Data Welding: Assimilating Publicly Available Datasets for Competitive Advantage

William Tastle
tastle@Ithaca.edu

James Bondra
jbondra@ithaca.edu

Ithaca College
Ithaca NY

Abstract

Datasets offer companies value by providing detailed information about potential or current customers. As data continues to rise as a new currency, it has become an increasingly lucrative business to purchase specialized mailing lists. This paper investigates a method by which very useful of datasets can be created, by small and medium-sized companies, at little or no cost. Data sources are identified and a method by which the welding can occur is delineated. Thus, any entrepreneur and/or business can create specific lists for virtually any geographical location using very cost-efficient means

Keywords: Open Government Data (GOD), voter data, census data

1. INTRODUCTION

A new currency (Effers, Hamill, Ali 2013; Shan 2017; Evans 2018) has arisen in the digital age named *data* with the term "currency" meaning a medium of exchange rather than as a temporal ordering. As business and trade are increasingly capturing the digital world, the knowledge and a general understanding about this anonymous environment becomes more important. Data can provide companies important information for forecasting, analyzing and improving efficiency of their businesses - if they can understand, read and correctly interpret it. As knowledge is power, entrepreneurs have engaged in this lucrative commodity by creating and selling specific datasets to small and medium businesses that do not have their own capacity or knowledge to generate their own lists. Many of these small and medium sized businesses are facing the problem that they can't afford commercially available mailing lists or data sets.

At the same time, an increasing amount of Open Government Data (OGD) is recognizable and very available to the public at large. Consequently, it seems tempting to use these available sources, if only one could find a way of making data at the national level applicable to much smaller geographies. These datasets are typically of a very general, all purpose, too frequently available only in summarize form. Thus, they are not tailored for specific economic use and must be modified to get a usable dataset containing useful information.

This paper provides a method of merging two publicly available datasets to create a specialized mailing list which is useful for marketing research and sales. Of the governmental data available, one is available from every county government in the US: the voter database. One might have to request it through Freedom of Information, but they are available and usually at a minimal cost, unless one seeks data from a rural county where the local Board of Elections may charge hundreds of dollars. This huge price differential is

uncontrolled and gives one pause to wonder why such extremes exist. The other dataset is that available from the Census Bureau. The welding of these two datasets can provide some rather specific information such as median income, level of education, and housing values to generate a targeted list.

2. BACKGROUND

Evolution and Strength of Open Government Data (OGD)

The availability of Open Government Data (OGD) developed in the 21st century has resulted in a wealth of potential information determined by one's ability to gather and analyze it in meaningful ways. In the period from 2009-2013 the initiatives of OGD has grown from two to over three hundred, while the membership in the Open Government Partnership increased from eight to fifty-nine countries in two years. At the same time, over 280 government catalogs have been published and over a million datasets have been released by governments around the world (Jetzek, et al. 2014, p. 101). This movement towards a more transparent government holds much promise (Below 2015). The official website of US government data (https://www.data.gov/open-gov/) lists these advantages of open data: increasing citizen participation in government, creating opportunities for economic development, and informing decision making in both the private and public sectors (Government Data U.S., 2017). The European Open Data Strategy launched in 2011 is expected to boost the European economy with €52 billion per year (Jetzek et al., 2014). In 2015, the value of the EU *data* economy was more than €285 billion (European Commission, 2017).

However, a wide range of fields can benefit from open data including "health, energy, education, public safety, finance and global development. The hope is that OGD will eventually lead to the generation of substantial value." (Jetzek, et al., 2014)

3. USED DATASETS

CENSUS DATA biggest publisher in the world of data

The United States Census Bureau is a principal agency of the U.S. Federal Statistical System. Their mission "is to serve as the leading source of quality data about the nation's people and economy' (United States Census Bureau 2017). Since the adoption of the Constitution in 1790, a census of the population has been completed every 10 years and since 1902 has been under Census Bureau control, a division of the U.S. Department of Commerce. The agency collects and publishes a huge amount of statistical data about the U.S. population for different kinds of purposes. The surveys include topics like households, businesses and politics as well as other demographic facts. These data are important in the redrawing of Congressional districts but can also be abused by gerrymandering districts to fit political agendas. Due to privacy concerns, the Census is not allowed to publish data specific to a particular address. Therefore, households are assigned to different block numbers within each county which aggregates all personal information.

For the purpose of this paper, the data about the *Income of American Households* and the *Value of Home* are used in the illustration of data welding. Thus, a detailed examination of *appendix 1* will give step-by-step instructions as how one can acquire those data sets.

VOTER DATA in the U.S

For the second source, the voter database in available from every Department or County Board of Elections Office throughout the US, sometimes through a Freedom of Information application. As the United States does not have a federal election agency and consequently no national voter list, each state is responsible for their voter database. Thereby, each state can gather the data from the individual counties, but it is the individual county that has full control over the database. The county, or in some cases the State, places their own charge for the generation of the CD or other medium on which the data is saved and those charges range from free to hundreds of dollars. *Appendix 2,* contains an example of the standard voter export file format with the high variety of different data. States may have their own format end will provide it with the data. Particularly, this detailed dataset offers a tremendous value to those who can organize the data into meaningful segments. By merging a voter data set with another dataset such as the census data, can reveal valuable and useful information in the identification of potential consumers, their characteristics, civic involvement, school district, political affiliation, age, and possibly phone number.

4. WELDING DATASETS
An Illustration

By means of this example it will be shown that the merger of local voter data with census track data can yield very useful results such as the median income and median house values by

ISSN: 2473-4901
v5 n4901

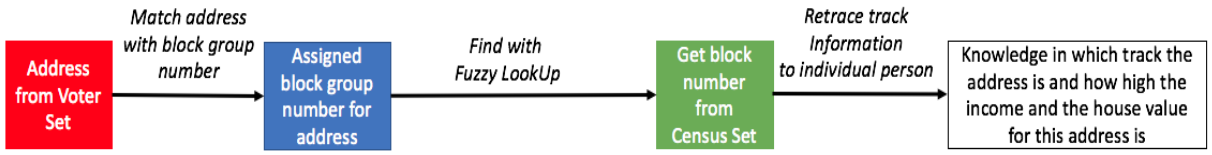

Figure 1 Welding data tables using Fuzzy Lookup

specific census districts, and the names, addresses, date of birth, geopolitical district, gender, and frequency of voter participation, to generate a detailed mailing list in which to target specific regions. As result, the list provides detailed information about the median value of that home in a particular census district. Therefore, the address from the voter set must be used to find the block group in the Census set. After identifying the associated block group number containing the information about house value and income, it can reverse engineered to the specific person via the address. As a result, we can generate knowledge about the economic value of each person (*see figure 1*).

1. Given the Voter database and the Consensus track file, format both lists to Excel tables (CTRL+T) and name them *VoterRawData* and *CensusData*, respectively.
2. Secondly, clean the data, removing duplicates, correcting misspellings and the like. Power Query in Excel is extremely useful in data cleaning. Unnecessary columns of data can be eliminated to reduce the complexity of the operations. For this example, the only necessary fields are *unique ID (name), the street number, street name, city, state and zip code.*
3. After cleaning the data, the resulting file should be given a different name, such as *VoterAdjData*.
4. Next, the addresses from the *VoterAdjData* table must be matched with the corresponding block group number and this requires the identification of the Census Block identifier. It is through this association that specific names and addresses be matched with their assigned block group identifier. Concatenate the street number, street name, city, and state in Excel to create an aggregated address. Using these aggregated addresses, we can identify the block group numbers for each, using the Census Geocoder API located: https://geocoding.geo.census.gov/geocoder/geographies/addressbatch?form. *Appendix 3* provides a step-by-step-instruction for using Geocoder. Save the list generated from the Geocoder and name it ***VoterGeocode.***
5. Copy both tables (*VoterGeocode* and *Census)* in one workbook in two sheets.
6. To assign the address (with the block group number) from the *Voter* data and the Census Data, use the Excel add-in Fuzzy LookUp. This add-in allows for the identification of textually similar data records in two different tables
7. Run FuzzyLookup and select the left and the right tables from the drop-down menus. Matching rows from the right table will be returned for each row in the left table.
8. Select the columns to match on. Now select the *block group number* column from the *VoterAdjData* and the *block group column* from the *Census* data. Run the add-in.
9. Following the Fuzzy Add-in instructions, select one maximum number of matches and adjust the similarity threshold until the resulting data seems as usable as possible. This may take a few tries
10. Move the current cell selected in the Excel spreadsheet to an empty cell which has empty space to the right below it.
11. Press the "Go" button to perform the match

(Microsoft 2012)

5. OUTPUT AND CONCLUSION

After using the *FuzzyLookup*, a table with all address and the matching block group numbers will be generated. The actual value of this list is that now the block group number is directly associated with the economic information when means the addresses are assignable to a specific

person with along with a reasonable estimation of their property value and income.

Such a list offers businesses a tremendous value when it comes to market research or actual promoting and selling new products. Companies can use these merged sets to promote specific products depending on the individual information and their assets. With this method marketers would have an overview about their potential customers and which areas would be the best target for goods or services they are offering.

With this skill, small and medium sized companies are no longer dependent on firms offering the same information at excessive prices.

6. ACKNOWLEDGEMENTS

The authors graciously thank graduate student Teresa von Haken for her efforts helping us with this paper.

7. REFERENCES

Below, B. (2015), "Shedding light on government, one dataset at a time." 3rd International Open Data Conference. http://oecdinsights.org/2015/05/29/shedding-light-on-government-one-dataset-at-a-time/.

Eggers, W., Hamill, R. and Ali, A. (2013), "Data as the new currency: Government's role in facilitating the exchange." Deloitt Insights, https://www2.deloitte.com/insights/us/en/deloitte-review/issue-13/data-as-the-new-currency.html

European Commission (2017). Building European Data Economy. Retrieved from: https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy

Evans, M. (2018), "Why Data is the Most important Currency Used in Commerce Today." Forbes, https://www.forbes.com/sites/michelleevans1/2018/03/12/why-data-is-the-most-important-currency-used-in-commerce-today/#6b1ef16d54eb.

Government Data U.S. (2017). Open Government Data. Retrieved from https://www. data.gov/open-gov/

Thorhildur, J., Avital, M. & Bjorn-Andersen, N. (2014). Data-Driven Innovation through Open Government Data. *Journal of Theoretical and Applied Electronic Commerce Research, 9 (2), 100-120.*

Microsoft (2012). Microsoft Fuzzy Lookup Add-In for Excel. Retrieved from: https://atidan.files.wordpress.com/2013/08/fuzzy-lookup-add-in-for-excel.pdf

Shan, P. (2017), "Data: The New Currency." D!gitalist Magazine. https://www.digitalistmag.com/cio-knowledge/2017/12/11/data-new-currency-05592449.

United States Census Bureau (2017). What we do. Retrieved from: https://www.census.gov/about/ what.html

Wikipedia (2017). Voter Database. Retrieved from: https://en.wikipedia.org/wiki/Voter_database

ISSN: 2473-4901
v5 n4901