Business Rules for Standardizing Name Fields

Initial Cleanup of All Names

1. Make all names upper case.
2. Strip leading and trailing spaces from all names.
3. Delete any of the following instances of the following words: 1ST, (1ST), 2ND, (2ND), 3RD, (3RD), 4TH, (4TH), etc. Note: the ERDC has found in the data it processes that these words are used in the same data source to differentiate between people who have the same names. They have not been found to be generational suffixes.
4. Convert letters that contain diacritics (accents) to letters that do not contain diacritics (e.g. "Ñ" is converted to "N").
5. Convert all parentheses and dashes into spaces.
6. Delete all characters that are not either spaces, zeros or alphabetic characters.
    a. The regular expression to perform these deletions is as follows:
        i. s/[^\s0A-Z]//
7. Convert embedded zeros to "O"s in cases where a letter is followed by two zeros.
    a. For example, "S00 B00" is converted to "SOO BOO".
    b. The regular expression to make these conversions is as follows:
        i. s/([A-Z])(00)/$1OO/
8. Convert embedded zeros to "O"s in cases where letters are immediately adjacent on both sides of the zero.
    a. For example, "DEV0N" is converted to "DEVON".
    b. The regular expression to make this conversion is as follows:
        i. s/([A-Z])(0)([A-Z])/$1O$3/
9. Convert embedded zeros in cases where two or more letters is followed by a zero.
    a. For example, "ALBERT0" is converted to "ALBERTO".
    b. The regular expression to the make these conversions is as follows:
        i. s/([A-Z]{2,})(0)/$1O/
10. Convert embedded zeros in cases where a zero is followed by two or more letters.
    a. For example, "0LLIE" is converted to "OLLIE".
    b. The regular expression to the make these conversions is as follows:
        i. s/(0)([A-Z]{2,})/O$2/
11. Delete all remaining zeros. Note: Implicit in the process is the fact that in all cases where a single letter followed by a single zero, the zero is deleted (e.g if a middle name is composed of the following characters, "J0", the zero will be deleted, leaving behind the "J".
12. Convert sequences of two or more spaces to a single space.
13. Convert into the empty string any name that is one of the following: UNKNOWN, ESTATE.

14. For last names, strip out any spaces that follows any of these words:
    a. AL, DE, DEL, DER, EL, MC, LA, LE, MAC, ST, VON, VONDER, VAN, VANDER
15. Where an ethnicity indicator, such as an "Ethnic" variable, indicates a person as being Hispanic or Latino, convert all "MA" words to "MARIA".

Isolate Generational Suffix

1. For the first name field, if the field consists of two or more words, remove the last word into the suffix field if the last word consists of one of the following: II, III, IV, V, VI, VII, VIII, ESQ, JR, SR.
2. For the first name field, if the last word ends with JR or SR and is three or more characters long, remove the JR or the SR into the suffix field. If there is only one word in the field, this condition still holds. The last word in a single word field is that word.
    a. The regular expression to identify the suffixes is as follows:
        i. /(II| III| IV| V| VI| VII| VIII| ESQ|.JR|.SR)$/
3. For the last name field, if the field consists of two or more words, remove the last word into the suffix field if the last word consists of one of the following: II, III, IV, V, VI, VII, VIII, ESQ, JR, SR.
4. For the last name field, if the last word ends with JR or SR and is three or more characters long, remove the JR or the SR into the suffix field. If there is only one word in the field, this condition still holds. The last word in a single word field is that word.
    a. The regular expression to identify the suffixes is as follows:
        i. /(II| III| IV| V| VI| VII| VIII| ESQ|.JR|.SR)$/
5. For the last name field, if any word other than the first word is either JR or SR, remove the word into the suffix field.
6. In the first name field, if a word in this field is both seven characters or longer, and terminates in III, remove the III from the word, and put the III into the suffix field.
7. In the last name field, if a word in this field is both seven characters or longer, and terminates in III, remove the III from the word, and put the III into the suffix field.
8. For the middle name field, if the field consists of nothing but III, VII, VIII or JR, remove the value of field to the suffix field. This will result in the middle name field becoming blank.

Cleanup of Middle Name

1. If middle name is blank and the second word of the first name consists of a single letter, move this letter into the middle name field.
2. If second word of first name is the same as the middle name, delete the second word of the first name.
    o i.e.: First name = "T JOSEPH", Middle name = "JOSEPH" becomes
        First name = "T", Middle name = "JOSEPH"
3. If middle name equals "NMN" (No Middle Name), or middle name equals "NMI" (No Middle Initial), then set middle name to an empty string.

Notes

· The listed steps should be followed in the order in which they appear.

- "Word" in this document is defined as a character token that is delimited by spaces, and/or the beginning of a field and/or the end of a field. Each name field can contain more than one word.
- With the exception of rule 15 in "Initial Cleanup of all Names", ERDC has implemented these business rules in the SAS macro **%StandardizeNames**. This macro can be downloaded from http://www.erdc.wa.gov/briefs/technical/.