

# PDSURNAME LAYOUT AND DATA DEFINITIONS

Below are the complete layout specifications and data definitions of all files provided with *pdSurname*.

Each line below contains the following information: **FIELD NUMBER**: field position number. **FIELD NAME**: name of field. **FIELD TYPE**: field data type; “Chr” = alpha/numeric characters, “Num” = numbers. **FIELD LENGTH**: length of field. **DECIMAL PLACES**: number of decimal places (if any). **START POSITION**: field starting position. **END POSITION**: field ending position. **DESCRIPTION**: data definition of field contents.

## LAYOUT OF PDSURNAME NAMES DATABASE

Field Count: 13

Total Length: 418

Record Count: Pro and Standard: 336,955

FIELD NUMBER	FIELD NAME	FIELD TYPE	FIELD LENGTH	DECIMAL PLACES	START POSITION	END POSITION	DESCRIPTION
1	PEACOCK_ID	Chr	16		1	16	Unique identifier for each record
2	NORMAL	Chr	35		17	51	Normalized surname spelling
3	STANDARD	Chr	35		52	86	Standardized surname spelling
4	NAME	Num	35		87	121	Stylized surname spelling
5	WHITE	Num	6	2	122	127	Percent white (not Hispanic/Latino)
6	BLACK	Num	6	2	128	133	Percent black (not Hispanic/Latino)
7	HISPANIC	Num	6	2	134	139	Percent Hispanic/Latino
8	ASIAN	Num	6	2	140	145	Percent Asian/Pacific (not Hispanic/Latino)
9	NATIVE	Num	6	2	146	151	Percent Native American/Alaskan (not Hispanic/Latino)
10	MULTIRACE	Num	6	2	152	157	Percent multiracial
11	RACECONF	Chr	1		158	158	Race confidence flag: 1 = Tabulation (not an estimate) 2 = Estimate—high confidence 3 = Estimate—medium confidence 4 = Estimate—low confidence
12	RANK	Num	6	0	159	164	Name rank from 2000 U.S. Census
13	LANGUAGE	Chr	254		165	418	Language or languages of origin and use

## LAYOUT OF PDSURNAME RELATIONSHIP FILE

Field Count: 11

Total Length: 80

Record Count: Pro: 109,595,846; Standard: 80,742,846

FIELD NUMBER	FIELD NAME	FIELD TYPE	FIELD LENGTH	DECIMAL PLACES	START POSITION	END POSITION	DESCRIPTION
1	NAME1	Chr	35		1	35	First surname in the related name pair
2	NAME2	Chr	35		36	70	Second surname in the related name pair
3	REL	Chr	1		71	71	<b>Relationship:</b> 1 = Close onomastic variant 2 = Near onomastic variant 3 = Distant onomastic variant P = Phonetic match F = Fuzzy logic match (Pro only) Blank = Only matched open source algorithms; possibly a false match
4	SCORE	Chr	2		72	73	<b>Match quality score:</b> 01 (best) to 99 score; fuzzy logic matches receive a score of 00
5	DMP	Chr	1		74	74	<b>Double Metaphore:</b> P = Primary line match S = Secondary line match
6	MP	Chr	1		75	75	<b>Metaphone</b> M = match
7	NY	Chr	1		76	76	<b>New York State Identification and Intelligence System (NYSIIS):</b> N = match
8	CP	Chr	1		77	77	<b>Caverphone:</b> C = match
9	SX	Chr	1		78	78	<b>Soundex:</b> S = match
10	DMSX	Chr	1		79	79	<b>Daitch–Mokotoff Soundex:</b> P = Primary line match S = Secondary line match
11	ABC	Chr	1		80	80	<b>Alphabetical order flag:</b> A = Name pair is in alphabetical order (or, in the Pro edition only fuzzy logic file, fuzzy name first and correct spelling second) R = Names pair is in reverse alphabetical order (or, in the Pro edition only fuzzy logic file, fuzzy name first and correct spelling second)

## IMPORTING DATA INTO YOUR SYSTEM

*pdSurname* is designed to be compatible with any database system. It comes in multiple file formats, uses only the ANSI character set, and has a well-defined layout.

### FILE FORMATS

The database is available in three common file formats. Each format contains the same data.

Available file formats are:

#### CSV-COMMA SEPARATED VALUES

---

Files in Comma Separated Values (CSV) format (also known as Comma Delimited) separate fields with commas, and alpha/numeric character fields are usually delimited with double quotes (in case some of the field content includes commas). This format is the most commonly used. It is a native format for Microsoft Excel and is compatible with nearly all database management systems and spreadsheets.

#### TXT-FIXED LENGTH

---

Files in Fixed Length (TXT) format (also known as Standard Data Format or SDF) use constant field positions and lengths for all records. In other words, each field starts and ends at the same place in the text file and each record is on a separate line. While not as popular as comma separated values, this format is preferred by many due to its input precision and is widely used to transfer data between different software programs. It is compatible with most database management systems and spreadsheets.

#### DBF-DATABASE

---

Files in DBF database format (also known as xBase) are native to Microsoft FoxPro and Visual FoxPro, dataBased Intelligence dBase, Alaska Software XBase++, Apollo Database Engine, Apycom Software DBFView, Astersoft DBF Manager, DS-Datasoft Visual DBU, Elsoft DBF Commander, GrafX Software Clipper and Vulcan.NET, Multisoft FlagShip, Recital Software Recital, Software Perspectives Cule.Net, and xHarbour.com xHarbour. They are also compatible with any database management system that can import the DBF (xBase) format, such as Microsoft Access, Microsoft SQL Server, and numerous others.

### CHARACTER SET

The ANSI character set is utilized for all database records. This includes ASCII values 0 to 127 and extended values 128 to 255. These are also known as the extended Latin alphabet. Some users may need to configure their database system to import the extended values. In many cases the option will be labeled the "Latin-1" character set.

## DATABASE VERSION NUMBER

Depending on the file format, the version number of each copy of *pdSurname* is written in the first or second row of the first or second column of all database files in **X.X.X** format. The first number is the main version number of the release. The number after the first dot is the update for the version indicated. The number after the second dot references a minor revision.

## COPYRIGHT NOTICE

*pdSurname* is Copyright © 2015 Peacock Data, Inc. All Right Reserved.