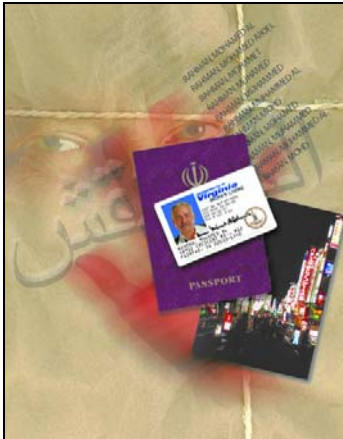


Celatro™

pattern matching library

Arabic Name Search Plug-In



Abstract

Name-matching algorithms typically rely on spelling variations to determine if two names represent the same person. This approach often yields false positives and false negatives, especially when the names have been transliterated into English:

- *Salim* and *Salam*; close spelling but morphologically different
- *Bouchaib* and *Abu Shouwayb*; different spelling but morphologically identical

Arabic Name Search (ANS) is a Celatro plug-in that out-performs other name-matching algorithms because it combines morphological and semantic analysis to produce highly accurate, rank-ordered results.

Arabic Name Taxonomy¹

- Arabic names consist of a complex chain of components that, among other things, identify:
 - Lineage (patrilineal, geographical, etc.)
 - Tribal or clan affiliation
 - Profession
- Arabic names can be much longer than Western names. For example, in Saudi Arabia a legal name for getting a driver's license must have 4 parts.
- Arabic name components themselves consist of subcomponents:
 - Prefixes: 'son of'; 'servant of'; 'father of'; etc.
 - Suffixes: 'of the faith'; family name markers; etc.
 - Optional use of definite article with family names
- Arabic names include five main categories of name elements:
 - *Kunya* (honorific name): uses prefixes 'abu', 'umm' (father of, mother of). Conveys respect. Precedes personal name when full name is used. *Abu Yusuf*, *Umm Jaffar* etc.
 - *Ism* (personal name): personal name given shortly after birth. *Muhammad*, *Ibrahim*, *Hassan* etc.
 - *Nasab* (patronymic name): uses prefixes 'ibn', 'bint' (son of, daughter of). Follows the personal name in usage. Usually limited to three generations. *Osama bin Laden*, *Sumaya bint Khubbat*.
 - *Laqab* (descriptive name): usually religious, relating to nature or describing another admirable quality. Follows isms in usage. *Harun al Rashid* (Harun the Rightly-Guided), *Nur-ad-din* (Light of the religion).
 - ... Special form of laqab: 'abd' (servant of) + 99 names of Allah, e.g.: *Abd-Allah* (the servant of God), *Abd-al-Aziz* (servant of the Almighty).
 - *Nisba*: (bynames):
 - ... Occupational: *Hassan al-Hallaj* (Hassan the dresser)
 - ... Geographical (derived from place of birth): *Mohammad al-Isfahani* (Mohammad of Isfahan), *Uday Saddam Hussain al-Tikriti*
 - ... Of descent (tribe or family name): *Yaqub al-Ayyubi* (Yaqub the Ayyubid).

Why Are Arabic Names Difficult to Match?

Finding matches for a given Arabic Name is a computationally difficult problem:

- In Arabic cultures, the same person can be identified by different names depending on the social situation.

- Some name components are optional (i.e. their presence does not add semantic value to a name), whereas others uniquely identify a person:
al-Mughrabi and *Mughrabi* can denote the same person
Abusalam and *Abdusalam* cannot be the same person
- Transliterated Arabic names are often broken up into first, middle, and last name parts to fit into data structures designed for Western names. Often, this split is done arbitrarily and/or incorrectly:

<u>First Name</u>	<u>Middle</u>	<u>Last Name</u>
<i>Abdul</i>	<i>Rahman</i>	<i>Aldakhil</i>
<i>Abd</i>	<i>Ulahman</i>	<i>Al Dakhil</i>
<i>Abd-Ulahman</i>	<i>Al</i>	<i>Dakhil</i>

- In this example, neither record will match if the first, middle, and last names are only compared against each other.
- Differences in regional Arabic dialects result in many variations for a name in the source language.
- Different transliteration rules for different target languages (e.g. English vs. French) create further variations:

Shaheen (E) vs. Chahine (F)

Kareem (E) vs. Carime (F)

- Contractions are very common in Arabic, which means that certain names can be spelled as one, two, or three words in English:

Abd-al-Rachman vs. Abd-Urrachman vs. Abdurrachman

Sal-ad-Din vs. Sal-Addin vs. Saladin

For these and other reasons, transliterated Arabic names may look similar, but in fact might be different (false positives). Conversely, they may look very different but, in fact, be the same (false negatives):

Al Salim Hasanawi vs. Al Salam Hasanawi (different)

Mohammad bin Ahmad Omarawi vs. Mohd Alomarawie (same)

Sayih El Sayed Bouchaib vs. Saa Abd el Esem Assaid Abu Shouwayb (same)

How ANS Addresses these Complexities

ANS operates on names in the Celatro Common Name Format (CNF). The CNF treats names as contiguous entities, rather than artificially splitting them up into arbitrary components. In this way, ANS avoids ambiguities associated with artificially splitting up Arabic names into the western First-Middle-Last name format. Additional simplification of the User Interface (UI) is also possible, since name queries can be entered on one line, instead of forcing users to guess the structure.

As a Celatro plug-in, ANS results can also be aggregated with other search criteria to produce sophisticated composite queries. Internally, ANS benefits from many key Celatro technologies to facilitate its scoring and ranking abilities.

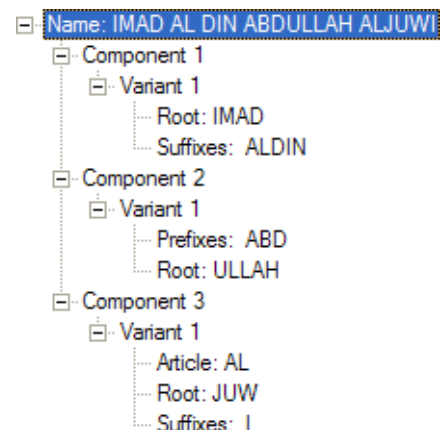
A key feature of ANS's innovative approach is the combination of an intelligent morphological parser with a comprehensive dictionary of Arabic root variations. ANS is able to recognize all name components, and correctly identify their morphological structure. Consider the name *Imad Al Din Abdullah Aljuwi*:

Features of the ANS Morphological Parser

- The ANS morphological parser correctly identifies all subcomponents, even when they are not separated from one another with white-space or hyphens:

Abdulazziz

Abu Moussa



Nur-al-Din
Saladdin
Arrachman
al-Zarqawi

- ANS recognizes all common prefixes and their spelling variations. Prefix ordering, in the case where more than one prefix precedes a name, is correctly handled in accordance with Arabic grammar (e.g. *Abu Abdurrahman Muhammad*, not **Abd Aburachman Muhammad*).
- ANS recognizes all common suffixes and their spelling variations. Difficult cases such as *al Din*, which can appear before some other suffixes, are correctly handled.
- Roots are compared against a comprehensive dictionary of Arabic names, which contains over a million variations of 16,000+ roots. When combined with prefixes, suffixes and article variations, the number of names that can be matched based on the dictionary alone is well over 68 million. Unusual roots not found in the dictionary are still correctly identified and ranked using proprietary fuzzy logic.
- ANS recognizes definite articles in all their variations, i.e. regardless whether they are in front of a sun or a moon consonant.
- ANS parses a name to an abstract, semantic representation, e.g., name elements serving the same morphological role produce the same parse. This allows for correct matching of names that are spelled differently, or correctly disambiguating names that are different but look the same:

Bouchaib vs. Abu Shouwayb

Abou Orbie vs. Belarb

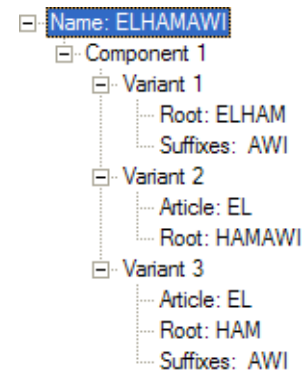
- ANS can correctly match partial names. For example, ANS correctly matches *Ahmed Ghailani* to:

Ahmad Ghailan

Ahmad Khalfan Gilani

Abubakary Khalfan Ahmed Ghailani

- When more than one parse is possible, ANS identifies all variants. Multiple parses ensure that a match will be attempted for every possible root in both the query and the search text. For example, *ibn Mouaweyah* has only one valid parse (prefix + root + suffix). *Ibn Muawiyah*, on the other hand, has two: prefix + root, and prefix + root + suffix. Had the parser stopped after the first parse of *ibn Muawiyah*, the parse that matches *ibn Mouaweyah* would not have been found and a perfectly valid match would have been ignored.



ANS Advantages: Rapid, Morphologically and Semantically Derived Results

ANS achieves exceptional precision and recall through the use of an innovative, high-performance morphological parser with an integrated semantic component. This architecture yields several key advantages over other name matching technologies:

- ANS directly compares two names and generates a similarity score. In the case where a single name is compared against a list of candidates, the parse for the first name is preserved and re-used for each candidate name. This technique yields results significantly faster than approaches which generate large permutation lists for the first name that are then compared one-by-one to the second name(s).
- Rather than simply generating a list of "hits", ANS produces *ranked* results. This allows for
 - Result ordering
 - Sophisticated combinatorics using fuzzy or Bayesian logic
 - User-specification of threshold limits
- As a Celatro plug-in, ANS leverages all the capabilities of the Celatro technology line, including²:
 - Handling of structured and unstructured text
 - Works correctly with both ASCII and Unicode data
 - Highly tuned algorithms
 - Component based architecture
 - Algorithm aggregation

-
- 1 Adapted from "Period Arabic Names and Naming Practices" by Da'ud ibn Auda (David B. Appleton) © 2003; online at <http://www.sca.org/heraldry/laurel/names/arabic-naming2.htm>.
 - 2 Visit <http://www.celatro.com> for more details about how Celatro™ technology can improve your applications.

