Database of Arab Names
قاعدة بيانات الأسماء العربيّة

Announcing v3.0 with 6.5 Million Entries

by Jack Halpern (jack@cjki.org**)**

Introduction

The number of personal name variants in Arabic is very large. **The CJK Dictionary Institute** (CJKI) has been using various techniques to collect and detect named entities and process name variants, some of which are described in my paper, "The Role of Lexical Resources in CJK NLP Applications" (.pdf). The number of applications that require the ability to process Arabic names is rapidly increasing in various sectors of the IT and security industries. To meet these needs, CJKI is maintaining the world's largest database of Arab personal name and their variants, referred to as the Database of Arab Names (DAN).

Database of Arab Names (DAN) v3.0

CJKI's Database of Arab Names (DAN) is a comprehensive database of Arab names and their romanized variants. It contains a large and constantly growing collection of romanized Arabic names mapped to the original Arabic script. DAN continues to undergo extensive proofreading by our team of native speaker editors on the basis of over 25 million romanized variants derived from a large variety of sources, including websites, corpora, books, phone directories, rule-based generation, dictionaries and encyclopedias.

In addition to its comprehensive coverage, DAN offers such unique features as *every Arabic name is manually vocalized* by our team of native Arab editors and *all romanized variants are validated for their frequency of occurrence* on the web. The database contains web frequency statistics for each of the millions of variants, adding great value for security applications.

For a more technical discussion about DAN and the linguistic issues related to Arabic names, please see the following paper which was presented at the 2nd International Conference on Arabic Language Resources and Tools in Cairo: "Lexicon-Driven Approach to the Recognition of Arabic Named Entities" (.pdf).

Key Features of DAN v3.0

- Contains over seven million validated romanized name variants.
- Constantly expanding based on raw data consisting of more than 25,000,000 variants.
- Proofread by native editors trained in Arabic phonology.
- Validated against the web and corpora.
- Fully vocalized Arabic names mapped to their Arabic spelling variants.
- Web-based frequency statistics for each name variant.
- Supports various romanization systems, such as the official IC standard.
- Fully supports OFAC names, their official aliases and unofficial variants.
- Various attributes such as TYPE (surname, etc.) and GENDER.

Practical Applications

Processing and normalizing names and their numerous variants is useful in a variety of real world applications, such as:

- Security applications, such as criminal watch lists and retail fraud.
- Cyber security applications, such as for preventing identity theft.
- Law enforcement applications including most-wanted and deportation lists.
- Border Security and immigration control applications.
- Information Retrieval, such as query processing by search engines.
- Named Entity Recognition and information extraction.
- Machine Translation, such as transcribing unknown proper nouns.
- Anti-money laundering and fraud detection by financial institutions.

Since this database includes a large number of spelling variants (some names have over 1,000 variants), it is of special interest to security agencies and security applications such as anti-money laundering, making it easy to find names spelled in a multitude of ways (e.g., *Usama bin Ladin, Osama ben Laden, Osama bin Laden,* etc.). It is interesting to note that the "underwear bomber" Umar Farouk Abdulmutallab might have been caught with the help of DAN. This .pdf file shows sixty variants of that name that could have been identified with the help of DAN.

Specifications

DAN data is normally provided as a plain text file encoded in UTF-8 with fields delimited by tabs, or in any other encoding and format by request. The format, data fields and other linguistic and technical specifications are determined on the basis of the customer's needs. That is, we do not provide an off-the-shelf data package, but customize the database and fine tune it to customers' specific requirements or applications.

Below is a brief description of the data fields that can be provided in the DAN database.

Other fields not shown here, including "Name Probability" (the probability of a name being a general vocabulary item -- for more information please see this .pdf) can also be provided.

No.	Field	Description
1	ID	A unique ID for each Arabic name in Arabic script.
2	SUBID	A code that together with the ID uniquely identifies the Arabic + romanized name combination.
3	HEADWORD	One of the romanized versions of an Arabic personal name.
4	ARABIC	Unvocalized Arabic name in UTF-8.
5	OFAC	Flag indicating whether the name (name element) is in present in OFAC. [O] Name appears in OFAC as is [V] Variant of name that appears in OFAC
6	ICS	Transcription following the widely-accepted Intelligence Community Standard romanization.
7	BUCKWALTER	Transliteration using the Buckwalter system.
8	TYPE	A code describing the kind of name (often unavailable and not always reliable):

		[SN] surname
		[GN] given name of unknown gender
		[B] both surname and given name
		[U] name of unknown type
9	**GENDER**	A code indicating the gender of the name (often unavailable and not always reliable). [M] male given name [F] female given name [B] both male and female given name [U] name of unknown gender [NA] not applicable if surname or gender unknown.
10	**FREQUENCY**	Number of web occurrences for the headword as a string (not 100% reliable because it might be a word in some other language by coincidence, especially for short headwords). [nnnnnnn] Web frequency [U] Unknown

Maintenance and Upgrades

1. **Minor Upgrades.** CJKI will provide free minor (corrections, small expansion) upgrades when they become available, or fix errors and make minor additions, free of charge.

2. **Major upgrades.** Major upgrades, such as a significant expansion of entries, need to be negotiated separately.

3. **OFAC Updates.** CJKI can provide regular updates to reflect changes and additions to OFAC and other watch lists like the UN list.

4. **Merge New Names.** Another aspect of maintenance is the merging of new names or name variants not in DAN. That is, if your company sends us Arabic names not found in DAN nor in our updates, we will merge those names into our data provide you with an update of DAN that includes those names in the format of your choice.

Arabic Place Names

CJKI can provide a database of Arabic place names and their variants, with focus on selected Middle East countries including Iran, Syria, Iraq, Afghanistan and Palestine. (Please see this sample.)

Support and Consulting

CJKI will provide, free of charge, technical and linguistic support for the specific data modules and services that your company licenses from CJKI.

CJKI also provides technical and linguistic support related to the development and fine tuning of systems that incorporate our Arabic data. This could include providing information and reports on linguistic and technical issues, especially as they relate to Arabic information processing and linguistic issues related to the Arabic script and Arabic name variation.

Business Model

We are quite flexible in all matters related to the business model and licensing fees.

These are determined on a case-by-case basis depending on the customer's specific needs and budget. Though our fees are quite reasonable, they can be significantly lowered for internal (rather than commercial) use. Typically, our contracts normally include the following basic terms, all of course subject to negotiation.

1. Non-exclusive worldwide license
2. Free minor (corrections or small expansion) upgrades when available
3. Major upgrades (significant expansion) need to be negotiated separately
4. Various kinds of maintenance, such as OFAC updates
5. Reasonable free technical and linguistic support
6. The data may be used for commercial applications.
7. The data may not be relicensed to third parties

Data Sample

The first six fields described in Section 3 above are shown below for the variants of the popular name عبدالرحيم *Abd Al Raheem*. (A complete sample of the over 1,000 variants for "Abd Al Raheem" is also available.) A sample of other fields is available upon request.

ID	SUBID	HEADWORD	ARABIC	BUCKWALTER	WEB_FREQ
V000107	U000001	'Abad Al Rahim	عبدالرحيم	EbdAlrHym	0000000002

The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001 JAPAN
Phone: 048-473-3508   FAX: 048-486-5032
E-mail: jack@cjki.org   Web: www.cjk.org

V000107	U000002	'Abad Al-Rahim	عبدالرحيم	EbdAlrHym	0000000002
V000107	U000003	'Abad-Al Rahim	عبدالرحيم	EbdAlrHym	0000000002
V000107	U000004	'Abad-Al-Rahim	عبدالرحيم	EbdAlrHym	0000000002
V000107	U000005	'Abd A Rahim	عبدالرحيم	EbdAlrHym	0000000002
V000107	U000006	'Abd A-Rahim	عبدالرحيم	EbdAlrHym	0000000002
V000107	U000007	'Abd Al Raheem	عبدالرحيم	EbdAlrHym	0000000093
V000107	U000008	'Abd Al Raheim	عبدالرحيم	EbdAlrHym	0000000002
V000107	U000009	'Abd Al Rahiem	عبدالرحيم	EbdAlrHym	0000000001
V000107	U000010	'Abd Al Rahim	عبدالرحيم	EbdAlrHym	0000036000
V000107	U000011	'Abd Al Rakhim	عبدالرحيم	EbdAlrHym	0000000114
V000107	U000012	'Abd Al Reheem	عبدالرحيم	EbdAlrHym	0000000002
V000107	U000013	'Abd Al Rehem	عبدالرحيم	EbdAlrHym	0000000009
V000107	U000014	'Abd Al Rehiem	عبدالرحيم	EbdAlrHym	0000000001
V000107	U000015	'Abd Al Rehim	عبدالرحيم	EbdAlrHym	0000000002
V000107	U000016	'Abd Al-Raheem	عبدالرحيم	EbdAlrHym	0000000093
V000107	U000017	'Abd Al-Raheim	عبدالرحيم	EbdAlrHym	0000000002
V000107	U000018	'Abd Al-Rahiem	عبدالرحيم	EbdAlrHym	0000000001