Alfonso Braulio Escribá • Katrina Falk Walker • Laura Roman Canal

Computing Lab / Data Warehousing and Business Intelligence

Analytics on the US Nonprofit Sector - *Milestone III*

Filtering the Raw Data

After locating our data files on the National Center for Charitable Statistics (NCCP) website [1], we created an R script[2] to automate the download of 702 csv files. Our first challenge was reconciling the variables given the fact our files were comprised of 8 different types of 990 forms for 51 different states over 3 years. Since our dashboard only focuses on 501c(3)[3] NPOs, we were able to discard three different types of forms which do not pertain to 501c(3)s. Next, we omitted forms from 2012 because they had significantly fewer variables than the following years.[4] Moreover, because forms from the year 2013 had the most data and we did not think that two consecutive years would yield any significant insight to our analytics, we decided to focus on data from 2013 solely. Thus, our database is comprised of data from 213 files: those extracted from the forms 990 pf, 990-pc, ez-pc, pf-pc and pf-pf for each state in the year 2013.

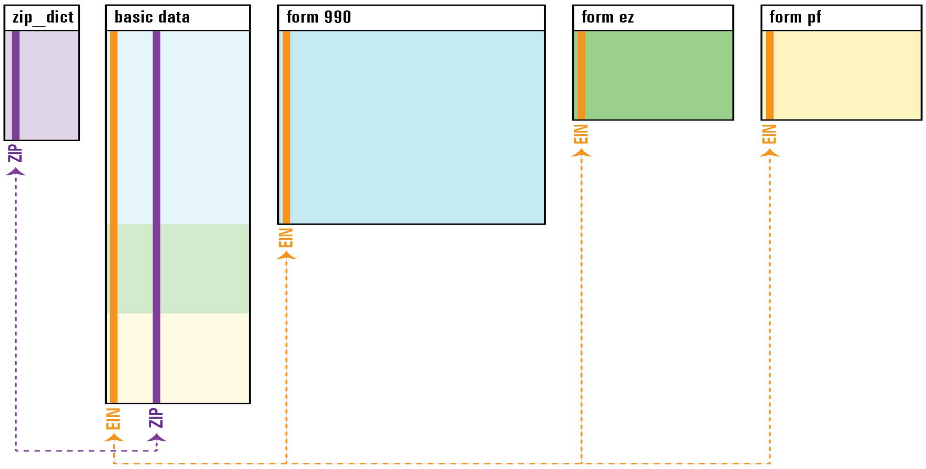
Creating the Database

Because the 990-pf and 990-pc forms have the same columns, we merged them into a single common table. For the same reason, we also merged pf-pc and pf-pf forms into a single table. After selecting the relevant variables from each type of form, we wrote an R-script[5] that generates 3 different tables: one that reads and selects columns of interest from each of the three forms. Another table was created to store basic descriptive variables that applied to all tables. To eliminate redundancy we also created a dimension table with the relation zip-city-state. The resulting data-tables in our database are:

- 990 long forms for private foundations and public charities: 496,905 NPOs and 203 variables
- 990-pf forms: 124,148 NPOs and 196 variables
- 990-ez forms, which are shorter forms for smaller NPOs: 156,591 NPOs and 84 variables
- Basic table: (common descriptive information to all organizations)
- Zip-City-State dimension table

Importing the Data

After creating the schema, we created an SQL-script that enabled us to dump the data from the aforementioned csv files into the database.[6] We established some primary and foreign keys constraints as well as some indexes. As we move into the analytics and visualization part of this project we will continue to refine our scheme so as to optimize user querying over our sizable tables.



[1] Due to a vulnerability in the NCCP website we were able to download the 990 files and circumvent monetary transactions. See http://nccsweb.urban.org/extracts/

[2] See Milestone III Addenda. https://github.com/KatrinaWalker/US-Nonprofit-Sector-Dashboard

[3] 501c(3) are the largest type of 501c() organization. See https://goo.gl/ekzb8x.

[4] For instance, in 2013 990-pc files for California had 264 variables while in 2012 they had 84.

[5] See Milestone III Addenda. https://github.com/KatrinaWalker/US-Nonprofit-Sector-Dashboard

[6] *ibid*