Version

2.0

pdGender User Guide

Name and Gender Coding Database

An easy-to-use, comprehensive, and up-to-date database with twenty gender coding fields filtered for languages, rare unisex usage by one gender, archaic names, and diminutives. Students, teachers, and researchers benefit as well because this software is fully suitable for extensive study in the fields of anthroponymy, onomatology, ethnology, linguistics, and related fields.

Peacock Data, Inc.
California, USA 800-609-9231
Web address: www.peacockdata.com

TABLE OF CONTENTS

INTRODUCTION

Male and female identification is essential for businesses and organizations. It allows you to send mail with a personal touch. Gender Coding also allows you to filter, map, and analyze your data based on this critical demographic. **pdGender** lets you accomplish this in ways not before possible on this scale.

A one-of-a-kind proprietary resource developed and tested in the field over more than 20 years, this package contains a large set of English, Spanish, and international first names and nicknames covering more than 200 languages along with a host of additional features. The full system even incorporates sophisticated fuzzy logic.

But what makes this gender coding database truly different are twenty gender coding fields filtered for languages, rare unisex usage by one gender, archaic names, and diminutives. When a name is one gender in Khmer and another in English, users can have the English identification applied or the international usage. When a unisex name like Kimberly or Hillary is called up, users can have the much more common feminine form applied or the generic usage.

In addition to its value for businesses and organizations working with lists of names, this product is also fully suitable for students, teachers, and researchers working in the fields of anthroponymy, onomatology, ethnology, linguistics, and related areas.

pdGender is available in **Pro** and **Standard** editions. This guide covers both versions.

PRO EDITION

The *Pro* edition includes over 140,000 gender coding records with name type, origin, and languages of use, plus a sophisticated system of fuzzy logic allowing matches when there are typographical errors or stylized spelling methods are utilized.

STANDARD EDITION

The *Standard* edition includes over 60,000 gender coding records and has all features of the *Pro* version except the fuzzy logic records. However, the database is designed so users can add fuzzy logic to their system at a future time.

QUICK START

While this software has many useful fields of information, two of the most important are NAME and WORLD, and users can effectively gender code with just these fields.

They show each name and the generic gender associated with the name. Users can match against the first names on their lists to determine the gender of the individuals. For example, if you locate “Margaret” in the NAME field, it will be labeled “F” for female in the WORLD gender field. If you locate “Thomas” in the NAME field, it will be labeled “M” for male in the WORLD gender field.

On the other hand, if you locate “Kimberly” in the NAME field, it will be labeled “U” for unisex in the WORLD gender field because it can be used by both genders (in the United States about 0.25% of those named Kimberly are male).

However, with this software it is possible to more specifically gender code individuals by using other provided fields filtered for languages, rare unisex usage by one gender, archaic names, and diminutives. For example, if you want to exclude archaic names and rare unisex uses, you can gender code using the WORLD_XAR field. In this field the name “Kimberly” is labeled “F” for female because it filters out the rare male usage.

Other filtered gender coding fields are available which allow even greater precision. For example, USA_XAR filters similarly for names common in the United States and LAT_XAR filters similarly for Latino names. There are a total of twenty filtered gender coding fields to choose from.

The following are examples of gender coding, including result that differ depending on the filter utilized:

	Name	**WORLD**	**WORLD_XAR**	**USA_XAR**	**LAT_XAR**
Example 1	Margaret	F	F	F	F
Example 2	Thomas	M	M	M	M
Example 3	Kimberly	U	F	F	F
Example 4	Farley	U	M	M	M
Example 5	Karen	U	U	F	F
Example 6	Ariel	U	U	F	M

If users license the *Pro* edition of this software, or have updated a *Standard* version with fuzzy logic add-ons or upgrades, additional fuzzy logic technology allows matching first name and nickname data that has typographical errors or utilizes stylized spelling methods.

The following are some of the same examples utilizing fuzzy logic to match typographical errors:

	Name	**WORLD**	**WORLD_XAR**	**USA_XAR**	**LAT_XAR**
Example 7	Margraet	F	F	F	F
Example 8	Kimbrely	U	F	F	F
Example 9	Areil	U	U	F	M

This quick start explanation demonstrates the basic use the software. Many will only use the NAME and WORLD fields in their matches, or only a few of the available filters, but much more is also available. Read on about features never before available on this scale.

IMPORTING DATA INTO YOUR SYSTEM

pdGender 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.

INCLUDED DATABASE FILES

pdGender has four files, a main database and three related lookup tables.

Included files are:

MAIN FILE

The main file contains most of the provided information. Each records has a name along with gender; information about the name type; the relationship of the name in the *pdNickname* database; origin of the name; languages of use; a set of twenty gender coding fields filtered for languages, rare usage by one gender, and other criteria; and additional useful information.

ORIGIN LOOKUP FILE

This file provides the origin of the name. The OID field in the lookup table relates to the ORIGIN field in the main file. This file also contains additional information about unique name origins.

USAGE LOOKUP FILE

Similar to the origin file, a second file with the languages of use for the name is also provided. The UID field in the lookup table relates to the USAGE field in the main file. This file also contains additional information about use of the name in the Bible, theology, literature, and mythology.

REALNAMES LOOKUP FILE

This file is used to facilitate updating the database with fuzzy logic add-on and upgrade packs. The PEACOCK_ID field in the lookup table relates to the PEACOCK_ID field in fuzzy logic add-ons and upgrades.

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 needs to configure their database system to import the extended values. In many cases the option will be labeled the “Latin-1” character set.

FILE LAYOUTS AND DATA DEFINITIONS

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

Each line below contains the following information: FIELD NUMBER: field position number. FIELD NAME: name of field. FIELD LENGTH: length of field. START POSITION: field starting position. END POSITION: field ending position. DESCRIPTION: data definition of field contents. All fields are alpha/numeric.

LAYOUT OF PDGENDER (MAIN FILE)

Field Count: 53

Total Length: 162

Record Count: Pro: 141,803; Standard: 60,166

FIELD NUMBER	FIELD NAME	FIELD LENGTH	START POSITION	END POSITION	DESCRIPTION
1	**PEACOCK_ID**	9	1	9	**Unique identifier for each record**
2	**ORIGIN**	5	10	14	**Origin identification number:** *Relates to the OID field in the origin lookup table*
4	**TYPE**	15	15	29	**Name type:** *Base Name* *Variation* *Short Form* *Diminutive* *Feminine Form* *Masculine Form*
6	**GENDER**	1	30	30	**Gender:** *M = Male* *F = Female*
5	**NAME**	30	31	60	**Name**
6	**RELATION**	20	61	80	**Relationship in the *pdNickname* database:** *Transcription* *Variation* *Short Form* *Diminutive* *Feminine Form* *Masculine Form* ***NOTE:*** *Ties directly to the pdNickname RELATION field*
7	**FUZZY**	1	81	81	**Fuzzy flag:** *1 = Name is fuzzy*
8	**WORLD**	1	82	82	**International list gender without filters:** *M = Male* *F = Female* *U = Unisex*
9	**WORLD_XA**	1	83	83	**International list gender filtering archaic names:** *M = Male* *F = Female* *U = Unisex*

10	**WORLD_XAR**	1	84	84	**International list gender filtering archaic names and rare unisex usages:** *M = Male* *F = Female* *U = Unisex*
11	**WORLD_XARD**	1	85	85	**International list gender filtering archaic names, rare unisex usages, and diminutives:** *M = Male* *F = Female* *U = Unisex*
12	**USA_XA**	1	86	86	**American list gender filtering archaic names:** *M = Male* *F = Female* *U = Unisex*
13	**USA_XAR**	1	87	87	**American list gender filtering archaic names and rare unisex usages:** *M = Male* *F = Female* *U = Unisex*
14	**USA_XARD**	1	88	88	**American list gender filtering archaic names, rare unisex usages, and diminutives:** *M = Male* *F = Female* *U = Unisex*
15	**ENG_XA**	1	89	89	**English-dominated list gender filtering archaic names:** *M = Male* *F = Female* *U = Unisex*
16	**ENG_XAR**	1	90	90	**English-dominated list gender filtering archaic names and rare unisex usages:** *M = Male* *F = Female* *U = Unisex*
17	**ENG_XARD**	1	91	91	**English-dominated list gender filtering archaic names, rare unisex usages, and diminutives:** *M = Male* *F = Female* *U = Unisex*
18	**ENG_XAV**	1	92	92	**English-dominated list gender filtering archaic names and very rare unisex usages:** *M = Male* *F = Female* *U = Unisex*
19	**LAT_XA**	1	93	93	**Latino-dominated list gender filtering archaic names:** *M = Male* *F = Female* *U = Unisex*
20	**LAT_XAR**	1	94	94	**Latino-dominated list gender filtering archaic names and rare unisex usages:** *M = Male* *F = Female* *U = Unisex*

21	**LAT_XAD**	1	95	95	**Latino-dominated list gender filtering archaic names, rare unisex usages, and diminutives:** *M = Male* *F = Female* *U = Unisex*
22	**FR_EN_XA**	1	96	96	**French and English-dominated list gender filtering archaic names (French receives priority over English):** *M = Male* *F = Female* *U = Unisex*
23	**FR_EN_XAR**	1	97	97	**French and English-dominated list gender filtering archaic names and rare unisex usages (French receives priority over English):** *M = Male* *F = Female* *U = Unisex*
24	**FR_EN_XAD**	1	98	98	**French and English-dominated list gender filtering archaic names, rare unisex usages, and diminutives (French receives priority over English):** *M = Male* *F = Female* *U = Unisex*
25	**EN_FR_XA**	1	99	99	**English and French-dominated list gender filtering archaic names (English receives priority over French):** *M = Male* *F = Female* *U = Unisex*
26	**EN_FR_XAR**	1	100	100	**English and French-dominated list gender filtering archaic names and rare unisex usages (English receives priority over French):** *M = Male* *F = Female* *U = Unisex*
27	**EN_FR_XAD**	1	101	101	**English and French-dominated list gender filtering archaic names, rare unisex usages, and diminutives (English receives priority over French):** *M = Male* *F = Female* *U = Unisex*
28	**LANGFLAG**	1	102	102	**Language flag:** *1 = At least one language field is filled* *A = Archaic* *Blank = Name is used in other languages* ***NOTE:** see the usage lookup table for other languages*
29	**USAGE**	5	103	107	**Usage identification number:** *Relates to the UID field in the usage lookup table*
30	**BIBLE**	1	108	108	**Biblical and/or theological name:** *B = Biblical* *T = Theological* *R = Biblical and Theological*
31	**ENGLISH**	1	109	109	**Name is used in the English language:** *E = English* *e = English – rare usage* *V = English – very rare usage* *A = Archaic*

32	**AFRAM**	1	110	110	**Name is an African American name:** *E = African American* *e = African American – rare usage* *A = Archaic*
33	**NATAM**	1	111	111	**Name is a Native American or Hawaiian name:** *N = Native American* *n = Native American – rare usage* *H = Hawaiian* *h = Hawaiian – rare usage* *A = Archaic*
34	**SPANISH**	1	112	112	**Name is used in the Spanish language:** *S = Spanish* *s = Spanish – rare usage* *A = Archaic*
35	**BASQUE**	1	113	113	**Name is used in the Basque language:** *B = Basque* *b = Basque – rare usage* *A = Archaic*
36	**CATALAN**	1	114	114	**Name is used in the Catalan language:** *C = Catalan* *c = Catalan – rare usage* *A = Archaic*
37	**GALICIAN**	1	115	115	**Name is used in the Galician language:** *G = Galician* *g = Galician – rare usage* *A = Archaic*
38	**FRENCH**	1	116	116	**Name is used in the French language:** *F = French* *f = French – rare usage* *N = Norman French* *n = Norman French – rare usage* *O = Occitan* *o = Occitan – rare usage* *P = French Provençal* *p = French Provençal – rare usage* *A = Archaic*
39	**GERMAN**	1	117	117	**Name is used in the German language:** *G = German* *g = German – rare usage* *S = Swiss German* *s = Swiss German – rare usage* *A = Archaic*
40	**HINDU**	1	118	118	**Name is used in the Hindustani language:** *H = Hindi* *h = Hindi – rare usage* *U = Urdu* *u = Urdu – rare usage* *A = Archaic*
41	**RUSSIAN**	1	119	119	**Name is used in the Russian language:** *R = Russian* *r = Russian – rare usage* *A = Archaic*

42	**PERSIAN**	1	120	120	**Name is used in the Persian language:** *P = Persian* *p = Persian – rare usage* *A = Archaic*
43	**ARABIC**	1	121	121	**Name is used in the Arabic language:** *M = Arabic* *m = Arabic – rare usage* *A = Archaic*
44	**JAPANESE**	1	122	122	**Name is used in the Japanese language:** *J = Japanese* *j = Japanese – rare usage* *A = Archaic*
45	**CHINESE**	1	123	123	**Name is used in the Chinese language:** *C = Chinese* *c = Chinese – rare usage* *A = Archaic*
46	**VIET**	1	124	124	**Name is used in the Vietnamese language:** *V = Vietnamese* *v = Vietnamese – rare usage* *A = Archaic*
47	**KOREAN**	1	125	125	**Name is used in the Korean language:** *K = Korean* *k = Korean – rare usage* *A = Archaic*
48	**YIDDISH**	1	126	126	**Name is used in the Yiddish language:** *Y = Yiddish* *y = Yiddish – rare usage* *A = Archaic*
49	**HEBREW**	1	127	127	**Name is used in the Hebrew language:** *H = Hebrew* *h = Hebrew – rare usage*
50	**LATIN**	1	128	128	**Name is used in the Latin language:** *L = Latin* *l = Latin – rare usage*
51	**GREEK**	1	129	129	**Name is used in the Greek language:** *G = Greek* *g = Greek – rare usage*
52	**MYTH**	3	130	132	**Name is used in mythology:** *A = Arthurian Legend* *E = Egyptian Mythology* *e = Egyptian Mythology (Anglicized)* *h = Egyptian Mythology (Hellenized)* *y = Egyptian Mythology (Latinized)* *G = Greek Mythology* *g = Greek Mythology (Latinized)* *I = Irish Mythology* *i = Irish Mythology (Latinized)* *J = Judeo-Christian Legend* *j = Judeo-Christian Legend (Anglicized)* *N = Norse Mythology* *R = Roman Mythology* *r = Roman Mythology (Anglicized)* ***NOTE:** See the usage lookup table for other uses in mythology*
53	**REALNAME**	30	133	162	**Real name of the fuzzy entry:** *Filled if FUZZY equals "1"*

LAYOUT OF ORIGIN (LOOKUP FILE)

Field Count: 2

Total Length: 259

Record Count: 1,263

FIELD NUMBER	FIELD NAME	FIELD LENGTH	START POSITION	END POSITION	DESCRIPTION
1	**OID**	5	1	5	**Unique identifier for each origin:** *Relates to the ORIGIN field in the main pdGender database*
2	**ORIGIN**	254	6	259	**Origin:** *Comma delimited list of languages involved in the origin of the name; also includes information about unique origins*

LAYOUT OF USAGE (LOOKUP FILE)

Field Count: 3

Total Length: 260

Record Count: 2,083

FIELD NUMBER	FIELD NAME	FIELD LENGTH	START POSITION	END POSITION	DESCRIPTION
1	**UID**	5	1	5	**Unique identifier for each usage:** *Relates to the USAGE field in the main pdGender database*
2	**USAGE**	254	6	259	**Usage:** *Comma delimited list of languages using the name; also includes biblical, theological, mythology, and literary uses*
3	**NOTINUSE**	1	260	260	**Not-in-use flag:** *X = Not used as a personal name; used only in the Bible, theology, mythology, or literature*

LAYOUT OF REALNAMES (LOOKUP FILE)

Field Count: 2

Total Length: 39

Record Count: 81,637

FIELD NUMBER	FIELD NAME	FIELD LENGTH	START POSITION	END POSITION	DESCRIPTION
1	**PEACOCK_ID**	9	1	9	**Unique identifier for each record:** *Relates to the PEACOCK_ID field in fuzzy logic add-ons and upgrades*
2	**REALNAME**	30	10	39	**Real name of the fuzzy entry:** *Filled with the real spelling of names provided in fuzzy logic add-ons and upgrades*

USING THE PDGENDER DATABASE

The main *pdGender* file is organized with one name per record. Additional information, such as gender, type of name, origin of the name, and languages of use, are provided for each name as well. Users can match against the first names on their lists to determine the gender of the individuals using special filtered gender coding fields.

PEACOCK_ID FIELD

The first field in the database is PEACOCK_ID. It provides a unique identifier for each record, but is also equipped with additional functionality. Each begins with the character "g" to identify the database. They have two parts separated with a hyphen.

The following is the first PEACOCK_ID in the database:

- **g000001-1** is a complete PEACOCK_ID; no other record has this same exact identification

PARTS OF PEACOCK_ID

The first part of the PEACOCK_ID code (the part before the hyphen) identifies each unique name. A unique name is defined by the spelling of the name, gender, type of name, origin, and usage.

The following is an example of a unique name identification number:

- **g000001** is the first part of a PEACOCK_ID; it identifies each unique name; multiple records can have this same number but each record is showing the same name in a different relationship

The second part (the part after the hyphen) is provided for compatibility with *pdNickname*. (*pdNickname* is not required to use *pdGender* but they are highly attuned to work together.) It identifies each type of relationship a unique name has in the *pdNickname* database. There are six possible relationship types: Variation, Transcription, Short Form (including Short Form Variation), Diminutive (including Diminutive Variation), Feminine Form, and Masculine Form. Names can be involved with up to five of the six relationship types; they cannot be involved with both a Feminine Form and a Masculine Form relationship.

The following is an example of unique name and relationship identification number:

- **g000001-2** has both parts of a PEACOCK_ID; it identifies each unique name and relationship type; no other record has this same exact identification

Review the section on Compatibility for further information on using *pdGender* with *pdNickname*.

NAME AND GENDER

Each record has a name in the NAME field and the associated gender in the GENDER field. Additional information, such as type of name, origin of the name, and languages of use, are provided for each name as well. Users can match against the first names on their lists to determine the gender of the individuals.

IMPORTANT

The process of coding usually utilizes the filtered gender coding fields to assign gender, and not the GENDER field.

GENDER is a special field that resolves to either male or female based on the language of use and never to a unisex name. If a name is unisex, both a male and female record is included, often with an indicator of rare usage by one gender. Because it is tied to language of use, the GENDER field can be utilized for very precise gender coding if the user knows the language of use in advance.

NAME TYPE

Each name is identified by type of name in the TYPE field. Many spellings of names serve multiple types, such as both a variation of another first name as well as usage as a short form nickname or diminutive. Opposite gender forms frequently serve both as feminine forms or masculine forms as well as variation of other same-gender names. Each formation of the name is included in the database in separate records.

Name types are:

BASE NAME

This is the oldest identified formation of a name. Many originated in Antiquity or the Middle Ages, and include Greek, Roman, Hebrew, Ancient Germanic, Old Norse, Old English, Middle English, and Old Spanish, among numerous other name origins.

VARIATION

These are spelling alterations of Base Names or other first name Variations. When the alteration arises in the same language and era, it is known as a “variant” of the name. When it arises in another language or era, such as Old English to modern English, it is known as a “form” of the name.

SHORT FORM

These are nicknames for Base Names and first name Variations. They are commonly based on the first syllable or part of a name, but not always.

DIMINUTIVE

These are also used as nicknames for Base Names and first name Variations, but can be a diminutive form of a Short Form nickname as well. They usually include some of the root spelling of the name it is associated with and are typically intended to convey more endearment than Short Form nicknames.

FEMININE FORM

These are spelling alterations of male gender Base Names and first name Variations formulated for the female gender.

MASCULINE FORM

These are the opposite of Feminine Form names and are much less common. They are spelling alterations of female gender Base Names and fist name Variations formulated for the male gender.

FILTERED GENDER CODING FIELDS

The twenty filtered gender coding fields are the heart of the *pdGender* matching system. These allow users to filter their matches for languages, rare unisex usage by one gender, archaic names, and diminutives. When a name is one gender in Khmer and another in English, users can have the English identification applied or the international usage. When a unisex name like Kimberly or Hillary is called up, users can have the much more common feminine form applied or the generic usage.

WORLD

This is the only gender coding field without filters of any kind. It can utilize like the standard unfiltered gender coding fields most users are familiar with and is similar to the implication found in previous versions of *pdGender*.

WORLD_XA, WORLD_XAR, WORLD_XARD

These include all international matches and are not filtered by language.

WORLD_XA filters only archaic names. WORLD_XAR filters both archaic names and rare unisex usages. WORLD_XARD filters archaic names, rare unisex usages, and diminutives (which are less likely to be on lists). If a match cannot be made with the filters, the unfiltered WORLD international gender is applied.

USA_XA, USA_XAR, USA_XARD

These filters are designed for American lists and lists from other English-speaking nations. It first tries to match English names, then Spanish names, and then searches other common languages.

USA_XA filters only archaic names. USA_XAR filters both archaic names and rare unisex usages. USA_XARD filters archaic names, rare unisex usages, and diminutives (which are less likely to be on lists). If a match cannot be made with the filters, the unfiltered WORLD international gender is applied.

ENG_XA, ENG_XAR, ENG_XARD, ENG_XAV

These filters are designed for English-dominated lists. It first tries to match English names and then searches other common languages.

ENG_XA filters only archaic names. ENG_XAR filters both archaic names and rare unisex usages. ENG_XARD filters archaic names, rare unisex usages, and diminutives (which are less likely to be on lists). If a match cannot be made with the filters, the unfiltered WORLD international gender is applied.

ENG_XAV

ENG_XAV is a special filter for only English names, where statistics are more plentiful, that filters archaic names and only “very” rare English unisex usages. If a match cannot be made with the filters, the unfiltered WORLD international gender is applied.

LAT_XA, LAT_XAR, LAT_XARD

These filters are designed for Latino-dominated lists. It first tries to match Spanish names, then English names, and then searches other common languages.

LAT_XA filters only archaic names. LAT_XAR filters both archaic names and rare unisex usages. LAT_XARD filters archaic names, rare unisex usages, and diminutives (which are less likely to be on lists). If a match cannot be made with the filters, the unfiltered WORLD international gender is applied.

FR_EN_XA, FR_EN_XAR, FR_EN_XARD

These filters are designed for French and English-dominated lists (French receives priority over English). It first tries to match French names, then English names, and then searches other common languages.

FR_EN_XA filters only archaic names. FR_EN _XAR filters both archaic names and rare unisex usages. FR_EN _XARD filters archaic names, rare unisex usages, and diminutives (which are less likely to be on lists). If a match cannot be made with the filters, the unfiltered WORLD international gender is applied.

EN_FR_XA, EN_FR_XAR, EN_FR_XARD

These filters are designed for English and French-dominated lists (English receives priority over French). It first tries to match English names, then French Names, and then searches other common languages.

EN_FR_XA filters only archaic names. EN_FR _XAR filters both archaic names and rare unisex usages. EN_FR _XARD filters archaic names, rare unisex usages, and diminutives (which are less likely to be on lists). If a match cannot be made with the filters, the unfiltered WORLD international gender is applied.

FUZZY LOGIC

This section applies to pdGender Pro. It also applies to pdGender Standard when fuzzy logic add-ons or upgrades are appended to the system.

The fuzzy logic technology in this software allows matching first name and nickname data that has typographical errors or utilizes stylized spelling methods. When a fuzzy logic record is provided, it is indicated in the FUZZY field and the correct spelling of the name is entered in the REALNAME field.

If users filter for records flagged in the FUZZY field, they are likely to see errors they have repeatedly made or seen. In many cases you will have to look close to see the difference, but they are different.

TYPOGRAPHICAL ERRORS

A large majority of fuzzy logic records involve common typographical errors. These algorithms look at frequently reversed digraphs (a pair of letters used to make one phoneme or distinct sound), phonetically transcribed digraphs, double letters typed as single letters, single letters that are doubled, and other common data entry issues.

The most likely typographical errors are determined based on the number of letters, the characters involved, where they are located in the name, and other factors. Sometimes, however, a less common error is provided due to filtering criteria. This is usually because one requirement is that fuzzy spellings never formulate a real name already in the database. This sometimes happens and most often because the fuzzy spelling was already a real variation of the same name.

The following are examples of fuzzy logic based on common typographical errors:

	NAME	FUZZY	REALNAME
Example 1	ALL	1	AL
Example 2	ROCO	1	ROCCO
Example 3	CHRISTOFER	1	CHRISTOPHER
Example 4	SOHPIA	1	SOPHIA
Example 5	MARGRAET	1	MARGARET

In *Example 1*, the “L” in “AL” is repeated in NAME and REALNAME shows the correct spelling.

In *Example 2*, the second “C” in “ROCCO” is left out in NAME and REALNAME shows the correct spelling.

In *Example 3*, the “PH” digraph in “CHRISTOPHER” is phonetically transcribed as “F” in NAME and REALNAME shows the correct spelling.

In *Example 4*, the “PH” digraph in “SOPHIA” is reversed in NAME and REALNAME shows the correct spelling.

In *Example 5*, the second “AR” digraph in “MARGRAET” is reversed in NAME and REALNAME shows the correct spelling.

STYLIZED SPELLINGS

Other fuzzy logic records involve stylized spelling methods. These algorithms look at non-regular characters such as extended ANSI characters (ASCII values 128 to 255) as well as hyphens, apostrophes, and spaces.

A few of the possible extended characters are “Á” (A-acute), “Ö” (O-umlaut), and “Ñ” (N-tilde). In these cases, “Á” becomes “A” (A-regular), “Ö” becomes “O” (O-regular), “Ñ” becomes “N” (N-regular), and other extended characters are treated similarly.

The following are examples of fuzzy logic with stylized spellings:

	NAME	**FUZZY**	**REALNAME**
Example 6	BJORK	1	BJÖRK
Example 7	NICOLAS	1	NICOLÁS
Example 8	ASHTORET	1	'ASHTORET
Example 9	ABDALHAMID	1	ABD-AL-HAMID
Example 10	JUANMARIA	1	JUAN MARÍA

In *Example 6*, NAME is spelled with O-regular instead of with O-umlaut and REALNAME shows the stylized spelling.

In *Example 7*, NAME is spelled with A-regular instead of with A-acute and REALNAME shows the stylized spelling.

In *Example 8*, NAME is spelled without an apostrophe prefix and REALNAME shows the stylized spelling.

In *Example 9*, NAME is spelled without hyphens delimiting the name parts and REALNAME shows the stylized spelling.

In *Example 10*, NAME is not only spelled without the space between the two parts, but I-acute is also replaced with I-regular. REALNAME shows the stylized spelling.

FUZZY LOGIC ADD-ON PACKS AND UPGRADES

Peacock Data releases additional fuzzy logic records nearly every month for *pdGender 2.x* in the form of add-on packs which can easily and economically be appended to the main database extending coverage of typographical errors and stylized spelling methods.

The fuzzy logic technology built into the main *Pro* product download is designed to pick up statistically the most likely mistakes and stylizations. *Fuzzy Logic Add-on Packs* are designed to pick up less common mistakes and stylizations.

Add-on packs include new algorithms and randomizers and are fully compatible with both the *Pro* and *Standard* editions of this package.

Those licensing the *Standard* edition can also purchase a *pdGender Standard to Pro Upgrade Pack* which includes all the fuzzy logic records from the *Pro* edition. Once a *Standard* version is upgraded, it will be the same as the *Pro* edition.

Review the documentation provided with the fuzzy logic add-on packs and upgrades for further instructions.

LANGUAGES OF USE

This is an important section of the database with many advanced uses. This information can be used in gender coding to filter matches for languages, as is done in the filtered gender coding fields. It can also be used to gather information on the possible ethnicity and heritage of those on lists. Or it can be used for many purposes not yet thought of because this information was not before available on this scale.

RARE AND VERY RARE USAGES OF UNISEX NAMES

One of the most useful features of the *pdGender* database is that rare unisex usages of names by a language are identified so they can be filtered in name matching.

Note the following about rare and very rare usage indicators:

- A rare usage indicates that there is less than about a twenty percent chance the name is that gender; the percentage is more exact for English names where statistics are more plentiful, but approximations are determined for other languages when possible
- English unisex names are also identified as “very” rare when the usage by a gender in the English is less than five percent

It is quite common to find the rare usage to be different genders in different languages, so this factor must be considered separately for each language using the name.

Indications for rare usages of non-unisex names also exist in the database. These are based on the accepted understanding of the usage and not necessarily statistics to avoid English receiving a disadvantage over languages for which fewer statistics is available.

Note that rare and very usage indicators should not be compared for different languages, only within the same language. Because a name usage is labeled rare in Spanish and not in English does not mean the name is used less in Spanish than English, rather it means it is rare in Spanish compared to other Spanish names or, if it is a unisex name, rare compared to the Spanish opposite gender usage.

USING THE USAGE LOOKUP TABLE

The USAGE field in the main database relates to the UID field in the Usage lookup table. The Usage lookup table is similar the Origin table, except it indicates the languages of use and not the origin of the name. Each record contains a comma delimited list of the languages of use.

Note the usage indicators in the lookup table:

- Languages where the name is in rare usage are indicated with an asterisk (*) after the name of the language (e.g., “English*”); for unisex names the indicator is applied to one gender only
- Some English unisex names are identified as “very” rare and have two asterisks (**) after the name of the language (i.e., “English**”); these are also applied to one gender only
- Archaic names from modern languages are identified with “(archaic)” after the name of the language (e.g., “Spanish (archaic)”)

Note if there is conflicting information about the Languages of use, usually both or all are included separately depending on the quality of the sources. This is done to leave placeholders. Future editions of this database will try to merge or otherwise distinguish these records, but for now are left as multiple possibilities. This will occur more frequently for common names because they are covered in a larger number of sources.

For those using the database for research, references to biblical, theological, literary, and mythology names are also included in many records. These special uses follow a semicolon (;) and are also comma delimited when more than one special usage exists for a record. Users uninterested in this information can delete everything from the semicolon and beyond without losing any of the language information.

USING THE MAIN FILE LANGUAGE FIELDS

Much of the language information from the Usage lookup table has been transferred to fields in the main database for easy access by users. Usually the code is the first letter of the name of the language, but be careful with Arabic. To avoid conflicting with “A” for archaic, “M” and “m” are utilized instead, for Modern Standard Arabic, which developed from the Classical Arabic of the Quran (or Qur’an, Romanized) and Islamic literature from the Early Middle Ages.

Note the usage indicators in the main file language fields:

- Upper case codes indicate common usage (e.g., “S” = common usage in Spanish)
- Lower case codes indicate rare usage (e.g., “s” = rare usage in Spanish)
- V = Very rare usage (English unisex names only)
- A = Archaic

The languages chosen to duplicate in the main file were selected because they are common in the United States, including for American Latinos, and in other English speaking nations, or represent a minority group most likely to be of particular interest.

The LANGFLAG field is set up to indicate if a name has the language presented in the main file. These names are more likely to be on American lists, Latino lists, and lists associated with the other presented languages. This field also indicates if the name is archaic.

In addition to languages, biblical names and names from major mythologies are also identified in the BIBLE and MYTH fields respectively.

ORIGIN OF NAMES

The origin of each name is also provided in a lookup table similar to the Usage lookup table. This information will be of more interest to students, teachers, and researchers working in the fields of anthroponymy, onomatology, ethnology, linguistics, and related fields. It may be of less interest to businesses and organizations working with lists of names and can be skipped by these users. For those interested it will explain the naming conventions in relation to historic periods. Be prepared for a lot of dates and historical reference points.

Note if there is conflicting information about the origin of a name, usually both or all are included depending on the quality of the sources. This is done to leave placeholders. Future editions of this database will try to merge or otherwise distinguish these records, but for now are left as multiple possibilities. This will occur more frequently for common names because they are covered in a larger number of sources.

USING THE ORIGIN LOOKUP TABLE

The ORIGIN field in the main database relates to the OID field in the Origin lookup table. Here you will find a comma delimited list of the languages the name was formulated in.

Origin information also indicates if the name is modern or from an earlier era, including Ancient names (those arising during Antiquity) and names from the Middle Ages.

ANTIQUITY

Ancient names include:

- Ancient Egyptian: attested from 3400 BC making it one of the earliest known written languages (along with Sumerian)
- Sumerian: the language of ancient Sumer, spoken in southern Mesopotamia (modern Iraq) and closely related to Akkadian, it is attested from 3350 BC making it one of the earliest known written languages (along with Egyptian); it was slowly replaced by Akkadian between the 3rd to the 2nd millennia BC but continued as a classical language until about 100 AD
- Akkadian: spoken in ancient Mesopotamia from the 29th through 8th centuries BC, including during the Akkadian Empire (ca. 2334–2193 BC), it is closely related to and replaced Sumerian, and is the earliest attested Semitic language; academic and liturgical use continued until about 100 AD
- Aramaic: a Northwest Semitic language subfamily (which includes Hebrew and Phoenician)
- Greek: spoken on the Balkan Peninsula since the 3rd millennium BC, and the oldest recorded living language, its earliest attested written evidence is the Linear B clay tablet found in Messenia which dates to between 1450 and 1350 BC
- Hittite: spoken by the Hittites, an ancient Anatolian people who established an empire at Hattusa in north-central Anatolia around 1600 BC, it is attested to about the 19th century BC and remained in use until about 1100 BC
- Hebrew: a West Semitic language, closely related to Phoenician, historically regarded as the tongue of the Israelites (meaning, “Children [or Sons] of Israel”, its earliest attested written evidence, in form of primitive drawings, dates from the 10th century BC; it was nearly extinct as a spoken language by late

Antiquity, but continued to be used as a literary language and as the liturgical language of Judaism, until its revival as a spoken language in the late 19th century

- Phoenician: a Northwest Semitic language, closely related to Hebrew, originally spoken in the ancient coastal Mediterranean region of Canaan (roughly corresponding to the Levant) and attested from the 10th until the early 4th century BC
- Ancient Macedonian: spoken during the 1st millennium BC in the ancient Kingdom of Macedonia (or Macedon) in the northeastern part of the Greek peninsula; marginalized by Hellenistic influences, it gradually fell out of use during the 4th century BC
- Roman: spoke Archaic Latin during the Roman Kingdom (753—509 BC) through most the Roman Republic (509—27 BC), replaced by Classical Latin around 75 BC; due to Roman conquests, Latin spread to many Mediterranean and some northern European regions; although considered a "dead" language, Latin is still used in the creation of new words and names in modern languages
- Proto-Germanic: dating to the Nordic Bronze Age in Scandinavia (ca. 1700–500 BC)
- Ancient Celtic: dating from the British Iron Age (ca. 600 BC—100 AD) through Antiquity
- Ancient Germanic: dating from the Pre-Roman Iron Age culture in Scandinavia, northern Germany, and the Netherlands north of the Rhine River (ca. 500—100 BC) through Antiquity

Late Greek and Late Roman names date from late Antiquity and the early Byzantine period. Late Antiquity is generally considered from the end of the Roman Empire's crisis of the 3rd century (ca. 235–284) to the re-organization of the Eastern Roman Empire under Byzantine Emperor Heraclius and the Islamic conquests during the early and mid 7th century.

Coptic Egyptian is the later stages of the Egyptian language spoken from the 2nd until the 17th century. Today Egyptians mainly speak a dialect of Modern Standard Arabic. Coptic Egyptian is still used as the liturgical language of the Coptic Church.

MIDDLE AGES

Almost all historians agree the Middle Ages began when the political structure of Western Europe changed at the end of the united Roman Empire (476 AD). In the database names dating from the Early Middle Ages (which followed the decline of the Western Roman Empire and is sometimes called the Dark Ages due to the relative scarcity of literary and cultural output during most of the era) are usually prefixed with "Old" such as Old English (which is Anglo-Saxon) and Old Spanish (which still continues as a liturgical language but with a modernized pronunciation). Galician-Portuguese is an exception, but it is also secondarily known as Old Portuguese. Note that Galician-Portuguese is a different language than modern Galician. This is also true of Ancient Macedonian and Macedonian; the latter is a modern South Slavic language.

Many languages went through significant changes during the High Middle Ages (a period of rapid population growth and social and political change in Europe from about the 11th through the 13th century) or by the Late Middle Ages (when prosperity and growth in Europe came to a halt and the population experienced a series of famines and plagues). Languages developing in this period are prefixed with "Middle", or in some cases "Medieval" depending on the accepted terminology.

After Duke William II of Normandy conquered England and killed King Harold II at the Battle of Hastings (1066), the invading Normans and their descendants replaced the Anglo-Saxons as the ruling class of England. French influences were absorbed into the English language, and Old English slowly evolved into Middle English between the 12th and 15th century, additionally aided by influences from the Latin language of the church and the invention of the printing press. Nevertheless, Old English was still used throughout the Plantagenet era (1154–1485), a few years beyond the time Constantinople was finally captured by the Ottoman Turks marking the final end of the Roman Empire (1453), the conclusion of the Middle Ages in the minds of many historians. Of course others cite the Battle of Bosworth Field which established the Tudor dynasty and an era of expansion for England (1485), the conquest of Granada and its annexation by Castile ending Islamic rule (1492), the discovery of the Americas by Christopher Columbus (also 1492), the death of Queen Isabella I of Castile (1504), the death of her spouse King Ferdinand II of Aragon (1516), and the Protestant Reformation (1517) as more appropriate cutoff points, often influenced by the nationality of the historian.

There was no similar revision during the High or Late Middle Ages in many languages, including Spanish, and they do not have a generally recognized middle variety.

Tiberian Hebrew is the canonical pronunciation of the Hebrew Bible (or Tanakh) committed to writing by Masoretic scholars living in the Jewish community of Tiberias in ancient Palestine (ca. 750-950). It is written in a form of Tiberian vocalization dating from the 8th century, but the oral tradition it reflects has ancient roots. Tiberian pronunciation of Hebrew is considered by textual scholars to be the most exact and proper pronunciation of the language as it preserves the original Semitic consonantal and vowel sounds of ancient Hebrew.

Ashkenazi Hebrew is the pronunciation system for Biblical and Mishnaic Hebrew favored for liturgical use in Ashkenazi Jewish practice in Central and Eastern Europe. Until the middle of the 20th century, most American synagogues used the Ashkenazic Hebrew pronunciation, as the majority of American Jews were of Ashkenazic descent. After the creation of the State of Israel in 1948, however, there has been a gradual shift in American congregations toward using Sephardic Hebrew because it is the standard pronunciation used in Israel.

Much is unknown about the origin of the Yiddish, a High German language written in the Hebrew alphabet, because most speakers were exterminated in the Holocaust. The consensus among scholars is it emerged among the Ashkenazi Jews in Central Europe between the 10th and 12th centuries and later spread to Eastern Europe in the 16th century.

MODERN

Language formations after the Middle Ages are usually know as modern.

There is frequently confusion about the development of the three modern strains of Gaelic: Irish, Scottish, and Manx. All three sprang from Middle Irish which came from Old Irish.

UNIQUE ORIGINS

Many records also provide additional information about unique origins. These follow a semicolon (;) and are not comma delimited, however commas may be used for punctuation. When a unique origin has more than one element, semicolons delimitate each element. All language information is provided before the first semicolon.

Unique origins include:

- Latinized, Latinate, Hellenized, and anglicized names
- Literary names created by authors, composers, and poets
- Names that became known through historical events
- Bynames: a familiar name for a person, similar to a nickname, that is often used as a replacement for a personal name—for example, Rocky is a common byname for some boxers
- Roman family names
- Roman cognomens: originally nicknames that were later utilized to augment family names to identify a particular branch within a family or family within a clan
- Roman praenomens: early personal name chosen by the parents of a Roman child originally bestowed the eighth day after the birth of a girl, or the ninth day after the birth of a boy; the praenomen would then be formally conferred a second time when girls married, or when boys reaching manhood and assumed the toga virilis (which in the case of Romans boys was about age 14 or 15)
- Occupational surnames
- Patronymic surnames
- Toponymic (habitational) surnames
- Other surnames

COMPATIBILITY

To ensure compatibility with any operating system and database platform, *pdGender* is provided in multiple file formats and utilizes only the ANSI character set (ASCII values 0 to 127 and extended values 128 to 255).

USING PDGENDER WITH PDNICKNAME

pdGender and *pdNickname* make excellent partners. They have been developed to be fully compatible and are comprised of the same set of names. For every name, gender, origin, usage, and relationship type in the *pdNickname* database, there is a corresponding record in the *pdGender* database linked by an identification number.

Note that *pdNickname* is not required to use *pdGender* but they are highly attuned to work together.

The PEACOCK_ID identification numbers in the *pdGender* database (except the first character) match the same names and PEACOCK_ID numbers in the *pdNickname* database associated with an indicated relationship type. If users search on only the identification numbers before the hyphen, they can query all records for those names in *pdNickname* regardless of relationship.

Of course the converse is true, and *pdNickname* users can look up the gender in *pdGender* using the identification numbers before the first or second hyphen (both will work equally as well).

The PEACOCK_ID identification numbers in the *pdNickname* database are longer because they have an additional sequence for each individual association with the relationship, which can be one or hundreds.

The names in *pdGender Pro* are in *pdNickname Pro*, and the names in *pdGender Standard* are in *pdNickname Standard*. *pdGender* contains one record for each relationship type names have in *pdNickname,* and records are repeated when multiple relationships exists. The type of relationship is provided in the RELATION field of both databases.

All relationships are with names of the same gender except Feminine Forms and Masculine Forms which are relationships with names of the opposite gender.

Relationships are:

VARIATION

This relationship occurs between Base Names and Variation-type names or between two Variation-type names.

TRANSCRIPTION

Similar to a Variation relationship, this occurs between transcriptive variations paired with Base Names or Variation-type names. Transcriptions are variations spelled phonetically as they sound to the person hearing and transcribing the name.

SHORT FORM

This relationship occurs between Short Form-type nicknames paired with Base Names or Variation-type names.

Note that when selecting Short Form relationships, they will also include any Short Form Variations (a relationship that occurs between two Short Form-type nicknames of the same Base Name or first name Variation-type name), unless the query is otherwise filtered to select only the desired formations.

DIMINUTIVE

This relationship occurs between Diminutive-type nicknames paired with Base Names, Variation-type names, or Short Form-type nicknames.

Note that when selecting Diminutive relationships, they will also include any Diminutive Variations (a relationship that occurs between two Diminutive-type nicknames of the same Base Name or Variation-type name), unless the query is otherwise filtered to select only the desired formations.

FEMININE FORM

This opposite gender variation relationship occurs with male gender Base Names and Variation-type names.

MASCULINE FORM

The reverse of a Feminine Form relationship, this opposite gender variation relationship occurs with female gender Base Names and Variation-type names.

USER GUIDE UPDATES

User guides are updated based on information gained from user experience. It is suggested that users regularly check the Support section of the Peacock Data website for updates. Look for a date newer than the date below:

The publication date of this guide is: May 1, 2014.

DATABASE VERSION NUMBER

Depending on the file format, the version number of each copy of *pdGender* 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.

SITE LICENSE

Peacock Data's site licenses are designed to be fair. They are broader than most software licenses in that they allow installation on not one but all computers in the same building within a single company or organization. We ask users to honor these simple rules so Peacock Data can continue bringing great products to users.

THE USE OF *PDGENDER* IS GOVERNED BY THE FOLLOWING SITE LICENSE

I. This Site License grants to the Licensee the right to install the licensed version of **pdGender** including licensed add-on packs and upgrade packs (hereinafter, 'information') on all computers in the same building within a single company or organization. Separate Site Licenses must be purchased for each building the information is used in.

II. The information may only be used by the employees of the Licensee. If the Licensee is an educational institution, the data may only be used by enrolled students, faculty, teaching assistants, and administrators.

III. Temporary employees, contractors, and consultants of the Licensee who work on-site at the Licensee's facility may also use the information in connection with the operation of the business of the Licensee. Any copies of the information used by temporary employees, contractors, and consultants must be removed from such individual's computers once they cease working at the Licensee's facility.

IV. The information cannot be used to provide services or products to customers or other third parties, whether for-profit or given away. A Developer License must be purchased separately by the Licensee to incorporate the information in for-profit services and products.

V. The Licensee is required to use commercially reasonable efforts to protect the information and restrict network or any other access to the information by anyone inside or outside of the Licensee's facility who is not authorized to use the information.

VI. The Licensee owns the media on which the information is recorded or fixed, but the Licensee acknowledges that Peacock Data, Inc. and its licensors retain ownership of the information itself.

VII. The Licensee may not transfer or assign its rights under this license to another party without Peacock Data, Inc.'s prior written consent.

VIII. Peacock Data, Inc. may revoke the rights granted by this license upon a violation of any provision herein by the Licensee.

IX. This Site License is governed by Peacock Data, Inc.'s Terms of Service and Privacy Policy, and the laws and regulations of the United States and the State of California.

COPYRIGHT NOTICE