

Mysteries of Posting Results on reddit

Adeline J. Wilcox, Retired

ABSTRACT

After reading a December 2016 USA TODAY story revealing data leaked from the U.S. Department of Veterans Affairs (VA), this retired VA employee doubted the quality of the leaked data. After I completed my analysis of the data published by USA TODAY, I wanted to share my results, but not by corresponding with USA TODAY. Naively, but successfully, I created a reddit account and posted my analysis on the subreddit named *dataisdepressing*. My post remained at the top of the front page of the *dataisdepressing* subreddit for nearly three months, despite receiving no comments and few upvotes.

For analyses of general interest that produce cheerless results, the *dataisdepressing* subreddit can be a means for sharing results or visualizations that can be displayed on a single screen. Reddit users submitting content to the *dataisdepressing* subreddit should not be disappointed if their post receives no more than a few hundred views.

INTRODUCTION

Called a social news aggregator [6], reddit allows its registered users to discuss and rate content. For those stuck at their computers, reddit offers abundant light reading and an online community. Posting on reddit is governed by rules known to users and algorithms unknown. Subreddits may have their own rules for posting. Generally, reddit users must vote and comment on subreddit posts before they are allowed to post to that subreddit.

MASTER DATA

Although posting my work on reddit would be years in the future, my journey to becoming a reddit user began when I bought a book titled *Master Data Management in Practice* [1]. Finding that the counts of healthcare facilities differed among the data files I worked with frustrated me. Cervo and Allen's emphasis on beginning data quality efforts with master data management inspired me to study the Veterans Health Administration (VHA) facility data stored in VHA Site Tracking (VAST) [3]. Cervo and Allen cite Garnter

Master data management is a technology-enabled business discipline that helps organizations achieve a "single version of the truth"

By September 2015, I had been studying VAST since the beginning of the year and had become confident of my knowledge of it.

OPEN DATA

To promote what he called open data, DJ Patil, the first White House Chief Data Scientist, held a forum at the U.S. Department of Veterans Affairs Central Office in early September 2015. Seeing advance publicity for this event on the VA intranet, I looked at the file touted as "open data". I found it full of errors such as duplicate records and identifier values different from those given in VAST. At the forum, I told DJ Patil the data were garbage. Neither he nor anyone else in the Obama Administration was able to improve the quality of these open data. Under the Trump Administration, VA has continued to publish this dirty data file on GitHub [2]. Today, I can still find the same duplicate and erroneous records I reported in 2015.

After the forum, I opened an issue on GitHub. I headed it `facilities.json is garbage`. In my issue, I reported the data quality problems I found. Because some GitHub users made unkind posts about me, I later closed the issue. The thread was eventually deleted by a GitHub user unknown to me.

To be helpful, I wrote SAS/STAT[®] code to produce a model file named `divipast.json`. The first JSON file I ever produced, it was imperfect but still better than VA's. I also saved the data in a CSV file and published it on Socrata.

ANALYSIS OF DATA PUBLISHED BY USA TODAY

I carried out my analysis using SAS University Edition. By selecting all of USA TODAY's published data, followed by copy and paste into a text file, I easily saved their data. Next I wrote a SAS program to read only those records holding data and write them to a CSV file. For reference data, I downloaded the data set I'd uploaded to `opendata.socrata.com` in 2015. Comparing VHA's nonstandardized data required hours of clerical matching.

With a `DATA _NULL_` step, I wrote an output file that I marked up with LaTeX code to prepare a publication quality table in PDF format. My data table fit on one page. My reddit text post [7] included a link to the data table which I had posted on imgur. Imgur is an online image hosting service designed to work with reddit [5].

SUBREDDITS FEATURING DATA ANALYSES

At this writing the subreddit named `dataisdepressing` has over 3,000 subscribers. Of course, this does not mean all subscribers will click on any link or post or even visit the subreddit. A quiet subreddit, only a dozen posts were submitted to `dataisdepressing` in the past year. An annotated graph I programmed with ODS GRAPHICS and posted later got more upvotes than my post on VHA data, even though the graph displayed county, not national, data. Data visualizations may be more appealing to reddit users than tabled results.

By comparison, `dataisbeautiful` is a subreddit with well over 12 million subscribers and many original content posts every day. Unlike the `dataisbeautiful` subreddit, `dataisdepressing` allows tabular presentation of data. Besides `dataisdepressing` and `dataisbeautiful`, reddit has more subreddits devoted to data analysis.

POSTING ON REDDIT

Due to the size and complexity of reddit, this is only an introduction to posting results there. When posting your own work on subreddits featuring data analyses, tag it [OC], identifying it as Original Content. To alert potential readers that your post is one they have not already seen elsewhere on the Web, end the title of your post with [OC].

While JPG files may be uploaded to reddit, PDF files must be uploaded to Imgur. The subreddit named `dataisdepressing` could be an opportunity for SAS programmers to present their work to a slightly larger audience than they may have found before.

Possibly because of my incomplete knowledge of reddit, when I posted to the `dataisdepressing` subreddit the second time, I needed help from the moderator.

CONCLUSION

Before I posted my results on `dataisdepressing`, VA did not care about correctly reporting the number of its healthcare facilities. The VHA web site gave 168 VAMCs while the rotating slide show on the same web site gave only 152. Today, on www.va.gov/health/, the rotating slide show and web page both give 170 medical centers. Whether these 170 VAMCs were properly classified cannot be determined from the publicly available data. That these values now match may not even be a positive development. VHA should make its VHA Site Tracking database [4] current with ground truth and publish a data set holding complete records on all its healthcare facilities.

Data presentations, particularly data visualizations on a single page, may find a popular audience on reddit.

References

- [1] Dalton Cervo and Mark Allen. *Master Data Management in Practice*. John Wiley & Sons, Inc. Print, Hoboken, NJ, 2011.
- [2] Department of Veterans Affairs. `VAFacilityLocation.json`. GitHub, February 2017. Accessed 27 September 2017.
- [3] Robert A. Petzel. VHA Site Classifications and Definitions. VHA Handbook 1006.02, 30 December 2013.
- [4] David J. Shulkin. *Assignment And Maintenance Of Station Numbers And Attributes*. VHA Directive 1044. Department of Veterans Affairs, Veterans Health Administration, Washington, DC, 15 March 2016.
- [5] Wikipedia. `Imgur` — wikipedia, the free encyclopedia, 2017. Online; accessed 25-September-2017.
- [6] Wikipedia. `Reddit` — wikipedia, the free encyclopedia, 2017. Online; accessed 25-September-2017.
- [7] Adeline Wilcox. VA Medical Centers Missing from USA TODAY Report [OC]. reddit, 30 January 2017.

Credit Notices

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Contact Information

Your comments and questions are valued and encouraged.

Contact the author: Adeline J. Wilcox
 10901 Fleetwood Drive
 Beltsville, MD 20705-2502
 301.931.0356
 awilcox1 at mindspring dot com