Exact Match (two names are identical "Tom Jones" ⇔ "Tom Jones")	✓	✓	✓	✓
Normalization (ability to identify matching names whose characters normalize to the same letters "LINDSTROM-JONES" ⇔ "Lindström-Jones")	Partial	✓	Partial	✓
Stop Words (ability to remove "noise words" from names "Mr. Tom Jones" ⇔ "Tom Jones")	None	None	None	✓
Nicknames (ability to recognize common nicknames such as "Thomas" ⇔ "Tommy")	None	None	Partial	✓
Fuzzy Match (statistical model for fuzzy matching)	None	None	None	✓
Truncation (ability to recognize long names cut short "McDonald" ⇔ "McD")	Partial	✓	✓	✓
Cross-lingual (ability to match the same name written in different languages and scripts "一郎" ⇔ "Ichiro")	None	None	Partial	✓
String Similarity (ability to detect similarity due to edit distance "John" ⇔ "Jhon")	None	✓	None	✓
Deletion (ability to take into consideration a missing name component "John Richard Williams" ⇔ "John Williams")	✓	✓	Partial	✓
Out-of-Order Deletion (ability to take into consideration a missing name component in conjunction with other name components having moved "George Herbert Walker Bush" ⇔ "George Bush Walker")	✓	✓	Partial	✓
Initialism (ability to handle organizational name acronyms "ABC" ⇔ "American Broadcasting System")	None	None	None	✓
Initials (ability to handle replacement of a name with an initial "John F. Kennedy" ⇔ "John Fitzgerald Kennedy")	Partial	✓	Partial	✓

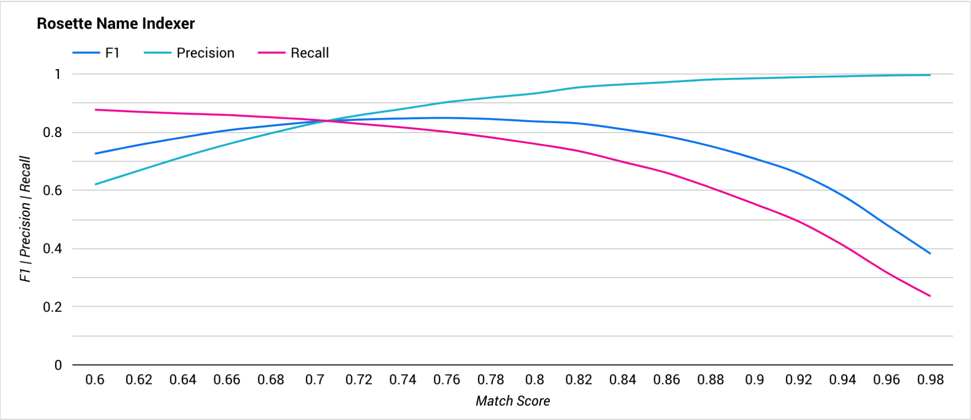
Coverage of Match Phenomena (continued)

	OS Fuzzy Matching	OS Search Engine	Record Engine	Rosette
Reordering (ability to consider components that are a match, but penalize for a mismatch in the order of components "George Herbert Walker Bush" ⇔ "George Walker Herbert Bush")	✓	✓	✓	✓
Insert Spaces (ability to handle name components that appear to have been "glued" together "MuhammadMulan Park" ⇔ "Mulan Park")	Partial	Partial	None	✓
Rotation (ability to avoid over-penalizing for reordered name components "George Herbert Walker Bush" ⇔ "Walker George Bush Herbert")	✓	✓	✓	✓
Concatenation (ability to consider if concatenating tokens produces a better match "Fred Will Sun" ⇔ "Fred Wilson")	None	Partial	Partial	✓
Gender Mismatch (ability to detect when a male name is being compared to a female name and adjust the score accordingly "Joe Smith" ⇔ "Joan Smith")	None	None	None	✓

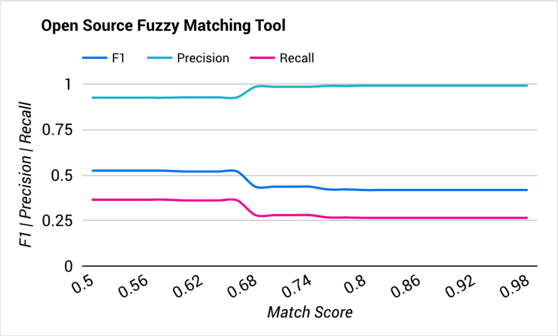
Useful Match Scores

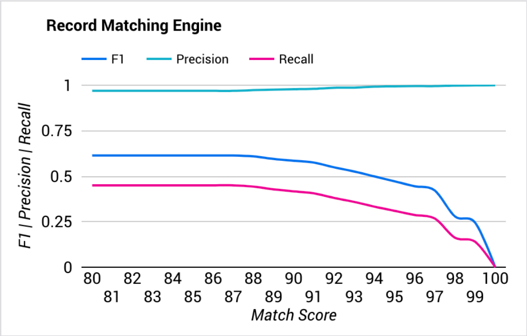
Rosette outputs a nuanced match score as a decimal ranging between 0 (no match) and 1 (perfect match); this match score can be used to balance precision and recall.

As the graph below shows, the precision and recall of Rosette meet at a point around .72; users of Rosette can look at lower scores to see more possible matches, and at higher scores to find only the most similar matches.



By contrast, the **open source matching tool** and the **record matching engine** operate in a binary "match" (score=1) or "no match" (score=0) paradigm without a range to indicate degrees of match. In this case, it is less than clear what threshold will produce the desired balance of precision v. recall.





Furthermore, the **open source search engine** does not provide a comparable score. Thus it is not possible to compare match scores across multiple queries or configure business logic around the results.