**990 Mission Glossary, Version 1**

(Paxton, Velasco, and Ressler 2019)

***Data File:***

A list of misspelling/corrected spelling combinations as well as additional corrections designed to aid pre-processing of Form 990 and Form 990EZ Mission Statements (Part III) and Primary Exempt Purposes (Part I). Saved in .csv format as **Mission\_Glossary\_PaxtonVelascoRessler\_v1.csv**.

***Citation:***

When using the glossary, *please cite the correct* *version* to aid replication:

Paxton, Pamela, Kristopher Velasco, and Robert Ressler. 2019. Mission Glossary v.1. [Computer file]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

***Coding Description:***

e-filed 990 and 990EZ mission statements (Part III) and primary exempt purposes (Part I) released by the IRS through Amazon Web Services contain numerous spelling errors, inconsistent abbreviations, and other errors. The 990 Mission Glossary identifies and corrects 11,303 of these.

Misspellings were identified through an iterative, supervised, automated process. We used the Hunspell spell checker to identify potentially misspelled words and five suggested corrections. A supervised process approved words for correction based on a cascading set of rules that incorporated commonly misspelled words, the Levenstein distance score between the misspelled word and its suggested correction, frequency of each misspelling across the mission statements, and a set of 355 human-derived corrections based on patterns seen in the data (e.g., “intl” to “international” or “re” having been cut off from a word so that “habilatate” is corrected to “rehabilitate”). Humans supervised and approved all spelling corrections.

Before this supervised process, American geographic names taken from the geonames.org database were removed and returned afterward. This is because spellchecking would suggest correcting “Glenbrook” to “glen brook.”

Overall, our process focused on correcting words that are incorrect. Thus, compound words like ironworker or downpayment or unbanked that could be identified as misspelled with a suggestion to split them, are not split. Also, words seemingly put together purposefully by a nonprofit such as elderbridge or healthshare are not corrected by splitting the words.

Altogether, 11,303 misspelling/corrected spelling combinations were compiled into a glossary that, when applied, corrects over 250,000 words across mission statements 2010-2015. The first 10,948 rows of the glossary include spelling corrections and the last 355 rows are additional abbreviations, patterns, and corrections that we suggest.

Note that these corrections apply to mission statements that have modified in a few ways. The most important are that words are lowercase and characters such as &, -, and # are separated from words by bookending them with spaces. The full list of preprocessing steps undertaken before the glossary was created can be found in **00.data\_prep\_glossary.py**.

***List of Variables:***

word original word as it appears in a Form 990 or Form 990EZ mission statement or primary exempt purpose

fix suggested correction

***Suggestions:***

We welcome suggestions for additions to the glossary. Please use the form available at www.pamelapaxton.com/990-mission-glossary-suggestion-form to contribute your suggestions for version 2 of the glossary.

***Additional Questions:***

For further questions or concerns regarding the use of this data, contact:

Dr. Pamela Paxton

Department of Sociology and School of Public Affairs

University of Texas at Austin

305 E 23rd St, A1700

Austin, TX 78712-1086

Phone: (512) 232-6323

email: ppaxton@prc.utexas.edu