Test data project

version: 2016-07-30

This project provides a set of test data for use in software testing. I developed this test data for a project I'm working on and am releasing it open-content in the hopes it will help others.

This project contains two versions. The thousand level version contains a randomly generated table of one thousand names; the million version contains one million name records. The data in this project is USA-centric: names are in the standard USA format of first name, middle name, surname. Data uses characters from either the English alphabet or related Latin characters like Ö.

The data in this project is purposely designed to have many of the problems found in real-world data.

Sources

Surnames are from this page:

- http://www.census.gov/topics/population/genealogy/data/2000_surnames.html

Boys' and girls' names are from the following pages:

- http://www.babynamewizard.com/the-top-1000-baby-names-of-2013-united-states-of-america
- http://www.babycenter.com/0_100-most-popular-hispanic-baby-names-of-2011_10363639.bc
- http://www.alohafriends.com/names_traditional.html

Miko O'Sullivan
miko@idocs.com