Extracting Insights from FBO.Gov data – Part 3

Earlier this year, Sunlight foundation filed a lawsuit under the Freedom of Information Act. The lawsuit requested solication and award notices from FBO.gov. In November, Sunlight received over a decade’s worth of information and posted the information on-line for public downloading. I want to say a big thanks to Ginger McCall and Kaitlin Devine for the work that went into making this data available.

In the first part of this series, I looked at the data and munged the data into a workable set. Once I had the data in a workable set, I created some heatmap charts of the data looking at agencies and who they awarded contracts to. In part two of this series, I created some bubble charts looking at awards by Agency and also the most popular Awardees.

In the third part of the series, I am going to look at awards by date and then displaying that information in a calendar view. Then we will look at the types of awards.

For the date analysis, we are going to use all of the data going back to 2000. We have six data files that we will join together, filter on the 'Notice Type' field, and then calculate the counts by date for the awards. The goal is to see when awards are being made.

The Data Munging
The CSV files contain things like HTML codes and returns within cells. During initial runs, the Piggybank function CSVExcelStorage wasn't able to correctly parse the CSV files. Other values would end up in the date field and provide inaccurate results. To combat that problem, I ended up finding the SuperCSV java classes. Using their sample code, I was able to write a custom cell processor that would read in the CSV file, strip out carriage returns, and write the information back out in a tab-delimited format.

My friends at Mortar are taking a look at the data and the CSVExcelStorage function to see if my issue can be addressed.

Pig Script
The pig script is really simple. I read in all of the metadata files, filter them where there is an award and it isn't null. All six files are then joined and a new set of data is generated of just the date of award. These are then grouped and counted. No real rocket science there.

Here is a sample of the results:

Date	Amt
2012-09	
2009-09	
2010-09	

2010-09
2009-06

Visualization
For visualization of this data, I used an adaptation of Mike Bostock's D3.js calendar example. I used the same colors but made a few tweaks to the scale. A snapshot calendar is below:

In this visualization, larger award counts are green while lower award counts are in red. Anyone familiar with government contracting will notice that "end of the fiscal year" awards are always made in September. Contracting officers need to spend fiscal year money before it expires. In the calendar, it is easy to see this for each of the fiscal years.

Data by Categories
Let's look at the data by categories. There are 103 unique categories of award types in the data sets. You can see that list here. There are over 9600 combinations of agencies awarding contracts in a category. You can see the CSV file here.

One of the curious categories is the "Live Animals" category. The agency, category and number of awards are listed below:

Agency Category

Customs
National
National
Centers

Preventi
Army Co
MICC
Office of
Officer
United S
Agricultu
Animal a
Inspecti
Food an
Air Forc
Bureau
Firearms
Forest S
Public B
TRICAR
U.S. Sp
Comma
United S
United S
(USSS)
Army Co
Bureau
Bureau
Direct R
District o
Direct R
Farm Se
FedBid
Fish and
Fresno V
Office of
Pacific A
San Fra
U.S. Arn
VA Conn
System
Washing
Services

Bi-Grams in the Description

I decided to take a look at descriptions associated with the awards and see what bi-grams appeared when running this through NLTK. I created a simple Pig script to filter the awards, join them to the descriptions and then run the descriptions through Python/NLTK.

The Python/NLTK piece is below:

Results
A majority (109 out of 192) consisted of "No description provided". Thus the bi-gram results were "{(description provided),(no description)}". Of the remaining awards, there were some interesting results. For example:

{(, boarding),(10 paso),(5 potential),(and general),(boarding and)} was from a US Marshals Service Contract for "serives to include the transportation, boarding and general care of 10 paso fino horses".

{(country of),(of japan),(japan .),(the country),(& removal)} was from a Pacific Air Forces contract for "this is feral pig control & removal service at tama service annex. "

and then there was this one:
{(for the),(farm hands),(on april),(a contract),(ms. crundwell)}
{(for the),(the horses),(on april),(care for),(ms. crundwell)}

contract was awarded under far 6.302-2, unusual and compelling urgency. on april 17, 2012, rita crundwell, comptroller for the city of dixon, illinois, was arrested by the federal bureau of investigation (fbi) for wire fraud. ms. crundwell was accused of embezzling $53 million from the city of dixon. on april 18, 2012, the u.s. marshals service was notified by the fbi that ms. crundwell's bank accounts had been frozen. the defendant was using embezzled money for the care and maintenance of over 200 horses located in dixon, illinois and beliot, wisconsin and several additional locations. since her accounts were frozen, the farm/ranch did not have the means to care for the horses. farm hands continued to work after this date even though they were unsure if they would receive payment. if the farm hands were to walk off the job, there would be no means to care for the horses. in addition, weekly deliveries of hay and grain would cease. the usms issued purchase orders for hay and grain deliveries to continue until a contract award could be made for the management and care of the horses. the government recognized that there was an immediate need to care for the animals. if a contract was not awarded immediately, the lives of the horses would be at risk.

You can read about Ms. Crundwell here. There were two contracts awarded for $625,840 and for $302,850 to ensure the horses' safety.

In summary, I looked at the award dates to look for patterns and then I looked at the text descriptions to look for interesting data combinations.