WSDOT DISCLOSURE REQUESTS

Week One: A Look at the Disclosure Datasets

By Angela Gonzalez-Curci for WSDOT as part of University of Washington/Open Data Literacy Internship

# Purpose

The purpose of this document is to document my exploratory analysis WSDOTs Public Disclosure Request Record dataset as well as propose two options for analysis going forward.

# Analytics Workflow

I learned this workflow in a professional development course I took a few years ago and believe it to be very useful in describing a project and keeping it within scope:

Progress as of Week One

## Identify the Problem (What is the objective? What questions must the dataset help answer? What does success look like?) :

### Objective:

Finding frequently requested datasets and documents, without any public record exemptions, to be nominated for preemptive open disclosure. In short, find datasets and documents that can be put into WSDOT Open Data repository or on the agency’s website, bypassing the public request process.

### Questions:

* What are the most frequently requested datasets?
  + Without exemptions?
  + With minor or contextual exemptions what can be controlled for?
* Who is requesting most frequently, and what types of data sets?
  + What does requestor type tell us about what requester will do with received data or documents?
* In what media are approved request being send in, and what are their costs?
  + Could the media used tell us something about the size of requests? Could the cost savings of making certain records and datasets open be quantified?

### Goals

* The primary goal is to create a ranked list of the most frequently requested types of datasets and records, and then assess the types of exemptions associated with that type and find the least restricted datasets and records for open disclosure.
  + A secondary goal is to find a relationship between particular types of requestors and records in order to make inferences about the nature of the requests.
  + A secondary goal is to quantify the potential cost savings of preemptive disclosure.

## Obtain Data (Identify Datasets, Import Data, Determine Tools):

### Identify and Import Datasets

On 6/18/2018 two datasets were retrieved from the Records Requests FileMaker database. The Old dataset consists of 6967 entries of closed records requests, while the New dataset consists of 8706 entries of open, though not necessarily active, record requests. These datasets differ not only in length, but also in numbers of columns and column order. In order to combine the two different datasets into one master Request Set, the columns must be uniform. Therefore, the datasets had to be standardized before merging them.

In Excel, I created two dataset with 39 columns with columns organized in alphabetical order. All of the administrative data and duplicate columns were deleted, as well as descriptive data not needed for the question. One column, CopyQuantityTotal, was added to the Old dataset; it already existed in the New dataset. Another column, Approval Comments, which details the exact dataset or record given to a requestor, does not exist in the New set. Since I do not want to exclude more than half for the records in order to use this one attribute, I have not included it in this dataset. I may choose to look at this attribute independently.

I won’t go through all the chosen fields, but here is a list of included data topics:

* WSDOT Employee/Group assigned to a record
* Collision Information
* Company and Individual Who Requested Records
* Location of Requestor
  + State and Zip
* Quantity, Cost, and Material Delivered to Requestor
* Item Denied
  + Statue
  + Wording of Statute
* Disposition of Request

### Determine Tools

At more than 15,000 entries, this dataset is too large to be filtered and cleaned in Excel, the original file type received. However, a sample of the dataset can be used an analyzed using Excel if need be. This process will be explained further in the proposal section. Another option is to use Python to clean and analyze the full dataset. SQL could also potentially used, however without access to a specific server to store the dataset and to a database which the table belongs to, it is the least possible option.

## Understand Data (Data Documentation, Exploratory Analysis, Quality Assessment)

### Data Documentation

One major challenge I’ve faced so far is the lack of data documentation I have access to. I do not have direct access to the data definitions in FileMaker nor do I have access to WSDOT’s internal data dictionary, DOTS, yet.

That being said, I have received some useful documentation from Andy Everett, which has enable me to understand the public request data to some extent:

* Public Records Database Schema
  + The new schema for public record request database. The FileMaker program is shortly going to be phased out. Knowing the new schema will help me make relevant recommendations and figure out the data definition of some of the fields I am using
* Data Categories
  + This document shows all the different types of data which WSDOT creates. This will help me discern when a request is asking for a dataset or another type records.
* Asset Management Taxonomy
  + This documents the relationships between terms used in WSDOT internal asset management. This allows me to understand the scope of a requested document by looking at where the term used in a request fits in the taxonomy.
* Business Topics
  + List of business topics WSDOT deal with. This also allows me to place a requested item in context to its subject matter. Note: this is also a work in progress, Andy would like feedback on added topics and clearer structures.
* Keywords PDR Kent Study:
  + The recommendations of the study of Kenneth Lee, a graduate student at Kent State University in 2014. Based on WSDOTs records he created a record classification schedule for Public Request Records. I could potentially use and modify to classify the

These documents could also help me standardize Request Types found in the PRR dataset I am analyzing. Right now, the types are unstructured, so although similar request are known to the trained eye, they will not be easily analyzable by machine. Standardizing the request types will make, for an easy example, value entries: Employment Records (Name of Person), (Name of Person) Personal Records, and (Name of Person) Employee Record all have the same common value, which a computer can read. These standardized request types can then be used, and worked upon.

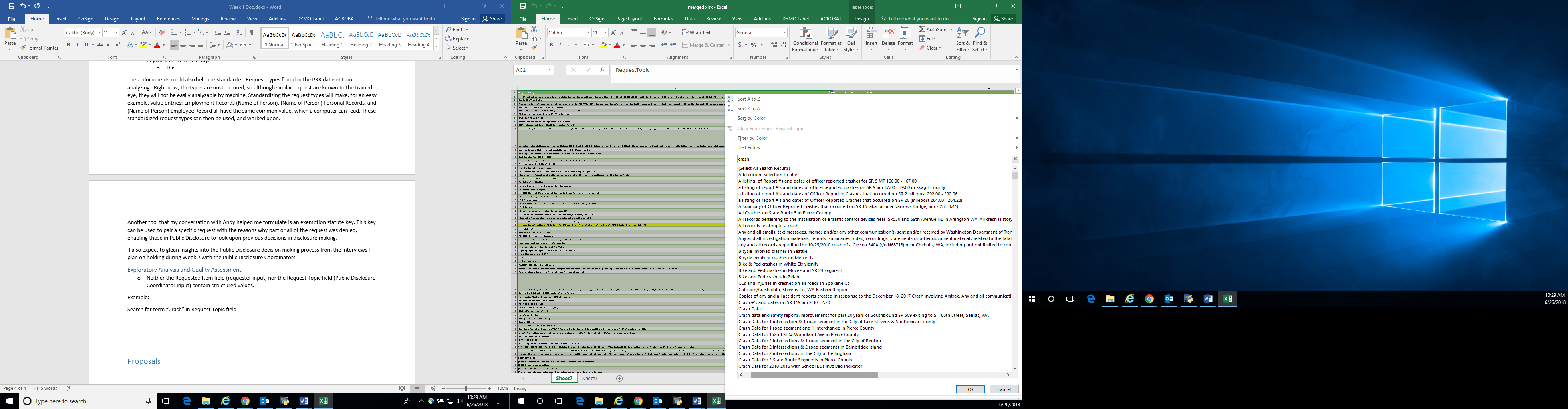
Another tool that my conversation with Andy helped me formulate is an exemption statute key. This key can be used to pair a specific request with the reasons why part or all of the request was denied, enabling those in Public Disclosure to look upon previous decisions in disclosure making.

I also expect to glean insights into the Public Disclosure decision making process from the interviews I plan on holding during Week 2 with the Public Disclosure Coordinators.

### Exploratory Analysis and Quality Assessment

Requested Data Fields (Requested Item and Request Topic)

* Neither the Requested Item field (requester input) nor the Request Topic field (Public Disclosure Coordinator input) contain structured values.
  + Example:
    - Search for term “Crash” in Request Topic field



Different wording, but similar requests. All requests are asking for crash histories of particular locations

Similar wording, but different requests. First wants all crash history near particular traffic control device, the second wants all records relating to a particular crash

* Other challenges include request for multiple type of records in one request entry
  + Example:
    - “All records pertaining to the installation of a traffic control device near SR530 and 59th Avenue NE in Arlington WA. All crash History”
      * Wants both records about the traffic control device and the crash history near there.
* Abbreviations and Symbols used in the place of words
  + Example:
    - Crash # vs. crash number
    - @ vs. at
* Similar terms with different meanings and different terms with same mean
  + Example:
    - Crash is not the same as Collision
    - Crash Data is part of Crash Report

Void and Non-Disclosure Request Entries

* This will be easier to clean, but still something of note. Both test records and incorrectly entered entries need to be scrubbed from the dataset.

# Proposals

As my understanding of the data evolves, it is clear that the unstructured nature of the requests is going to be a major challenge to my analysis. The following proposals deal with this challenge.

## Sample in Excel

Take a statistically significant random sample of PDR request and code requests by using and further developing Ken Lee’s classification schedule. Find frequency using pivot tables and finding answers to secondary questions after finding most frequently request types.

### Pro:

Does not require any extra software downloads

Uses techniques I have personally used professionally before

### Con:

Requires tedious manual work

Size of dataset (greater 10,000 entries) may make randomization difficult

## Natural Language Processing with Python

Use the technique of natural language processing to find most frequently terms in request strings, use Ken Lee’s classification schedule to create and develop types of frequently requested data set types.

Pro:

Can easily document process, and code used for replication.

Can use whole dataset

Con:

Will need to learn how to use technique. Know some Python, and did exercise in Natural Language Processing in R, but have never used the technique in Python. I’ll need to teach myself how to do it.

Requires extra software.