

Digitized Data Experiences: Cleaning the SoCal File

Kathleen Costello, Gianneschi Center for Nonprofit Research, CSU Fullerton

kcostello@fullerton.edu

www.fullerton.edu/GCNR

The NCCS-Guidestar National Nonprofit Organization Research Database was the data source for the report, “Southern California’s Nonprofit Sector,” published by the Center for Nonprofit Management in Los Angeles and the Gianneschi Center for Nonprofit Research at California State University, Fullerton.

Our prior reports on the nonprofit sector in Orange County and Los Angeles County relied on traditional data sources such as the Business Master File, NCCS Core File, and data from the Registry of Charitable Trusts. In the case of GCNR reports on Orange County, prior data sources consisted of a few financial variables and some descriptive variables, for up to 1,600 filing organizations.

We wanted to use the digitized data from the NCCS-Guidestar National Nonprofit Organization Research Database because we needed specific financial variables that were not available to us in other files. The dataset we received from NCCS-Guidestar included over 300 financial variables for 16,000 organizations.

Because we were the first research team to request releases from NCCS-Guidestar, certain delays in obtaining data were inherent in the process. We received data in batches over several months. Each batch reflected the additional number of records that NCCS-Guidestar had entered into the data since the prior batch. Comparing the population of records across batches was one of the challenges of the project.

Address and NTEE-CC Cleaning

During the times between receiving batches of data, the research team focused on our initial objectives, which were to assess the addresses and NTEE-CC classification codes for all organizations in the dataset. The address information was necessary in order to support our county-by-county analysis, and the NTEE-CC coding was necessary to support our reporting according to major service category. Approximately 25 percent of NTEE-CC codes in the dataset were re-coded based on the research team’s verification of organizations’ primary purposes. (For more on the NTEE-CC verification process, see “Coding California: Results of an NTEE-CC Verification Project” at <http://www.fullerton.edu/gcncr/CodingCalifornia.pdf>.)

1999 Filers absent from the 2000 Dataset

For some organizations we had a Form 990 for 1999 but not for 2000. We mailed letters to several hundred organizations to request a copy of their Form 990 for 2000. We entered data into the dataset for nearly 200 such organizations.

Financial Variables Cleaning

Once this phase was complete, the cleaned geographic/NTEE file was matched to financial variables in a final release of digitized data from NCCS-Guidestar. Cleaning of financial variables included correcting obvious typographical errors (such as transposed numbers, line items entered on the wrong lines, sums not being copied into adjoining cells, etc.), correcting math (when supporting evidence from other cells confirmed solutions), or filling in missing data.

Conclusions about Financial Data

The first conclusion we reached is that NCCS-Guidestar quite faithfully reproduced the data reported on Form 990/990-EZ. Our quality checks confirmed that the dataset was an exact representation of the information contained in the original returns – with very few exceptions of minor transcription errors, far less than to be expected in a project of this magnitude and complexity.

That is at the same time the good news and the bad news. Remaining true to their objective of replicating the data – warts and all – meant forgoing the opportunity to make obvious corrections during data transcription. How tempting it must have been to “improve” the data at this stage!

Nevertheless, NCCS-Guidestar provide us with a faithful facsimile of the truth of Form 990/990-EZ reporting. Some organizations reported negative revenues. About 84 percent reported itemized revenues or reported some details that could be reconciled (although nearly all reported at least “total revenue”). Only 73 percent of organizations itemized expenses (including 98 percent of those filing the long Form 990). Itemizing and/or reconcilable returns represented 98 percent or more of the total revenues and total expenses reported for all organizations in the study: the lowest compliance was among the Form 990-EZ filers and those with very low revenues and/or expenses.

Recommendations

E-filing should result in considerable reduction in the extent of computation errors that we see presently in Form 990/990-EZ returns. It could also mean fewer instances of reporting sums only without supporting itemizations in detailed line items, if users are prompted or forced to enter these details. E-filing also should reduce simple errors such as transposed digits, failing to copy sums from one line item to another line in the return, or inserting sums on the wrong lines.

E-filing technology will have a limited effect if embraced by only the largest organizations, whose current reporting we might consider to be the most reliable because of their access to qualified staff, counsel and accounting procedures. Although organizations with over \$5 million in revenues accounted for 85 percent of all Southern California revenues, they were only 6 percent of the organizations. We may have more to gain from improved reporting among the 94 percent of organizations that are small but whose capacity for accounting and reporting is uncertain. Although their combined dollars may not make a dent when compared to the largest organizations’, our ability to accurately reflect the activities of the majority of nonprofits would be greatly improved.

Southern California’s Nonprofit Sector

http://www.cnmsocal.org/Services/p_nonprofitlandscape.html

This full-color, 154 page report is the first comparative analysis of Southern California’s nonprofit sector. Features include:

- **Southern California Overview**
- **Executive Summary**
- **Comparative analysis of 7 counties:** Los Angeles, Orange, Riverside, San Diego, San Luis Obispo, Santa Barbara, and Ventura
- **Comparative analysis of 6 NTEE categories:** Arts, Culture & Humanities; Education; Environment & Animals; Health; Human Services; Religious
- **Over 100 figures and tables**