Description of document:	Food Safety and Inspection Service (FSIS) report and presentation from contract with MITRE Corporation/Todd Reed for Text Mining regarding 1) disease testing, and 2) ready to eat products, 2014
Requested date:	28-December-2018
Release date:	30-September-2019
Posted date:	24-February-2020
Source of document:	Freedom of Information Act Officer USDA, Food Safety and Inspection Service Room 2168 South Building 1400 Independence Ave., SW Washington, DC 20250 Fax: (202) 690-3023 E-mail:fsis.foia@usda.gov

USDA

United States Department of Agriculture

Food Safety and Inspection Service

1400 Independence Avenue, SW.
Washington, D.C.
20250

September 30, 2019

Transmitted via Electronic Mail

RE: FOIA-2019-00157
Mitre Corporation Report & Presentation

This is the final response to your Freedom of Information Act (FOIA) request, dated December 23, 2018, to the U.S. Department of Agriculture's Food Safety and Inspection Service (FSIS). You requested access to the report and presentation from the contract with Mitre Corporation/Todd Reed for Text Mining. We received your request in our Office on January 28, 2019. We apologize for the delay in responding to your request.

The FSIS FOIA staff works with subject matter experts across the Agency to locate responsive documents. For this request, we conducted a records search in the Office of Management. FSIS' search began on February 15, 2019.

We have located 58 pages that are responsive to your request. After a thorough review, we have determined that these pages may be released in their entirety. Accordingly, your request is granted in full.

You may appeal this determination within 90 days from the date of this letter. Your appeal should include copies of your original request and this response, as well as a discussion of the reasons supporting your appeal. The envelope should be plainly marked to indicate that it contains a FOIA appeal. If you decide to appeal this determination, please send your appeal to:

Carmen Rottenberg
Administrator
Department of Agriculture
Food Safety and Inspection Service
1400 Independence Avenue, S.W.
Room 1168, South Building
Washington, D.C. 20250-3700

Please be advised that your FOIA request, including your identity and the information made available, is releasable to the public under any subsequent FOIA requests. However, FSIS does not release your personal privacy

information, such as home addresses, telephone numbers, or Social Security Numbers, all of which are protected from disclosure under FOIA Exemption 6.

If you have any questions about the way this request was handled, please contact Emmanuel Olufotebi at 202-260-9433 or via email at Emmanuel.Olufotebi@usda.gov. If you have general questions about FSIS' FOIA procedures or regulations, please contact me at 202-690-2760 or via email at Arianne.Perkins@usda.gov.

Additionally, you may contact the Office of Government Information Services (OGIS) at the National Archives and Records Administration to inquire about the FOIA mediation services they offer. The contact information for OGIS is as follows: Office of Government Information Services, National Archives and Records Administration, 8601 Adelphi Road-OGIS, College Park, Maryland 20740-6001, e-mail at ogis@nara.gov; telephone at 202-741-5770; or facsimile at 202-741-5769.

Thank you for your interest in FSIS programs and policies.

Sincerely,

ARIANNE PERKINS

Digitally signed by ARIANNE PERKINS
Date: 2019.09.28 16:37:22 -04'00'

Arianne M. Perkins
Director, Freedom of Information Act Staff
Office of Public Affairs and Consumer Education
Food Safety and Inspection Service

Enclosure

Prepared for:

Data Analysis and Integration Staff,
Office of Data Integration and Food Protection,
Food Safety and Inspection Service, USDA

DRAFT: Categorizing Causes for Cancelled Pathogen and Residue Samples

July 18, 2014

Version 1.0

MITRE
7515 Colshire Drive
McLean, VA 22102

Record of Changes

No.	Date	Reference	A=Add M=Modify D=Delete	Description of Change
1	7/18/2014	First Draft		

Executive Summary

The Data Analysis and Integration Staff (DAIS) requested MITRE support in identifying the categories of reasons for cancelling samples and then determining the feasibility of automatically categorizing free text narratives from an important data stream. In the most recent fiscal quarter, nearly 20% of samples were cancelled. Cancelling a sample and (potentially) rescheduling it takes valuable time from an inspector's already busy day. Currently, inspectors can record the reason for a sample cancellation by a pull-down menu of options or by a free text narrative. DAIS wants to determine if information in the narratives can be leveraged to help improve FSIS's information technology (IT) system – the Public Health Information System (PHIS) – to decrease the number of cancelled samples, to decrease inspectors' frustration with the IT system and to increase inspectors' efficiency in scheduling and performing samples.

MITRE performed an initial content analysis to identify 21 specific categories of reasons for sample cancellation and five high level categories. MITRE developed a automated categorization script that correctly categorized 61.5% of records for the full set of categories and 78.5% of records for the high level categories.

This report documents the steps that MITRE took and the categories that MITRE identified. The categories and the script should be viewed as initial steps in this process to understand the feasibility of this approach as well as its possible utility for FSIS's mission. MITRE encourages feedback and further collaboration on the categories as well as the rules developed as part of the script. Along with this report, MITRE delivered the source code for the script and a table of data that includes the category for each narrative.

The categorization script that MITRE developed as part of this effort should continue to perform as well on unseen data as it did on the test set as long as there are no substantial changes in the language of the justifications. For several of the categories, MITRE does not foresee a likelihood of substantial change (for example, weather may always cause issues with transport of samples to labs). However, as PHIS is modified to meet the needs and expectations of the inspectors, MITRE can foresee new topics and issues arising from those modifications to PHIS. MITRE recommends that FSIS periodically review a sample of cancellation descriptions or apply techniques from corpus linguistics to identify new issues and new categories.

Overall, MITRE found that roughly 64% of sample cancellation narratives and roughly 13.5% of all sample cancellations in the most recent fiscal quarter – third quarter of Fiscal Year 2014 (3Q FY14) – may have been caused by IT issues. In the same period, MITRE found that nearly half of narratives in a sample of 200 narratives included some reference to PHIS scheduling double or triple (or more) samples when an inspector intended to schedule just one sample. This was up from 10% in a sample of narratives gathered from the period of May 2011 to February 2014. This suggests that an improvement to the user interface to PHIS or improved training could cut the number of canceled samples dramatically.

MITRE recommends that FSIS consider the following modifications to the current pull-down menu options:

- A pull-down menu options for cancellations to capture (at least) IT issues, such as: IT – Connectivity, IT – Computer, IT – Printer, IT – PHIS. FSIS might also consider requiring a brief narrative if the inspector selects PHIS. This will allow FSIS to continue to gather data on the application layer as improvements are made. FSIS may also

consider adding other reasons identified through this analysis, including: SUPPLIES and CHANGE_DATE.

- From the data that MITRE received, it appears that the inspector has a choice to select a pull-down "Not collected for miscellaneous reasons" or write a narrative about why the sample was not collected. MITRE recommends that a narrative always be required for the "miscellaneous" option. This information can then be used to periodically update the pull-down menu options.

Overall, MITRE recommends that FSIS:

- Follow up with inspectors to determine if the issue of "double-scheduling" could be fixed through modifications of the PHIS Graphical User Interface (GUI).
- Follow up with PHIS implementers and inspectors to determine if there might be a way to change the date information for a sample without canceling the sample and re-entering the data. (This issue may have already been fixed in PHIS.)
- Collaborate with MITRE to determine if the categorization script will help meet FSIS's information needs or if there is another technique that will help FSIS make sense of this important data stream as it changes. For example, advanced search techniques, clustering or corpus contrastive statistics may be sufficient to help make sense of new patterns in the data on a quarterly or semi-annual basis.

Table of Contents

List of Figures

List of Tables

1 Introduction

The Food Safety and Inspection Service (FSIS), Data Analysis and Integration Staff (DAIS), is responsible for scheduling and analyzing pathogen samples across all establishments under the jurisdiction of the U.S. Department of Agriculture (USDA). One of the strategic goals of the FSIS is to use "innovative methodologies…to protect public health efficiently and effectively."[1] DAIS tasked MITRE to determine if text processing could help identify patterns in one of its important free text data streams.

Inspectors receive pathogen and residue sampling tasks from headquarters in PHIS with a range of dates during which the samples must be taken. Inspectors then choose exact days during the range for taking the samples. As of this writing[2], there are two ways that inspectors can cancel a task:

- Type 1 Cancellation: an inspector can cancel a task before entering it in his calendar.
- Type 2 Cancellation: an inspector can cancel a task after entering it in his calendar.

When an inspector performs a Type 1 Cancellation, he or she can select a reason from a list of options in PHIS.[3] However, when an inspector performs a Type 2 Cancellation, he can select a reason from a list of options, or the inspector has the opportunity to write in a free text narrative describing why the task was cancelled. Because the free text option is currently available only when cancelling a sample after it has been entered on an inspector's calendar (Type 2), this report focuses mainly on Type 2 cancellations. We have included an appendix that offers some insight into the content of the free text narratives for Type 1 cancellations during the period from May 29, 2011 to May 20, 2012. Cancelling a task or a sample in PHIS and rescheduling it takes valuable time from inspectors' already busy schedules.

DAIS wants to identify the main classes of causes for cancellation expressed in the free text narratives, and DAIS wants to understand the frequencies of the various classes. The broader goal is to identify if improvements to PHIS may decrease the number of cancelled samples and thus increase inspectors' efficiency. A few of the inspectors' narratives express frustration with PHIS, and potential improvements to PHIS may help to alleviate this. Some examples of these types of narratives include:

- `PHIS SUCKS!!!`
- `scheduled WRONG task. Should have been Salmonella. This scheduling is VERY TEDIOUS & time consuming!!!!!`
- `PHIS scheduled 2 when I only had one available. glitch!!!`

[1] "Strategic Plan: FY 2011-2016." Available: http://www.fsis.usda.gov/wps/portal/informational/aboutfsis/strategic-planning/fy-2011-2016-strategic-plan (September 10, 2013).

[2] In recent months, a third method for cancelling tasks has become available: An inspector can cancel all remaining tasks on his calendar. The tasks remain in all tables with an appropriate "cancelled" flag.

[3] It appears that before May 20, 2012, an inspector could write in a free text narrative for the cause of cancellation. After that date, however, the vast majority of reasons (99%) for cancellation are covered by one of the options in

In Table 2-2, we present the pull-down options for Type 2 cancellations between October 1, 2012 and June 30, 2014. For Type 2 cancellations, there are separate fields in the database for pull-down responses and for free text narrative responses.

Table 2-2.

Identifying the common causes for cancellation will also enable DAIS to more closely align the pull-down menu options in PHIS with what inspectors are currently identifying as causes in their free text narratives.

MITRE recommended a two-step approach. In the first, MITRE read through a random sample of narratives and identified common classes of reasons for cancellation. MITRE then worked with DAIS to refine these initial classes. In the second step, MITRE used techniques from text mining to categorize each narrative into one of the classes of causes. This report documents the steps taken, the classes identified and the performance of the automatic categorization code. Overall, this report should be viewed as an initial step to determine if text mining might be a feasible and reliable method to gain insight into the causes of sample cancellations. At the very least, MITRE recommends collaborating with a broader group of stakeholders at FSIS to refine the classes of causes.

2 Data

With guidance from DAIS on properly joining tables in PHIS, MITRE exported data from PHIS and performed analysis on a local database. MITRE focused on two tables. The first was a table created by a join on ALGORITHMRUNRESULTS and several related tables. This table contained all assigned samples and Type 1 cancellations. The second table was derived from LABSAMPLECOLLECTIONS; MITRE only included records where the field ISREJECTED had a value of '1'. This table contained all Type 2 cancellations. Before focusing on the causes for cancellation, it will be helpful to offer general descriptive statistics of the sample collection rates and the canceled sample rates as recorded in PHIS.

In

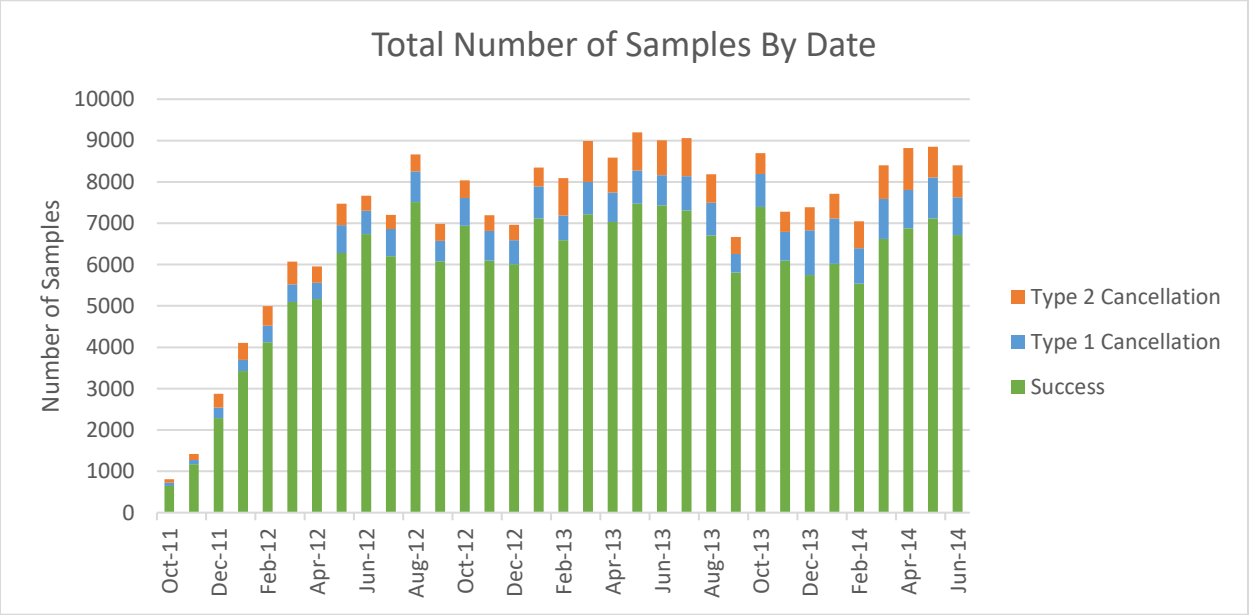


Figure 2-1, we show the number of samples by month and the number of cancelled samples by month as recorded in PHIS. Since May 2012, FSIS has recorded between 6,000 and 9,200 samples (successful and cancelled) per month in PHIS. For Type 1 cancellations, the date is the date cancelled, and for Type 2 cancellations, the date is the date that the sample was collected.

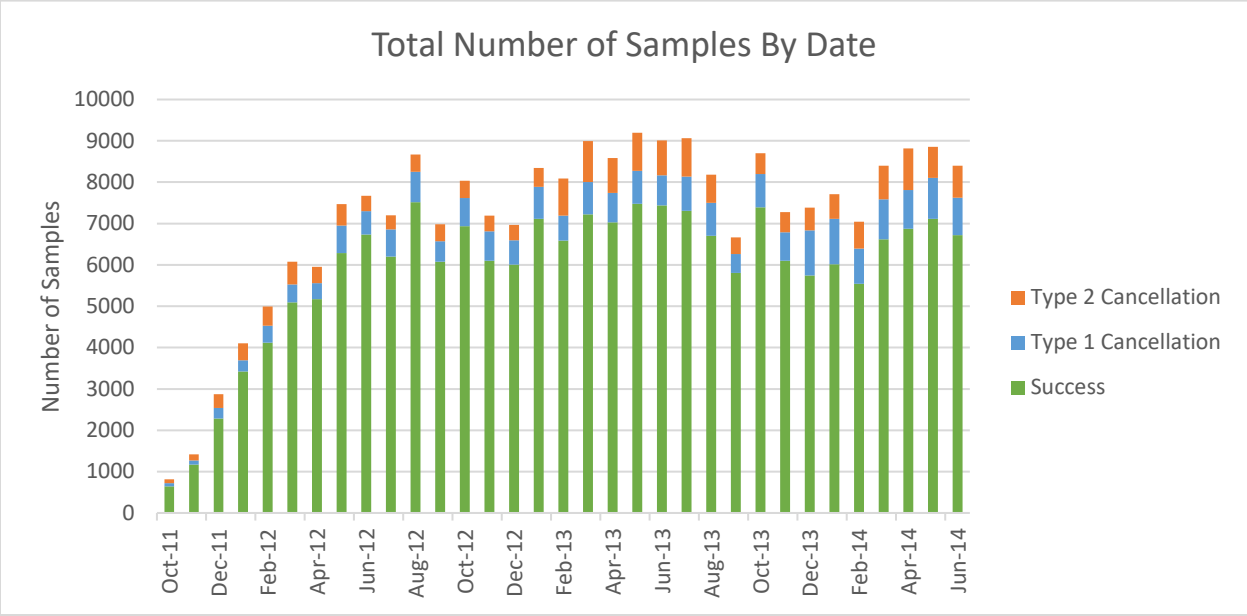


Figure 2-1: Number of Samples, Successes and Type 1 and 2 Cancellations

In Figure 2-2, we present the percentage of samples that were cancelled by type. To calculate the percentage, we divided the type of cancellation by the sum of the success samples, and the Type 1 and Type 2 cancellations per month. Since October 2011, the cancellation rate has fluctuated between 12% and 22%.

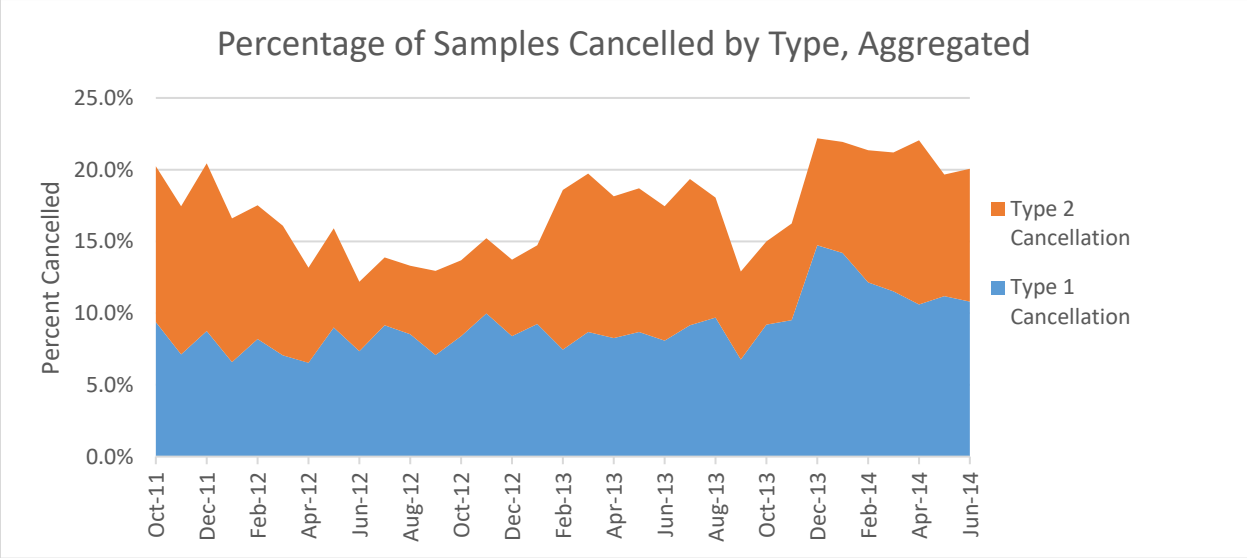


Figure 2-2: Cancelled Samples as a Percentage of All Samples, Aggregated

In Figure 2-3, we present a comparison between Type 1 and Type 2 cancellations. There appears to have been a major uptick in Type 2 cancellations in February 2013 and an uptick in Type 1 cancellations in December 2013.

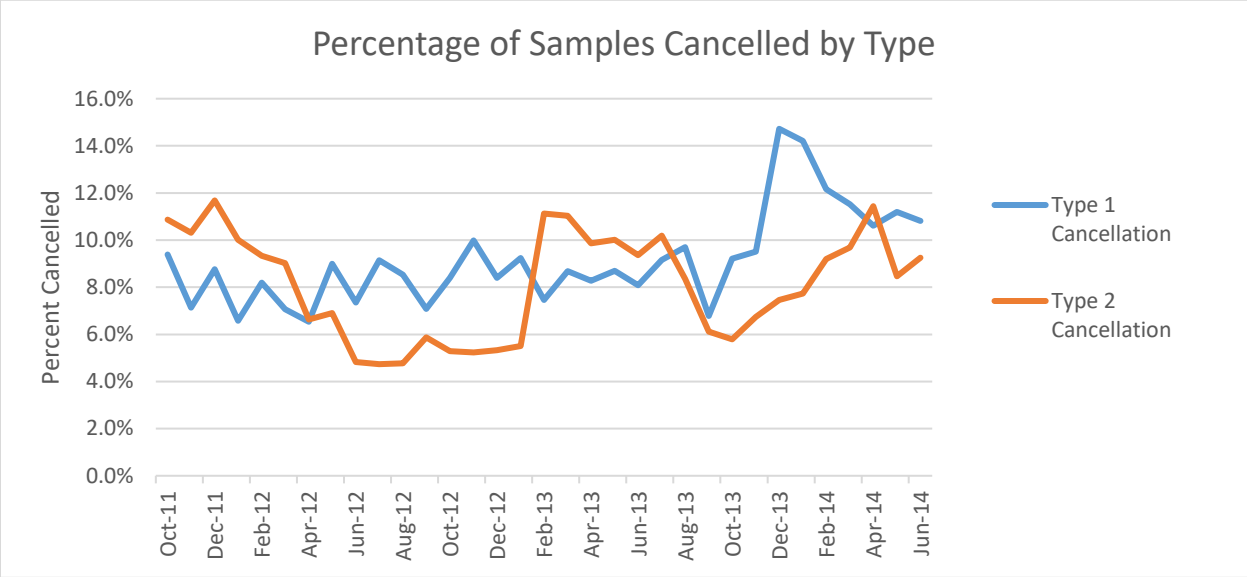


Figure 2-3: Cancelled Samples as a Percentage of All Samples

Table 2-1 shows the top 8 most common reasons for a Type 1 cancellation between May 21, 2012 and June 30, 2014. These "top 8" most likely represent the pull down menu options, the frequencies go down dramatically after May 21, 2012 for anything not in the top 8. For example the 9th most frequent option appears in only 3 records. Before May 21, 2012, there are many more options for the reason for cancellation; it is clear from the data that there was a free text option before this date.[4]

Table 2-1: Top 8 Options for Type 1 Cancellations, May 21, 2012 to June 30, 2014

Pull-down Option	Number of Records
Requested sample unavailable during sampling timeframe	8508
Requested sample/product never slaughtered/produced	5841
Not collected for miscellaneous reasons	3193
Insufficient time to collect sample	848
Plant closed/no kill	698
Inspection suspended officially	130
Inspection withdrawn	110
FedEx did not pickup sample	82

In Table 2-2, we present the pull-down options for Type 2 cancellations between October 1, 2012 and June 30, 2014. For Type 2 cancellations, there are separate fields in the database for pull-down responses and for free text narrative responses.

[4] See Appendix A for an analysis of the Type1 cancellation narratives before May 21, 2012.

Table 2-2: Pull-down Options for Type 2 Cancellations, October 1, 2012 to June 30, 2014

Pull-down Option	Number of Records
Requested sample unavailable during sampling timeframe	3917
Not collected for miscellaneous reasons	3451
Insufficient time to collect sample	1204
Requested sample/product never slaughtered/produced	1070
FedEx did not pickup sample	778
Plant closed/no kill	326
Inspection withdrawn	50
Inspection suspended officially	25

In Figure 2-4, we show the number of cancelled samples by month, subcategorized by whether the inspector used one of the pull-down options or chose to write a narrative.

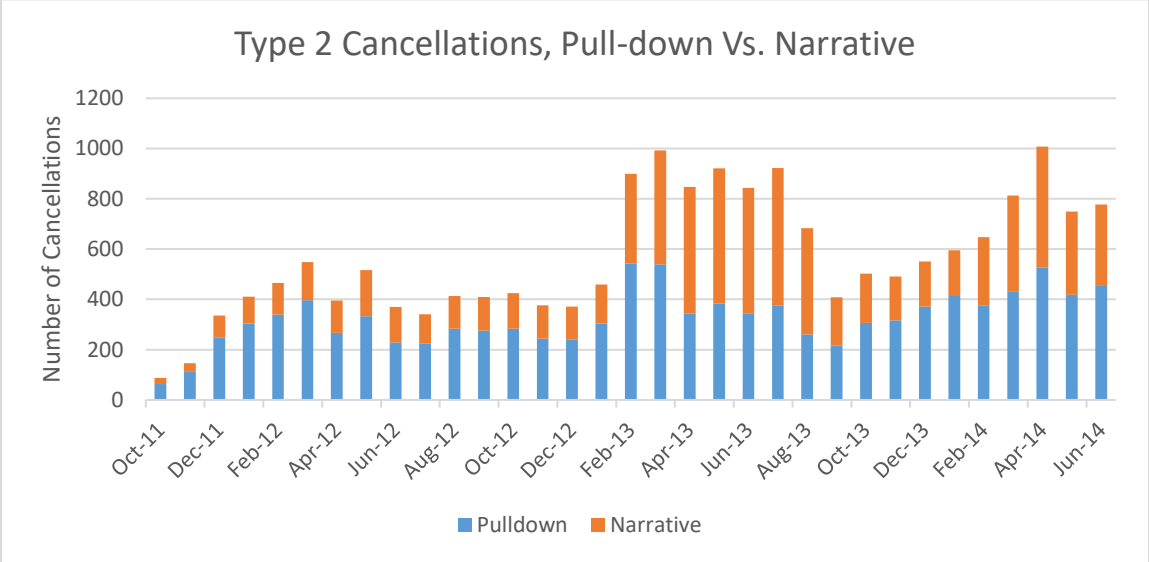


Figure 2-4: Stacked Bar Chart of Number of Cancellations by Month

As with Figure 2-2, Figure 2-4 shows an increase in Type 2 cancellations in February 2013. To gain a clearer view of the ratio of pull-down reasons and narrative reasons, we present the same data in Figure 2-5 as we did in Figure 2-4. Figure 2-5 shows that the number of narratives was generally lower than the number of pull-down reasons, but the number approached that of pull-down reasons in January 2013 and then surpassed the number of pull-down options in April 2013.

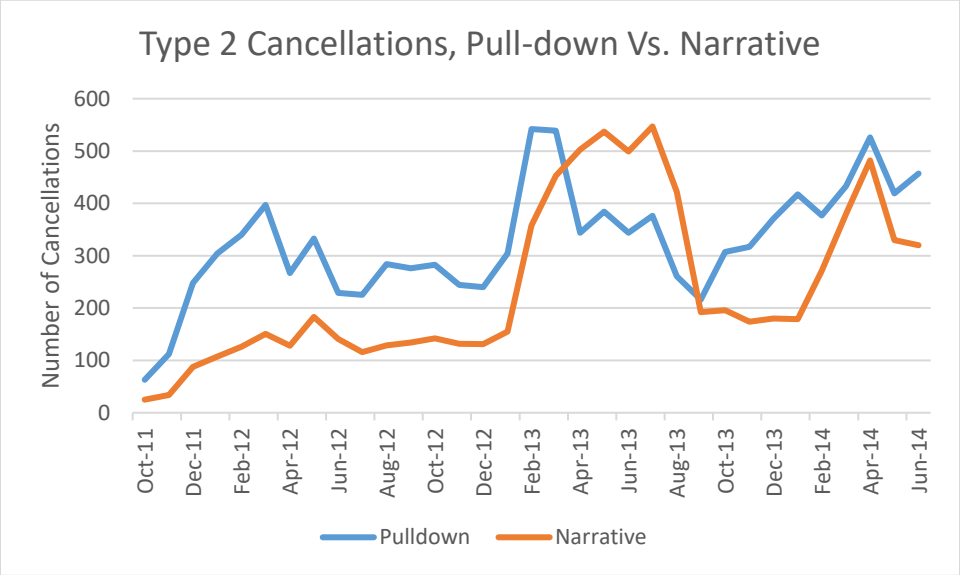


Figure 2-5: Line Chart of Number of Type 2 Cancellations by Month

Let us turn briefly now to the narrative descriptions. There were 7,960 Type 2 narrative descriptions between October 1, 2012 and June 30, 2014, and 5,994 unique narrative descriptions. There were several verbatim copies of a narrative. The narratives are short, and the median length for narratives with at least one character is 30 characters.

As a side note, we compared term frequencies in the narratives six months before February 2013 (August 2012-January 2013) with February and March 2013, and we found the following terms were much higher in the February and March timeframe than the comparison six months prior:

- `Reschedule`
- `Date`
- `Change`
- `Move`
- `Need`

These terms suggest that something may have changed in the user interface around this date having to do with changing dates.

3 Categorization Process

3.1 Overview

The first step in any categorization task is to identify the most common categories. Selecting appropriate categories can be somewhat arbitrary. Generally, the categories should be driven by the research question or the mission need. For example, if the stakeholder is concerned about staffing, there might be more granularity for staffing issues (appropriate inspector not available). If the stakeholder is interested in whether improvements to PHIS might reduce the number of cancelled samples, the categories might be binary: PHIS related or not PHIS related. For this effort, with guidance from DAIS, MITRE selected categories that are fairly general but focused on potential issues with the information technology system. We recognize that some stakeholders within FSIS would prefer other categories or perhaps more granularity on some

issues and less on others. These categories should be viewed as an initial step, and DAIS can use these categories as a first step to determine its categorization framework.

Given the scope of this subtask and the brevity of the narratives, MITRE chose to have one person develop the categories and then assign tags. In a larger categorization effort, it would be more common to have at least two people developing the categories and then tagging subsets of the data to determine how clear the category definitions were and to measure inter-annotator agreement. DAIS also provided initial categories, but MITRE chose to develop the categories independently of DAIS's categories initially and then to assess how well the categories overlapped.

3.2 Initial Categories

MITRE's first step was to create a training set of Type 2 cancellation narratives that included records with collection dates from the beginning of the data set – May 2, 2011to February 19, 2014. MITRE also created a test set of held out data from February 20, 2014 to June 30, 2014 for later testing. MITRE randomly selected 200 narratives from the training set and developed initial categories. After developing the initial categories, MITRE collaborated with DAIS to narrow the categories to those shown in Table 3-1. Finally, MITRE reviewed its initial categorizations for the 200 narratives and made small modifications based on the collaboration with DAIS.

Table 3-1: Initial Categories of Reasons for Cancellation in a Random Sample of 200 Records from May 2011 to February 2014

CATEGORY	COUNT	PERCENTAGE
CHANGE_DATE	55	28%
PRODUCT	21	11%
DOUBLE_SCHEDULED	19	10%
PHIS	19	10%
SUPPLIES	14	7%
STAFFING	13	7%
UNKNOWN	13	7%
INSUFFICIENT_TIME	10	5%
DATA_ENTRY	8	4%
CONNECTIVITY	7	4%
SHIPPING	4	2%
COLLECTION_MISHAP	4	2%
PRINTER	3	2%
COMPUTER	2	1%
WRONG_ESTABLISHMENT	2	1%
WRONG_PROJECT	2	1%
ESTAB_REFUSES	1	1%
LAB_NOT_AVAILABLE	1	1%
SUPERVISOR_DIRECTION	1	1%
WRONG_LAB	1	1%

3.3 High Level View of Categories

After further work with all of the training data, the following high-level causes emerged as candidate categories to the MITRE team. As we describe below, there can be overlap between some issues and there may not be sufficient information in the narratives to determine the cause.

- IT Issues
  - Something in the IT system (including database, user interface, connectivity or peripherals) did not work as expected
  - Headquarters' scheduling error
  - User error during data entry
- Sample Collection Issues
  - Sample not available
  - Inspector not available
  - Supplies not available
  - Sample collected with mishap
  - Supervisor direction to cancel sample
- Shipping Issues
  - Problem sending sample to lab
- Lab Issues
  - Lab not available
  - Lab rejects sample (same as sample collected with mishap?)
- Uncategorizable Issues
  - Change date
  - Unknown

When we bin the initial randomly selected set of causes from Table 3-1 into these higher categories, we find the following distribution as shown in Table 3-2. Within the timeframe of May 2, 2011 to February 19, 2014, there seems to be a roughly equal distribution of UNCATEGORIZABLE, SAMPLE_COLLECTION and IT.

Table 3-2: Higher Level Categories of Reasons for Cancellation in a Random Sample of 200 Records from May 2011 to February 2014

HIGHER LEVEL CATEGORY	COUNT	PERCENTAGE
UNCATEGORIZABLE	68	35%
SAMPLE_COLLECTION	64	32%
IT	63	31%
SHIPPING	4	2%
LAB	1	1%

Given the distributions in Table 3-2, FSIS may want to consider only three high level categories: IT, SAMPLE_COLLECTION_AND_TESTING (to include SAMPLE_COLLECTION, SHIPPING and LAB) and UNCATEGORIZABLE.

In the following sections we offer further detail of these issues.

3.3.1 IT Issues

This class of issues covers the broad spectrum of the inspectors' and headquarters' interactions with information technology generally. MITRE chose to include headquarters' scheduling errors as an IT issue because much of the scheduling is done automatically. MITRE chose to include user error during data entry as an IT issue because it might be possible to improve the user interface to reduce these errors.

In the following we include direct, unedited quotes from narratives in the `Courier` font.

3.3.1.1 Something in the IT System Did Not Work As Expected

- CONNECTIVITY
  - There was no internet connectivity or connectivity was limited.
    - `I had no internet at time of sample collection.`
    - `Sample could not be submitted,network down.`
- PRINTER
  - The printer was not working. Note that we found some examples where an inspector was probably referring to a physical printer problem, and some examples where an inspector was probably referring to the printing process via PHIS and its user interface. The current script is written to categorize the first as PRINTER and the second as PHIS.
    - `Printer died.`
    - `Program would not print out form or submit form to lab.`
- COMPUTER
  - The computer was not working. This is a general category. In the following example, the correct category might be CONNECTIVITY (if the lack of connectivity was causing the problem) or COMPUTER (if the computer was the problem preventing connection to PHIS).

    - `computer was down, unable to connect to PHIS`

- PHIS
  - An inspector reports a problem with PHIS without any specificity about user interface, connectivity, response time or any other factor of PHIS. This can overlap with "DOUBLE_SCHEDULED" if the inspector suggests that PHIS is the cause of the double scheduling. This can overlap with "STAFFING" or "INSUFFICIENT_TIME" if the inspector writes that a problem with PHIS caused delays sufficient to cancel the sample.
    - `PHIS bug`
    - `PHIS error`
    - `I'm told on one screen there is lab capacity, on another that there is none…This program was developed by morons.`
    - `PHIS tells me after thet fact the lab is at capacity!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! PLEASE FIX THIS PROGRAM !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!`
    - `PHIS error, could not get forms printed out in time for shipment.`

- DOUBLE_SCHEDULED
  - There is a fairly large number of reports in which the inspector says that a task was double (or triple) scheduled. It was not immediately clear whether the issue was caused by user error or by an error in the IT system. A few narratives suggest that at the least, improvements could be made to the user interface. In the majority of instances, the inspector says something like:
    - `Inadvertently scheduled two samples for the same day. I don't usually make mistakes of this nature, but I haven't been sleeping well. :(`
    - `When I tried to schedule this task it placed two samples on the calendar.`
    - `You tell me???? 2 tasks went to calendar when scheduling ONE!`
    - `Some how five samples was programed for the same day. Not by my doing.`
    - `Problem with software. I click to schedule the task and it doesn't appear to work. I click again with same result. After closing the task scheduler, I see that 4 sample tasks have scheduled for the same day.`

3.3.1.2 Headquarters' Scheduling Error

- WRONG_ESTABLISHMENT
  - This is a subtype of DATA_ENTRY. It is not always clear whether the source of the problem was Headquarters' assignment of the task or a data entry error by an inspector or supervisor. This overlaps with the pull-down options: a) Requested sample/product never slaughtered/produced and b) Plant closed/no kill. Examples include:

        - Mistakenly scheduled for wrong establishment by inspector not assigned to 866
        - Instructed by PHV that if the plant was on the school lunch program, FSIS does not take a Salmonella Set Sample

3.3.1.3 User Error

- DATA_ENTRY
    - This a general category for a user error while entering data. When more information is available, the more fine-grained category is used. This can overlap with WRONG_LAB and WRONG_PROJECT.
        - Input wrong information into form.
- WRONG_LAB
    - Wrong laboratory was selected. This is a subtype of DATA_ENTRY.
        - wrong lab
- WRONG_PROJECT
    - The inspector entered the wrong project in PHIS. This is a subtype of DATA_ENTRY.
        - MT-43 was placed in PHIS wrong It was to be an HC01-GB test for the 6-27-2013

3.3.2 Sample Collection Issues

3.3.2.1 Sample Not Available

- PRODUCT
  - Product was not available for testing during a given timeframe. The establishment may have stopped producing or have never produced the product altogether. They may have shifted the day of production to a day other than the one that was scheduled for sampling. The product might not be ready for sampling during the time frame. This combines three pull-down options:
    - "Plant closed/no kill"
    - "Requested sample unavailable during sampling timeframe."
    - "Requested sample/product never slaughtered/produced."
  - This can overlap with WRONG_ESTABLISHMENT in scheduling a sample for an ineligible establishment. Examples from the narratives include:
    - `The product was not available today.`
    - `Establishment no longer produces this product.`

3.3.2.2 Inspector Not Available

- STAFFING
  - In some narratives, the inspector mentions that there are not enough staff to take the sample. In others, there is notification of a reassignment of the sample to a different inspector. This category can overlap with INSUFFICIENT_TIME in cases where more staff might have enabled the inspector to have the time to take the sample.
    - `Annaul leave`
    - `Inspector off due to illness`
    - `Assigned to another inspector`
    - `working slaughter line`
- INSUFFICIENT_TIME
  - This category covers instances where an inspector reports that there wasn't enough time. This could be caused by the timing of operations in the establishment (as in the first example below and in the pull-down option "Requested sample unavailable during sampling timeframe"). This could also be caused by the inspector's lack of time to collect the sample (available as a pull-down menu option "Insufficient time to collect sample"). In the case of the latter, this category overlaps with STAFFING. This can also overlap with PHIS or other physical IT issues if those issues delayed an inspector from carrying out the task.
    - `Larger turkeys today. Did not get chilled until late and the Fed Ex cut-off time is Noon. Could not collect within that time frame.`
    - `sample not totally frozen for shipping`
    - `no time rescheuled`

3.3.2.3 Supplies Not Available

- SUPPLIES
  - The sampling supplies were not available or the supplies were otherwise not useable.
    - `No box to do the sample`
    - `Discovered that solution used to collect sample was expired. Sample rejected at establishment by SPHV`

3.3.2.4 Sample Collected with Error

- COLLECTION_MISHAP
  - Something went wrong during the taking and packaging of the sample. This can overlap with LAB_DISCARDED_SAMPLE. For example, if the inspector notices that a sample bag is leaking, he may cancel the sample (COLLECTION_MISHAP); however, if the lab notices the leak, the lab might cancel the sample (LAB_DISCARDED_SAMPLE). The only difference is in who noticed the problem.
    - `Integrity failure of the sample seal. Will re-submit.`

3.3.2.5 Supervisor Direction

- SUPERVISOR_DIRECTION
  - This could be direction from a front line supervisor, an unspecified supervisor or a district office.
    - `Supervisor Instruction`

3.3.3 Shipping Issues

- SHIPPING
  - There was a failure by Federal Express or the shipping agent to pick up or deliver the sample. Something prevented successful shipping (e.g. weather). This is very similar to the pull-down option "FedEx did not pickup sample."
    - `Sample was scheduled but FED-EX did not pick-up, sample was discarded.`
    - `Not received at lab.`

3.3.4 Lab Issues

- LAB_NOT_AVAILABLE
  - A lab is not open to receive the sample.
  - A lab is at capacity and can't process the sample.
- LAB_DISCARDED_SAMPLE
  - Lab discarded the sample. This can overlap with COLLECTION_MISHAP.
- ESTABLISHMENT_NOT_ELIGIBLE
  - The establishment is not eligible for a given sampling program.

3.3.5 Miscellaneous Issues

- CHANGE_DATE
  - The inspector chose to change a date. This covers the following types of dates (at least): slaughter date, the sample date and the parcel pickup date. There can be

many reasons for this change, and the narratives do not always make it clear what the underlying reason for the date change is. Some causes that are described include: a) Input error and the inspector can't change the date in the system; and b) He/she has to cancel the sample and start a new one with the correct date; and c) Wrong day, wrong shift
    - `Wrong slaughter date entered origionally and your stupid program wouldn't let me change the slaughter date.`
    - `Need to reschedule`
- UNKNOWN
    - This is the broad category for records that were unable to be categorized. This appears to be the same as the pull-down option "Not collected for miscellaneous reasons." Examples included:
        - `additional info`
        - `Unable to collect sample due to unspecified reason`

In addition to the categories above, some other potential candidates emerged. MITRE did not choose to include these in this first phase of categories, but we do want to document them here to help in future discussions with DAIS and FSIS.

- ESTABLISHMENT_DIRECTION
    - There are a few records that mention that the establishment asked that a sample not be taken on a given day. This is a modified name for the ESTAB_REFUSES category identified in the initial review of 200 records.
- HOLIDAY
    - It is not clear if this should be its own category or a subcategory of data entry or headquarters error; for example, if the inspector initially scheduled the sample for a holiday was that a mistake during data entry? Or, did headquarters schedule the task within a date range that was too narrow and the only time to schedule the sample was on a holiday?

4 Evaluation

4.1 Development Set

MITRE used the training set to develop rules to categorize the descriptions into one of the categories mentioned above. This was an iterative process in which we identified the most common words in uncategorized descriptions and built rules for those. For each rule, we performed basic analysis to make sure that the rules were general enough to capture the most common cases, but were not too general to mis-categorize records. As is common in this type of process, we stopped adding rules when we determined that there were no more general terms that would cover sufficient uncategorized records to justify the added complexity in rules. This is a subjective decision, and more work could be performed to add more rules.

4.2 Held Out Set

We randomly selected 200 records that were not included in the development set from 3Q FY14: April – June 2014. We manually tagged these 200 records for one of 21 categories.

In Table 4-1, we present the distributions of the truth categories in the held out test set. When we compare this with Table 3-1, we can see that DOUBLE_SCHEDULED, which was only 10% in the data from May 2011 to February 2014 is far more common (49%) in the recent data.

Table 4-1: Categories for 200 Randomly Selected Test Cases, April-June 2014

CATEGORY	COUNT	PERCENTAGE
DOUBLE_SCHEDULED	98	49%
UNKNOWN	21	11%
CHANGE_DATE	20	10%
DATA_ENTRY	15	8%
PRODUCT	10	5%
STAFFING	10	5%
PHIS	7	4%
COMPUTER	4	2%
SUPPLIES	4	2%
CONNECTIVITY	3	2%
LAB_NOT_AVAILABLE	3	2%
INSUFFICIENT_TIME	2	1%
COLLECTION_MISHAP	1	1%
SHIPPING	1	1%
WRONG_LAB	1	1%
ESTABLISHMENT_DIRECTION	0	
LAB_DISCARDED_SAMPLE	0	
PRINTER	0	
SUPERVISOR_DIRECTION	0	
WRONG_ESTABLISHMENT	0	
WRONG_PROJECT	0	

When we perform the same binning we did for Table 3-2 on the held out test data (Table 4-2), the high level distributions reflect the increase in DOUBLE_SCHEDULED. In the data from May 2011 to February 2014, the three categories – IT, UNCATEGORIZABLE and SAMPLE_COLLECTION – each represented roughly one third of the narratives. Within the most recent fiscal quarter, IT has nearly doubled to 64%, driven largely by DOUBLE_SCHEDULED.

Table 4-2: Higher Level Categories for 200 Randomly Selected Test Cases, April-June 2014

HIGHER LEVEL CATEGORY	COUNT	PERCENTAGE
IT	128	64%
UNCATEGORIZABLE	41	21%
SAMPLE_COLLECTION	27	14%
LAB	3	2%
SHIPPING	1	1%

For the purposes of this text processing and this evaluation specifically, differences between training and testing data can often lead to lower performance.

4.3 Evaluation Results

After manually tagging the test set, we ran the categorization script that was developed for the 21 categories of the training set on the test set. The script correctly identified the category in 123 cases (61.5%). The two biggest sources of error were:

1. The script categorized the record as PHIS, but the human categorized it as DOUBLE_ENTERED in 16 cases.
2. The script categorized the record as UNCATEGORIZABLE, but the human categorized it as DOUBLE_ENTERED in 14 cases.

When we reduce the categories to the higher level categories, specified in Table 4-2, the script correctly tagged 157 cases (78.5%). The three biggest sources of error were:

1. The script categorized the record as UNCATEGORIZABLE, but the human categorized it as IT in 23 cases.
2. The script categorized the record as SAMPLE_COLLECTION, but the human categorized it as UNCATEGORIZABLE in 8 cases.
3. The script categorized the record as IT, but the human categorized it as UNCATEGORIZABLE in 5 cases.

The results from both runs suggest that if DAIS chooses to apply this categorization script to new unseen data, MITRE should determine if adding rules to cover IT and SAMPLE_COLLECTION would improve performance.

5 Discussion of IT Issues and Type 2 Cancellations

When we compare the distributions of the high level categories between the training set (May 2011 to February 2014 in Table 3-2) and the testing set (April to June 2014 in Table 4-2), there is a significant change in distributions of the high level categories. This change suggests that a fairly recent change in the user interface of PHIS may be causing an increased occurrence of double-scheduling inspection tasks.

If we extrapolate from the test sample and estimate that 64% of Type 2 narratives between April and June 2014 were caused by an IT issue, we calculate that 725 of the Type 2 narratives identified IT as the reason for a cancellation. That number (725) represents 13.5% of all cancellations (5,368) in 3Q FY14, including Type 1 and Type 2 pull down cancellations. That number should be viewed as a lower bar because inspectors had the opportunity to record the cause as “Miscellaneous” for Type 2 cancellations, and this might include IT issues; further, for

Type 1 cancellations, the best option an inspector had after May 21, 2012 for categorizing an IT issue was the "Miscellaneous" category. In short, the "Miscellaneous" pull down option could be masking IT issues.

The results from running the categorization script on all of the data from 3Q FY14 are consistent with the results from the random sample shown in Table 4-2 and the findings about the limitations of the categorizer.

In Table 5-1, we present the results of running the categorizer on all of the 1,132 Type 2 cancellation narratives from the 3Q FY14.

Table 5-1: Results of the Categorization Script on 3Q FY14 Type 2 Cancellation Narratives

Higher Level Category	Category	Count	Percentage of Type 2 Narratives
IT	DOUBLE_ENTERED	292	26%
IT	PHIS	101	9%
IT	DATA_ENTRY	74	7%
IT	COMPUTER	33	3%
IT	CONNECTIVITY	6	1%
IT	WRONG_LAB	6	1%
IT	PRINTER	3	0%
IT	WRONG_ESTABLISHMENT	1	0%
LAB	LAB_NOT_AVAILABLE	25	2%
SAMPLE_COLLECTION	STAFFING	78	7%
SAMPLE_COLLECTION	SUPPLIES	41	4%
SAMPLE_COLLECTION	PRODUCT	29	3%
SAMPLE_COLLECTION	SUPERVISOR_DIRECTION	4	0%
SAMPLE_COLLECTION	ESTABLISHMENT_DIRECTION	3	0%
SAMPLE_COLLECTION	TIME_FRAME	3	0%
SAMPLE_COLLECTION	COLLECTION_MISHAP	2	0%
SAMPLE_COLLECTION	LAB_DISCARDED_SAMPLE	2	0%
SHIPPING	SHIPPING	11	1%
UNCATEGORIZABLE	UNKNOWN	311	27%
UNCATEGORIZABLE	CHANGE_DATE	107	9%

In Table 5-2, we show the counts and percentages of the higher level categories. The SAMPLE_COLLECTION, LAB and SHIPPING categories are all very close to the distributions found in the random sample of records during the same time frame (Table 4-2). In the evaluation of the performance of the categorizer, the categorizer's most common mistake was to identify IT issues as UNCATEGORIZABLE. We would therefore expect to find more UNCATEGORIZABLE and fewer IT records in the automatically tagged set than in the random sample, and this is what we find. If we apply the 64% rate found in the random sample to the total number of records (1,124), we would expect to find 725 IT records. From the evaluation, we found that the script correctly categorized 82% of IT records. We would expect the

categorizer to find 594 records (82% of 725). The difference between what it actually found (516) and what we would expect (594) suggests that there was some language shift between the training and testing data. We found through our random samples of the training and testing data that there were big differences in the distributions of reasons reported.

Table 5-2: Results of the Categorization Script on 3Q FY14 Type 2 Cancellation Narratives, Higher Level Categories

Higher Level Category	Count	Percentage of Type 2 Narratives
IT	516	46%
UNCATEGORIZABLE	418	37%
SAMPLE_COLLECTION	162	14%
LAB	25	2%
SHIPPING	11	1%

From the Table 4-2, Table 5-2 and the discussion above, we conclude:

1) In 3Q FY14, IT issues comprised roughly 64% of the Type 2 narrative reasons for cancellation. Overall, IT issues comprised at least 13.5% of all cancellations during this period.
2) The narratives for Type 2 cancellations are still in flux, and a small amount of effort should be applied to modify the script as new data comes in if FSIS would like to apply the script to new data.
3) The results of the script will likely undercount IT issues and over count UNCATEGORIZABLE issues.

6 Conclusions and Recommendations

MITRE was tasked to offer an initial set of potential categories for sample cancellations and to prototype methods for categorizing the free text explanations of the cancellation of a sample. This report represents a first step towards the overall goal, and MITRE looks forward to collaborating with FSIS to align the categories more closely with input from stakeholders. MITRE also looks forward to determining the appropriate granularity of information to meet FSIS's analytic needs for this project.

In section 3.3, MITRE offered a set of 21 categories and five high level categories. On a held-out test set, the script was able to correctly categorize 61.5% of records for the full set of categories and 76.5% or records into the high level categories.

The categorization script that MITRE developed as part of this effort should continue to perform as well on unseen data as it did on the test set as long as there are no substantial changes in the language of the justifications. For several of the categories, MITRE does not foresee a likelihood of substantial change (for example, weather may always cause issues with transport of samples to labs). However, as PHIS is modified to meet the needs and expectations of the inspectors, MITRE can foresee new topics and issues arising from those modifications to PHIS. MITRE recommends that FSIS periodically review a sample of cancellation descriptions or apply techniques from corpus linguistics to identify new issues and new categories.

In section 5, MITRE estimates that roughly 64% of Type 2 narratives and roughly 13.5% of all sample cancellations including Type 1, Type 2 pull-down and Type 2 narratives in the most

recent fiscal quarter – 3Q FY14 – may have been caused by IT issues. Given that inspectors have a "Miscellaneous" pull down option that might mask IT issues, this estimate should be viewed as a lower threshold.

MITRE recommends that FSIS consider the following modifications to the current pull-down menu options:

- Given FSIS's interest in identifying areas for improvement in the IT infrastructure, FSIS might consider adding pull-down menu options for cancellations to capture (at least) IT issues, such as: IT – Connectivity, IT – Computer, IT – Printer, IT – PHIS. FSIS might also consider requiring a brief narrative if the inspector selects PHIS. This will allow FSIS to continue to gather data on the application layer as improvements are made. FSIS may also consider adding other reasons identified through this analysis, including: SUPPLIES and CHANGE_DATE.
- From the data that MITRE received, it appears that the inspector has a choice to select a pull-down "Not collected for miscellaneous reasons" or write a narrative about why the sample was not collected. MITRE recommends that a narrative always be required for the "miscellaneous" option. This information can then be used to periodically update the pull-down mention options.

Overall, MITRE recommends that FSIS:

- Follow up with inspectors to determine if the issue of "double-scheduling" could be fixed through modifications of the PHIS GUI.
- Follow up with PHIS implementers and inspectors to determine if there might be a way to change the date information for a sample without canceling the sample and re-entering the data. (This issue may have already been fixed in PHIS.)
- Collaborate with MITRE to determine if the categorization script will help meet FSIS's information needs or if there is another technique that will help FSIS make sense of this important data stream as it changes. For example, advanced search techniques, clustering or corpus contrastive statistics may be sufficient to help make sense of new patterns in the data on a quarterly or semi-annual basis.

Appendix A: Free Text Narratives for Type 1 Cancellations before May 21, 2012.

The main report focused on narratives for Type 2 cancellations because narratives for Type 1 cancellations were no longer being collected after May 21, 2012. MITRE wanted to determine the frequency of IT issues in the early Type 1 cancellation narratives. MITRE randomly selected 200 of the 2,319 cancellation narratives dating between May 29, 2011 and May 20, 2012 and manually categorized the records. As we show in Table A-1, the majority of records align well with the existing pull-down menu options for the Type 1 cancellations. We have included the existing pull-down options in mixed case, and the new categories that we identified in the Type 2 cancellation narratives in upper case.

Table A-6-1: Reasons for Type 1 Cancellations between May 29, 2011 and May 20, 2012

Reason for Cancellation	Number of Records
Requested sample/product never slaughtered/produced	116
Requested sample unavailable during sampling timeframe	51
Plant closed/no kill	11
Insufficient time to collect sample	7
DATA_ENTRY	4
Not collected for miscellaneous reasons	4
PHIS	2
SUPERVISOR_DIRECTION	2
CONNECTIVITY	1
Inspection suspended officially	1
SUPPLIES	1

Prepared for:

Data Analysis and Integration Staff,
Office of Data Integration and Food Protection,
Food Safety and Inspection Service, USDA

DRAFT: Categorizing Ready To Eat Products

August 29, 2014

Version 1.0

MITRE
7515 Colshire Drive
McLean, VA 22102

Record of Changes

No.	Date	Reference	A=Add M=Modify D=Delete	Description of Change
1	8/29/2014	First Draft		

Executive Summary

The Data Analysis and Integration Staff (DAIS) tasked MITRE to help categorize Ready to Eat (RTE) food products to enable more effective risk assessment of these products." DAIS requested that MITRE aid this goal in two ways:

1. Categorize RTE food product samples according to DAIS's current categories
2. Propose a new categorization scheme, derived from data that allows for consistent categorization.

The RTE categorizations that are currently in DAIS's databases are based on the categories and policies active at the time the records were labeled. These categories have changed over time, and the first task is to bring the categorizations up to date with the current categories and policy.

The second task is to identify a new, consistent categorization scheme. This will allow DAIS to identify positive pathogen samples in clearly defined categories and to prioritize the collection of pathogen samples from products in those risky categories.

Assessing the pathogen risk of a food item is difficult when items with radically different names could actually have very similar properties.

To mitigate this, MITRE makes two recommendations:

1. Implement the proposed standard naming scheme for RTE products. This scheme will allow for free text entry in a predictable format that ensures all pertinent information is included.
2. Implement IT automation for the new naming scheme. This will allow for collection of information as structured data, thus facilitating more comprehensive, in-depth risk analysis.

Table of Contents

1 Introduction

The Food Safety and Inspection Service (FSIS), Data Analysis and Integration Staff (DAIS), is responsible for scheduling, collecting, and analyzing pathogen samples across all establishments under the jurisdiction of the U.S. Department of Agriculture (USDA). One of the strategic goals of the FSIS is to use "innovative methodologies…to protect public health efficiently and effectively."[1]

DAIS tasked MITRE to help categorize Ready to Eat (RTE) food products to enable DAIS's risk assessment in two different tasks. First, categorize RTE food product samples according to the current categories. Secondly, propose a new scheme, derived from data that allows for consistent categorization. A consistent categorization across all products will allow DAIS to identify the riskiest categories and to prioritize pathogen samples to products in the highest risk categories, facilitating more comprehensive, in-depth risk analysis.

DAIS provided MITRE with a combined table of 66,099 RTE pathogen sampling records from two legacy databases (Microbiological and Residue Computer Information System (MARCIS) and Performance Based Information System (PBIS)) and the current enterprise database (Public Health Information System (PHIS)). The table's columns included information about the pathogen sample and two columns to describe the sampled product. The first food description column contained a free text narrative of the product, and the second column, called SampleSource, contained a mix of fixed fielded options that were available across the timeframes covered by the databases. The available categories varied over time. In the data set MITRE received from FSIS, there were 315 unique categories. Some example categories include:

- Pork, Large Mass, Chopped & Formed
- Beef, Whole Sausage Type Product, Unpeeled
- Product-RTE
- Product-RTE-Acidified/Fermented
- Product-RTE-Fully Cooked, Meat/Nonmeat Combination

These categories are a mix of species (pork, beef), process (chopped, acidified) and product (sausage). Some of the categories are specific (Beef, Whole Sausage Type Product, Unpeeled), while others are general (Product-RTE-Fully Cooked, Meat/Nonmeat Combination). This inconsistent approach makes it difficult to assign new food items to categories and to accurately assess the risk of a category. Risk assessments are largely based on historical data, and inconsistent categories make the data difficult to analyze.

Because field inspectors have been presented with changing fixed field options over the years, there is also inconsistency in categories for the same food product name. For example, "White meat chicken salad," which appears 87 times in the dataset is categorized by different inspectors at different times as:

- Chicken, Multicomponent Products
- Chicken, Salads/Pate/Spreads
- Product-RTE
- Product-RTE-Fully Cooked, Diced/Shredded-Chicken

[1] "Strategic Plan: FY 2011-2016." Available: http://www.fsis.usda.gov/wps/portal/informational/aboutfsis/strategic-planning/fy-2011-2016-strategic-plan (September 10, 2013).

- Product-RTE-Fully Cooked, Meat/Nonmeat Combination-Chicken
- Product-RTE-Fully Cooked, Salad/Spread/Pate
- Product-RTE-Fully Cooked, Salad/Spread/Pate-Chicken
- Product-RTE-Fully Cooked, Salad/Spread/Pate-Other Poultry
- Product-RTE-Other/Miscellaneous-Chicken
- RTE Fully Cooked Salad-Spread-Pate - Chicken

MITRE's task was to categorize the RTE food products in the sample data by two major schemes and one minor scheme:

1. Species/Product
    a. Hot Dog (ProductName= any variation of Hot Dog; any variation of Corn Dog; any variation of Frank or Frankfurter; any variation of Wiener)
    b. Corn Dog (ProductName= any variation of Corn Dog)
    c. Bologna (ProductName= any variation of Bologna)
    d. Pastrami (ProductName= any variation of Pastrami)
    e. Chicken (ProductName= any variation of Chicken)
    f. Sausage (ProductName= any variation of Sausage)
    g. Ham (ProductName= any variation of Ham)
    h. Beef (ProductName= any variation of Beef)
    i. Pork (ProductName= any variation of Pork)
    j. Turkey (ProductName= any variation of Turkey)
    k. Deli Roll (ProductName= any variation of Roll; any variation of Wrap)
    l. Salad (ProductName= any variation of Salad)
    m. Pate (ProductName= any variation of Pate)
    n. Other = everything else
2. For three specific Species/Product categories from the above list (Chicken, Beef and Sausage), break into subcategories as appropriate:
        i. Ex. Chicken: breast, tenders/nuggets/strips, patties, etc.
        ii. Ex. Beef: jerky, brisket, other, etc.
        iii. Ex. Sausage: smoked, beef, pork, etc.
3. PHIS Categories
    a. No Product
    b. Product-RTE-Acidified/Fermented
    c. Product-RTE-Dried
    d. Product-RTE-Intact Whole or Parts
    e. Product-RTE-Fully Cooked, Diced/Shredded
    f. Product-RTE-Fully Cooked, Hot Dog Products
    g. Product-RTE-Fully Cooked, Meat/Nonmeat Combination
    h. Product-RTE-Fully Cooked, Patties/Nuggets
    i. Product-RTE- Fully Cooked, Salad/Spread/Pate
    j. Product-RTE- Fully Cooked, Sausage Products
    k. Product-RTE- Other Fully Cooked
    l. Product-RTE- Other/Miscellaneous
    m. Product-RTE-Salt Cured, Deli Sliced
    n. Product-RTE-Salt Cured-Other

If a row had a SampleSource value that was an existing PHIS category, we did not assign a new category. If a row had a value that was not an existing PHIS category, it was given a label from the list above by the software.

In addition to categorizing the data on the two major and one minor categorization schemes, we augmented the Product/Species scheme with additional elements that were present in the free text narratives. After analyzing the results of this categorization, the MITRE team propose a naming scheme to reduce the variation in product names and increase the amount of information they convey.

2 Data

The sample data provided by DAIS in the Excel spreadsheet included the following columns:

- FormID
- EstablishmentID
- ProjectCode
- CollectionDate
- ProductName
- SampleSource
- Result

Fifteen additional columns indicated the presence of specific categories (ham, pork, turkey, etc.).

The spreadsheet contained a total of 66,099 rows. Of these rows, 45,094 had a value that was an existing PHIS category and were not recategorized for that scheme. This leaves 21,005 rows to categorize with a PHIS label. All rows were categorized under the remaining two schemes (Species/Product, Specific Species (species/product subcategorization)).

The column ‘Result’ is an indicator of whether or not the test of the RTE food product tested positive for a pathogen. DAIS requested to use this column to determine the riskiest categories. However, with only 207 positive examples out of the 66,099 records, it is difficult to derive any significant insights from these records.

3 Categorization Process

3.1 Overview

With the categorization tasks defined, we started examining the data to determine the best method for the multiple categorization tasks.

We considered a machine learning approach; however, the requirements of this task and the nature of the data pose several challenges. First, most machine learning algorithms emit one label per input, and this task requires three labels. One solution is to combine all three labels into a single label per item, however, this would create a very large number of possible categories (196 plus Beef/Chicken/Pork subcategories), and each category would not have sufficient examples to produce a robust categorizer. Second, there is a small amount of information available to the machine learning algorithm for this task. Eighty-four percent of product names consist of one to four words (one word 5,537 (8.4%), two words 20,793 (31.5%), three words 17,880 (27.1%) and

four words 11,311 (17.1%)). Categorization of short texts is challenging without a resource to provide background information. Ultimately, we decided on a pattern-based approach.

3.2 Categorization Implementation

Without any linguistic processing, a pattern-based approach would require the creation of string-based rules for categorization. For example, a pattern might state: if the word is “turkey” or “trky” or “turkeys”, categorize the product’s species as “turkey.” There are efficient ways of representing these patterns, but each pattern would still have to be developed manually to cover the variation within the product names. MITRE examined two different methods for matching in spite of textual variances (spelling errors, naming conventions, etc.): double metaphone and vowel removal. Both methods seek to reduce the spelling variations so that the categorization rules do not require as much manual specification.

Double metaphone is an algorithm that creates a reduced representation of a word based on phonetics. For example, the words “turkey”, “turkeys” and “teriyaki” all have the double metaphone representation: “TRK”. Reduction techniques remove variance in data by reducing the size of the alphabet, which reduces the size of the vocabulary. This reduced representation is useful for finding matches where the data is noisy. Other techniques for reducing variance include: vowel removal, duplicate consonant removal and stemming.

Vowel removal is the process of removing vowels from words to produce a reduced representation of a word while duplicate consonant removal is the process of removing duplicate adjacent consonants. Stemming is a linguistic process to remove all suffixes to reveal the root word. An example of stemming: the words “view”, “review” and “reviewed” each have the representation: “view”, ”review” and “review” respectively. An example of vowel removal: the words “turkey”, “turkeys” and “teriyaki” each have the representation: “trky”, “trkys” and “tryk” respectively. An example of duplicate consonant removal together with vowel removal: the words “skipper”, “scrapped” and “skidded” each have the representation: “skpr”, “scrpd” and “skd” respectively.

Double metaphone was useful for variances due to spelling (“burrito”, “burrit**t**o”) and singular versus plural (“strips”, “strip”), however, analysis revealed problematic matches, including:

- turkey, teriyaki (TRK)
- patty, pate, patties, pudding (PT)
- angus, ounce (ANK)
- paste, pasta, pasty (PST)
- taco, dog (TK)
- fully, filling (FL)
- ham, hummus, home (HM)

One solution is to find these cases and add conditions in the rules to prevent these matches. However, these unwanted matches are an indication that double metaphone creates representations that are not appropriate for this task. As such, double metaphone was tested and discarded for this effort.

Vowel removal yields a phonetic-looking representation, but does not collapse words as aggressively as double metaphone. Below, we show the results of combining stemming, vowel removal and duplicate consonant removal for the words that were problematic for double metaphone:

- turkey, teriyaki (TRKY, TRYK)
- patty, pate, patties, pudding (PTY, PT, PTS, PDNG)
- angus, ounce (ANG, NC)
- paste, pasta, pasty (PST, PST, PSTY)
- taco, dog (TC,DG)
- fully, filling (FLY,FLNG)
- ham, hummus, home (HM, HMS, HM)

Even though vowel removal is not as aggressive as double metaphone, it still creates problem cases. This is illustrated in the cases of "paste" and "pasta" above represented with "PST" and of "ham" and "home" represented with "HM." For these situations, the only alternative is exact string match on the original text. This is used as sparingly to prevent tailoring rules to this data set.

For quantitative comparison, double metaphone yields 2,747 unique words, while with vowel removal there are 4,118 out of a total of 6,892 unique words in all 66,099 product names. With vowel removal, there will be additional rules to catch additional cases, however, there will be fewer conditions to prevent erroneous matches. This will lead to a simpler, more maintainable codebase. After some consideration of the options, MITRE chose to use the combination of stemming, vowel removal and duplicate consonant removal for this task.

4 Evaluation

Normally in a categorization effort we would split the original data into multiple sets before beginning the work. Splitting the data minimizes bias from using the entire data set in developing the categorization system. Minimizing bias is important to create a categorization system that generalizes well to unseen data. A typical approach divides data randomly into the following sets:

- Development
- Training
- Evaluation

The development set is a small portion (usually ~10% or less) that is used to develop the machinery of the categorization system. The bulk of the data (usually ~80%) is in the training set, which is used to build up and optimize the categorization decisions. Finally, the evaluation set is another small portion (usually ~10%) that is set aside and not examined throughout the development and training process. Once the categorization system is optimized on the training set, the evaluation set is used to assess how well it generalizes to unseen data.

For this effort, the MITRE team began to write the patterns prior to splitting the data. With the number of records and the large variance in the data, the MITRE team does not believe it will bias evaluation.

4.1 Development Set

MITRE used a training set to develop rules to categorize the descriptions into one of the categories mentioned above. This training set consisted of 65,791 records from the Excel spreadsheet from DAIS.

This was an iterative process in which we identified the most common words in uncategorized descriptions and built rules for those. For each rule, we performed basic analysis to make sure that the rules were general enough to capture the most common cases, but were not too general to mis-categorize records.

As is common in this type of process, we stopped adding rules when we determined that there were no more general terms that would cover sufficient uncategorized records to justify the added complexity in rules. This is a subjective decision, and more work could be performed to add more rules if additional precision or recall is required.

4.2 Evaluation Set

We randomly selected 300 records for our evaluation set that were not included in the development and training sets. The evaluation set contained nine duplicate product name entries, therefore, eight of these records were eliminated, leaving 292 records. We manually tagged these records for the PHIS categorization task, as it is the most difficult categorization task of the three.

For most annotation efforts we would have multiple subject matter experts annotate each record, and then we would adjudicate any discrepancies and merge them by hand to create the evaluation set. Due to the smaller size and scope of this effort, one annotator was sufficient. The evaluation set was distributed to subject matter experts which were then reviewed and found to be valid.

In Table 4-1, we present the distributions of the truth categories in the evaluation set.

Table 4-1: Categories for 292 Randomly Selected RTE Products

CATEGORY	COUNT	PERCENTAGE
Product-RTE-Salt Cured, Deli Sliced	56	19.18%
UNKNOWN	4	1.37%
Product-RTE-Other/Miscellaneous	6	2.05%
Product-RTE-Other Fully Cooked	19	6.51%
Product-RTE-Intact Whole or Parts	11	3.77%
Product-RTE-Fully Cooked, Sausage Products	50	17.12%
Product-RTE-Fully Cooked, Salad/Spread/Pate	7	2.40%
Product-RTE-Fully Cooked, Patties/Nuggets	22	7.53%
Product-RTE-Fully Cooked, Meat/Nonmeat Combination	45	15.41%
Product-RTE-Fully Cooked, Hot Dog Products	18	6.16%
Product-RTE-Fully Cooked, Diced/Shredded	25	8.56%
Product-RTE-Dried	16	5.48%
Product-RTE-Acidified/Fermented	13	4.45%

4.3 Evaluation Results

After manually tagging the evaluation set, we ran the categorization script that was developed for the categories of the training set on the evaluation set. The script correctly identified the category in 196 cases (67.1%) based on the current annotations in the evaluation set. The categories that are the most problematic are:

1. Product-RTE-Other Fully Cooked (20 incorrect categorizations)
2. Product-RTE-Other Fully Cooked, Diced/Shredded (14 incorrect categorizations)
3. Product-RTE-Other Fully Cooked, Meat/Nonmeat Combination (12 incorrect categorizations)

The "Other Fully Cooked" category is not well defined, and is conflated with Product-RTE-Other/Miscellaneous. These categories are used when some information is known about a product (species, preparation method, etc.) but the entry failed to be categorized by any other rule.

The "Diced/Shredded" category is easily conflated with the "salt cured, deli sliced" category. The products are similar and use similar terminology.

The results from this effort suggest that if DAIS chooses to apply this categorization script to new unseen data, MITRE should determine if adding rules would improve performance.

4.4 Evaluation Discussion

The MITRE team has examined the data throughout this task. This categorization task is difficult due to the small amount of information and the variance in the information that is available. The product names contain differing amounts of information and are underspecified in several different ways. The different ways include (exactly as it appears on the data examined):

1. Species not specified
    a. biscuit slices
    b. cheddar smoked sausage
    c. bbq strip
    d. cooked pizza topping
    e. Fully Cooked Taco Filling
2. Preparation method not specified
    a. boneless ham
    b. kalua brand pork
    c. stubb"s all natural chopped beef
    d. Ham
3. Other major components not specified for meat/nonmeat combination
    a. chicken salad
    b. chili brick
    c. meat loaf
    d. Meatball #2
    e. Russ Loaf
4. No information present in product name
    a. 1130 hours
    b. bar-b-que
    c. bbq hot
    d. bcc
    e. 10OZ KROGER WA

Some products are well specified including:

1. breakfast burrito beef, egg, hashbrowns, sausage & cheese
2. mister swifts tender mini pork hot dogs
3. chicken rosemary dinner w/potato and mixed vegetable

4. Fully Cooked Sliced Bone In Smoked Ham
5. Hickory Smoked Alligator and Pork Andoullie Sausage

Implementing a standard naming scheme for RTE products will decrease variance in product names and allow for a robust RTE product risk assessment. An alternative is to reconsider the set of categories according to current criteria and then assign each product to its new category using some matching method. As found in this effort, this method is problematic as the information may not be available to assign products to their new categories.

A more robust approach is to collect as much information as possible about each product as semi-structured data. Implementing a naming scheme allows for better information collection with a free text field, and enables reliable downstream processing. Then, as resources allow, additional IT automation could be introduced to move from semi-structured data collection via free text input with naming scheme, to full structured data input via drop down lists and checkboxes.

The proposed naming scheme is:

{brand} {preparation method} {process method} {main protein} {main food item} with {other components} in {liquid}

An explanation of the fields:

- **Brand** is the commercially known name of the finished product. Ex. Ortega, TGI Fridays; or the business name of the producer. Ex. Dietz and Watson, Smithfield Foods.
- **Preparation method** is the common name of the method used to cook the raw product. Ex. Barbecued, smoked, baked, fried, steamed, roasted.
- **Process method** is the common name of the method used to physically manipulate the product. Ex. Sliced, pulled, ground, diced, chopped.
- **Main protein** is the name of the main meat component in the product. Ex. Beef, chicken, pork, duck, ox.
- **Main food item** is the commercial name of the product. Ex. Breakfast sausage, wrap, chili, mortadella, turkey breast.
- **Other components** are a list of other major components (protein, produce) in the food item. Ex. Eggs, sausage, tomatoes, lettuce, cheese.
- **Liquid** is any fluid that is packaged with the food item for any reason such as flavor and preservation. Ex. Water, au jus, barbecue sauce, gravy.

Examples of fully qualified names under the proposed naming scheme (fabricated):

- Uncle Bob’s Smoked Pulled Pork Shoulder with pearl onions in BBQ Sauce
- Swindon Brand Cooked, Seasoned Beef Taco Meat
- Good Co. breakfast burrito with egg, cheese, peppers, onions and sausage
- ABC Kitchen Beef, Pork, Chicken Hot dogs

This naming scheme covers data that has been seen in the sample data provided to MITRE for this task. There may be cases that this scheme will not fit, and may require additional analysis.

5 Conclusions and Recommendations

MITRE was asked to perform two tasks in categorizing a set of RTE products. The first task was to categorize a set of RTE products according to two major schemes and one different scheme. The second task was to develop a set of categories based on the data. This report represents a first step towards the overall goal, and MITRE looks forward to collaborating with FSIS to align the categories more closely with input from stakeholders. MITRE also looks forward to determining the appropriate granularity of information necessary to meet FSIS's analytic needs for this project.

The difficult part of this task is the variance in the names used for these products. Some product names specify species, preparation method and other major components, while others only specify a finished food item. Assessing the pathogen risk for an item is difficult when items with radically different names could actually have very similar properties.

To mitigate this, MITRE makes two recommendations:

1. Implement the proposed standard naming scheme for RTE products. This scheme will allow for free text entry in a predictable format that ensures all pertinent information is included.
2. Implement IT automation for the new naming scheme. This will allow for collection of information as structured data, thus facilitating more comprehensive, in-depth risk analysis.

Prepared for:

Data Analysis and Integration Staff,
Office of Data Integration and Food Protection,
Food Safety and Inspection Service, USDA

DRAFT

Advanced Free Text Search to Support Analysis

September 19, 2014

Version 1.0

MITRE
7515 Colshire Drive
McLean, VA 22102

Record of

No.	Date	Reference	A=Add M=Modify D=Delete	Description of Change
1	9/19/2014	First Draft		

Executive Summary

The Data Analysis and Integration Staff (DAIS) requested MITRE support in identifying potential areas for improving analysts' efficiency through free text search. FSIS analysts are knowledge workers and need to be able to navigate through and find the relevant sections of numerous documents to accomplish their missions. Many analysts currently have no way of searching the sources of data they need, the tools available are not satisfactory or analysts are using publically available search engines and hoping that the documents they need have been published by FSIS and ingested by the public search engines.

As an initial step, MITRE delivered an advanced free text desktop search prototype for analysts to use and experiment with their own data. MITRE followed up with some analysts after a few months to inquire whether the search prototype was useful, how it helped (if it did) and what features were necessary to help analysts accomplish their tasks more quickly, such as requiring or greater depth and the ability to carry out innovative analysis that would not otherwise be possible.

This report offers a brief background and overview of search features, and it documents some information that emerged from the follow on conversations with analysts who experimented with the prototype to carry out their missions. This report should not be construed as a formal requirements gathering process, nor do we believe that the handful of analysts interviewed for this report are statistically representative of analysts at FSIS. Rather, this report identifies potential areas where free text search might help analysts accomplish their current tasks more efficiently and with greater understanding of their documents, or where advanced free text search might enable analysts to carry out innovative approaches not possible without advanced free text search.

Overall, among the analysts interviewed, MITRE found:

1) Analysts do not have sufficient tools to carry out free text search of common data sets or of team-specific batches of documents. Currently, to find mission critical information:
    a. For some use cases, analysts are opening individual documents and using "Find" to locate the information they need.
    b. Analysts are relying on google.com to find published FSIS documents.
    c. Analysts with Structured Query Language (SQL) skills and database access are performing some free text searches of records within the Public Health Information System (PHIS).
2) Microsoft Windows' integrated desktop search is not working for analysts who have tried it. Analysts do not trust the results from the AskFSIS search tool.
3) Analysts estimated that reliable search tools could speed up their performance of tasks by anywhere from a few hours to days or weeks.
4) At least one analyst performed an innovative analysis of non-compliance reports (NRs) that would not have been possible without advanced search capabilities.

MITRE recommends that FSIS consider a pilot project to demonstrate the state of the possible for advanced enterprise search of common data sets in PHIS on the FSIS network. MITRE also recommends identifying a search solution for team-specific batches of documents.

Table of Contents

List of Figures

1 Introduction

The Food Safety and Inspection Service (FSIS) is responsible for ensuring the safety of the nation's commercial supply of meat, poultry and egg products. To carry out this mission, inspectors verify that establishments are operating according to federal regulation and guidance from FSIS. However, there is also a large analytical component that supports FSIS's mission by evaluating the effectiveness of policies and performing analysis to ensure that FSIS's resources are aligned as closely as possible to the biggest threats to food safety. One of the strategic goals of the FSIS is to use "innovative methodologies…to protect public health efficiently and effectively."[1] The Data Analysis and Integration Staff (DAIS) requested MITRE support in identifying potential areas for improving analysts' efficiency through free text search.

To carry out FSIS's mission, analysts at headquarters perform analysis on fixed field data in the Public Health Information System (PHIS), but many analysts are also performing analyses on free text within documents or within text fields in PHIS. Many analysts currently have no way of searching the sources of data they need, the tools available are not satisfactory or analysts are using publically available search engines and hoping that the documents they need have been published by FSIS and ingested by the public search engines.

As an initial step to determine whether free text search might help analysts, MITRE recommended allowing analysts to experiment with a desktop search prototype and then following up with those analysts to learn more about their search needs. The search prototype, Rhapsode, was approved for use on FSIS computers in January 2014, and MITRE carried out training sessions in February and March. MITRE then followed up with analysts as needed throughout the rest of the fiscal year. Rhapsode was initially developed for analysts in the Department of Defense, and it has been found to aid analysts on some tasks dramatically. At the end of fiscal 2014, MITRE briefly interviewed analysts to inquire about the utility of search. Specifically, in these conversations, we focused on: data sources, query features, results, advanced search tools beyond basic search and overall utility of search for analysts' tasks.

The goal of this effort is not to transfer Rhapsode to FSIS or to make its use more widespread within FSIS. Rather, the goal of this effort is to help guide FSIS to design tools or develop requirements for tools that will help analysts. This report represents a small effort, and more work, including a pilot, may help further define the requirements for search capabilities at FSIS.

2 Quick Overview of Search

This report focuses on two main categories of search: enterprise search and desktop search. We assume that internet search is already handled by commercial websites, such as Google.[2] The differences between enterprise and desktop search are important, and we highlight a few of the key differences below.

[1] "Strategic Plan: FY 2011-2016." Available: http://www.fsis.usda.gov/wps/portal/informational/aboutfsis/strategic-planning/fy-2011-2016-strategic-plan (September 10, 2013).

[2] As a side note, as of this writing, FSIS appears to be aggregating results from Bing and USA Search when users search FSIS's publically available website.

An enterprise search system is typically set up and maintained on dedicated servers. A developer or the search team is responsible for scheduling the indexing of new documents, server backups and maintenance, and users typically interact with these systems through web-browsers. The information technology (IT) department typically controls which sources of data are indexed, and the user interface is often quite general. Enterprise search should scale well to large volumes of data, and enterprise search systems should be able to handle multiple queries. Most of the processing is handled by servers, not the user's computer.

For desktop search, a program might be integrated (as with Windows 7) or there may be a standalone program. For the standalone program, the user is required to gather the documents, build the index and then run the search program. Users who are used to enterprise and internet search may not be willing to take the steps of gathering documents and building an index. Finally, unlike enterprise search, desktop search uses and is limited by the desktop/laptop's processing resources (central processing units (CPUS)). Desktop search tools are typically not meant to scale beyond a few million documents. Desktops/laptops with few CPUs may not be able to index that many documents efficiently.

In the following sections, we turn to some key considerations for search systems. In addition to the traditional requirements for information technology systems – for example: up time, response time, redundancy, recoverability and security – search systems have some specific basic areas that should be considered in developing or procuring a system that will meet user needs. In our follow up phase, these were some of the areas where we focused our questions. Different users may have very different use cases, and it is helpful to define some high level commonalities.

2.1 Data Sources

Different analysts rely on different data sources. Organizations often have common data sources on which many analysts rely, and then there are pockets of information sources that may only be used by one group or by a few individuals. Organizations can benefit from having both an enterprise search system for the common sources and a desktop application for those analysts whose data sources are not handled by the enterprise.

2.2 Query Capabilities

Search engines differ in the how analysts are able to create queries. Some differences might include:

1) Complexity of free text search, including:
    a. Key term – find documents that contain "salmonella"
    b. Phrase (proximity search) – require one or more words near each other, such "salmonella action plan"
    c. Fuzzy term – find documents that contain "salmonella" or any term that is within one key stroke of "salmonella" such as "salmonella"
    d. Boolean logic – require both "salmonella" and "portungal" in the same document
    e. Complex combinations of the above
2) Fields/Facets available for search. Search engines can allow users to search on multiple fields (either as free text or by faceted search). For example, one might want "Portugal" in the document's title field and "Salmonella" in the document's body field (fielded search), or an analyst might want to be able to search for non-compliance reports from a specific district or state or from establishments of a specific size (faceted search). The

fields available for search depend on the nature of the data. For many enterprise systems that include data with different fields, this level of complexity does not make sense. However, for some analytic needs on common data sets, it might make sense to allow analysts to search with these added fields.

3) Temporal Queries – if a document has a notion of a date, the user can subset the results by a date range. Non-Compliance reports have an obvious date associated with them; for Directives, the date might be the date that it became active; for Notices there may not be a clear single date (although one could imagine indexing the date range during which the Notice was effective).
4) Geospatial Queries – as with dates, many data-streams have no geospatial component. Non-Compliance reports can be linked to an establishment, and PHIS and DAIS have tables of geo-spatial queries.

2.3 Presentation of Search Results

Another major component of a search system is the user interface and how search results are presented to the user. There has been much work on designing user interfaces. The coverage here focuses only on very high level considerations. The design choices should be driven by the nature of the data and the nature of analysts' use cases. For some use cases, document titles are sufficient return values; for others, snippets of matching text might be sufficient; for other use cases, analysts might want to see every time a term or phrase appears, not just the first few (or most important) times that a term appears in a document. Search engines might allow geospatial views of data or they might offer other high level summaries of the data (as in dashboards). Finally, users may need to download or export the original documents. For some analysts, clicking on a document at a time will be sufficient, for others, downloading all results in a large zip file is desired.

In Appendix A, we include some screen shots for a user interface that we tailored for the search of Non-Compliance Reports (NRs).

3 Rhapsode

Rhapsode is a desktop text search prototype that MITRE initially developed for analysts at the Department of Defense. Prototypes have benefits, including ease of development and flexibility. However, prototypes also have limitations. The code is not as robust as commercial code, MITRE offers no support for the prototype, and the user interface and the workflow is not as elegant as one would hope to find in commercial code. However, MITRE has found this prototype to be useful in several operational areas to help identify the types of capabilities that analysts need to carry out their missions.

Rhapsode relies on several open source software projects for the main components, including Apache Lucene, Apache Tika and Jetty. MITRE has developed extra features to these components and is currently in the process of transferring these components back to the open source software community so that all of our sponsors can easily benefit. The key additional components include:

1) An advanced query parser that allows for very complex queries.
2) Concordance search.
3) Term co-occurrence counts.

There are four primary tools within Rhapsode. The first is "Basic Search." This returns basic snippets for the first few times that a term appears in documents. Document results are sorted based on relevance.

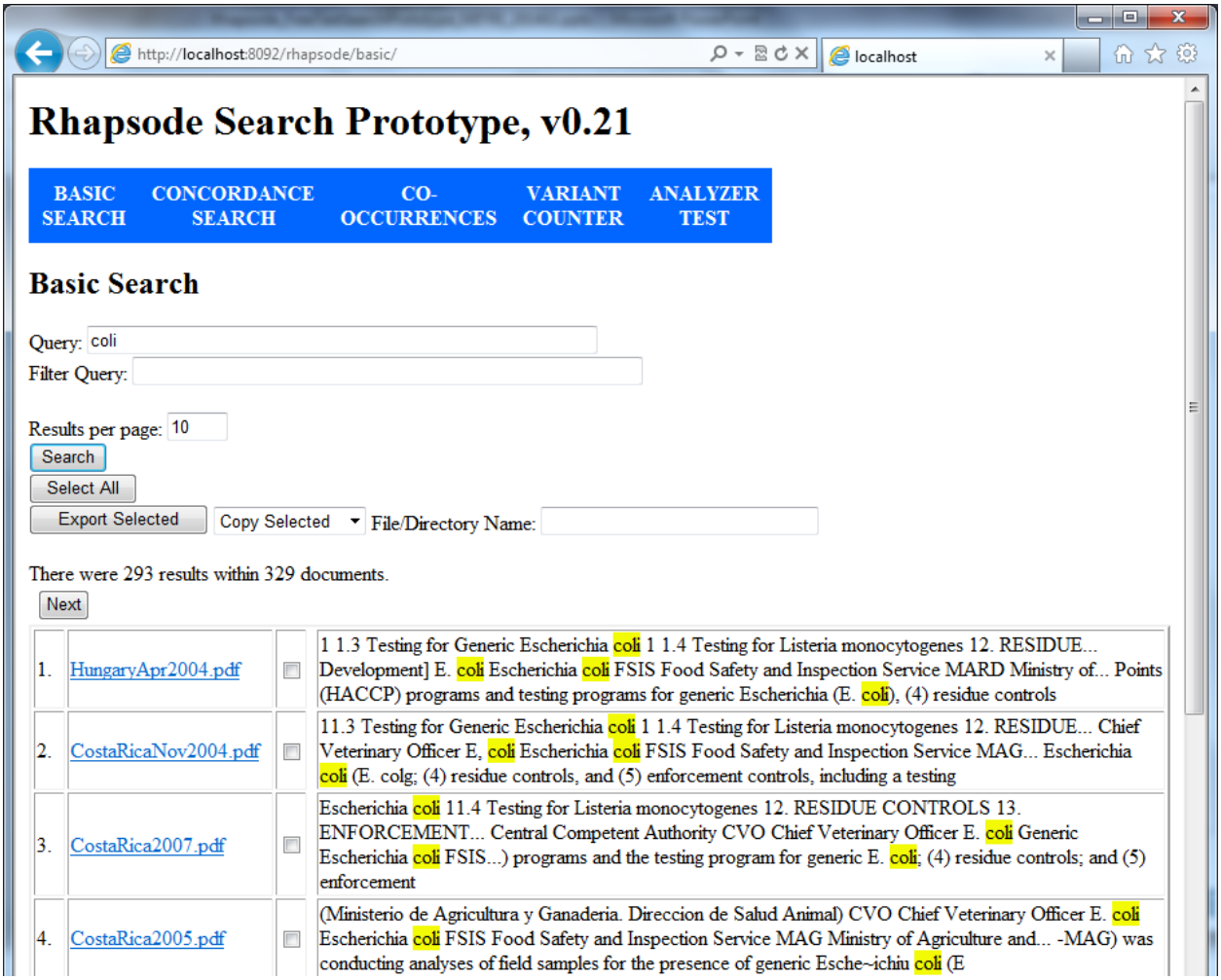
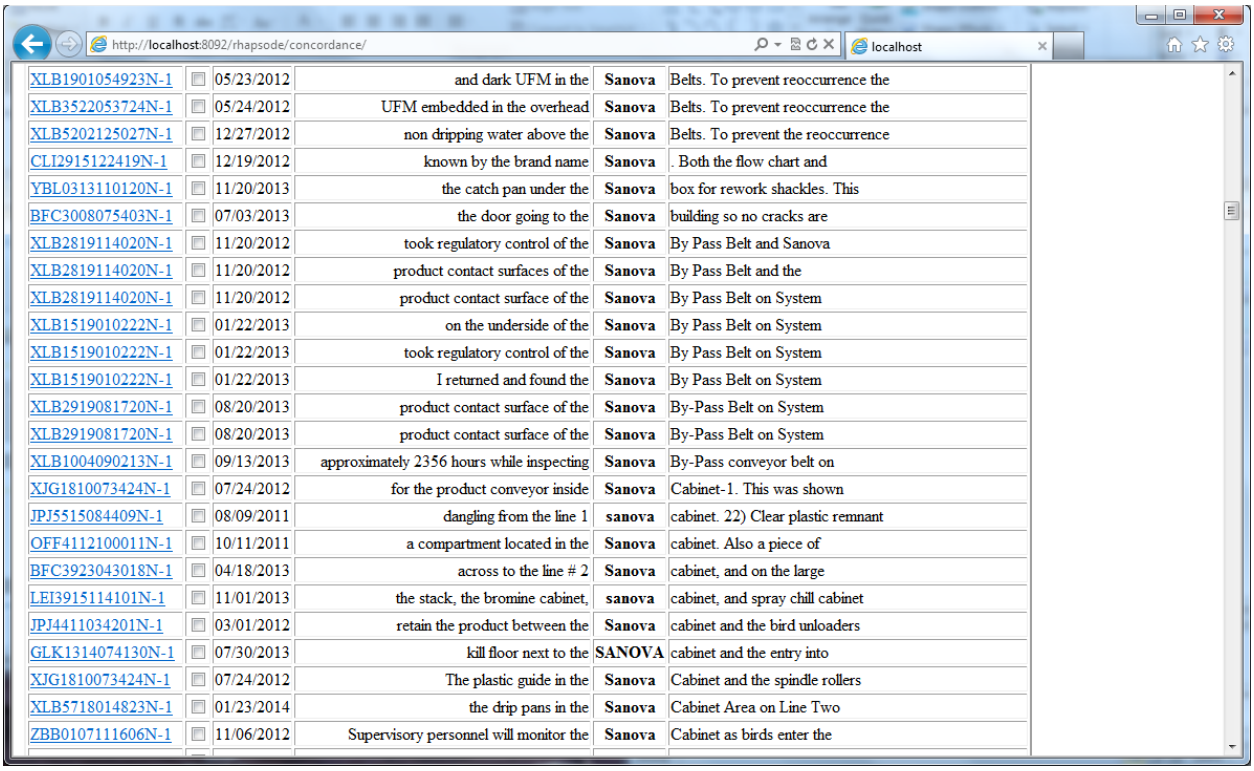


Figure 1: Basic Search Example in Rhapsode

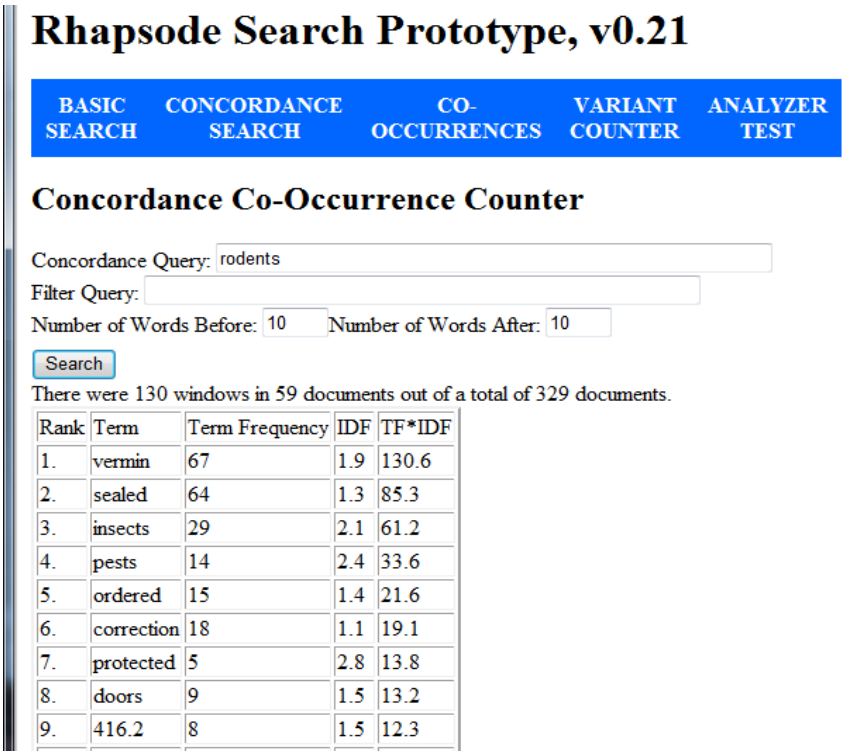
Concordance search (Figure 2) shows every time that a term appears, and there are several sort options to help analysts find patterns in words that appear before and after the target term. Figure 2 shows a few of the results for searching for the brand name “Sanova” in a collection of NRs. On this search, the results are sorted by words to the right of the key term: when sorting alphabetically by words after the key term, “belts” appears before “both” which appears before “box.” An analyst can quickly see that there are at least two different Sanova components that appear in NRS: a by-pass belt and a cabinet.



XLB1901054923N-1		05/23/2012	and dark UFM in the	**Sanova**	Belts. To prevent reoccurrence the
XLB3522053724N-1		05/24/2012	UFM embedded in the overhead	**Sanova**	Belts. To prevent reoccurrence the
XLB5202125027N-1		12/27/2012	non dripping water above the	**Sanova**	Belts. To prevent the reoccurrence
CLI2915122419N-1		12/19/2012	known by the brand name	**Sanova**	. Both the flow chart and
YBL0313110120N-1		11/20/2013	the catch pan under the	**Sanova**	box for rework shackles. This
BFC3008075403N-1		07/03/2013	the door going to the	**Sanova**	building so no cracks are
XLB2819114020N-1		11/20/2012	took regulatory control of the	**Sanova**	By Pass Belt and Sanova
XLB2819114020N-1		11/20/2012	product contact surfaces of the	**Sanova**	By Pass Belt and the
XLB2819114020N-1		11/20/2012	product contact surface of the	**Sanova**	By Pass Belt on System
XLB1519010222N-1		01/22/2013	on the underside of the	**Sanova**	By Pass Belt on System
XLB1519010222N-1		01/22/2013	took regulatory control of the	**Sanova**	By Pass Belt on System
XLB1519010222N-1		01/22/2013	I returned and found the	**Sanova**	By Pass Belt on System
XLB2919081720N-1		08/20/2013	product contact surface of the	**Sanova**	By-Pass Belt on System
XLB2919081720N-1		08/20/2013	product contact surface of the	**Sanova**	By-Pass Belt on System
XLB1004090213N-1		09/13/2013	approximately 2356 hours while inspecting	**Sanova**	By-Pass conveyor belt on
XJG1810073424N-1		07/24/2012	for the product conveyor inside	**Sanova**	Cabinet-1. This was shown
JPJ5515084409N-1		08/09/2011	dangling from the line 1	**sanova**	cabinet. 22) Clear plastic remnant
OFF4112100011N-1		10/11/2011	a compartment located in the	**Sanova**	cabinet. Also a piece of
BFC3923043018N-1		04/18/2013	across to the line # 2	**Sanova**	cabinet, and on the large
LEI3915114101N-1		11/01/2013	the stack, the bromine cabinet,	**sanova**	cabinet, and spray chill cabinet
JPJ4411034201N-1		03/01/2012	retain the product between the	**Sanova**	cabinet and the bird unloaders
GLK1314074130N-1		07/30/2013	kill floor next to the	**SANOVA**	cabinet and the entry into
XJG1810073424N-1		07/24/2012	The plastic guide in the	**Sanova**	Cabinet and the spindle rollers
XLB5718014823N-1		01/23/2014	the drip pans in the	**Sanova**	Cabinet Area on Line Two
ZBB0107111606N-1		11/06/2012	Supervisory personnel will monitor the	**Sanova**	Cabinet as birds enter the

Figure 2: Concordance Search Example in Rhapsode

There is a "Co-Occurrence" tool that runs the concordance tool in the background and calculates terms that statistically interestingly co-occur with the query term (Figure 3). This is similar to recommender systems – "If you like this book, you might also like this book." However, this tool focuses on terms and returns those that often appear near the target query. Figure 3 shows the results of searching for "rodents." Terms that statistically interestingly co-occur with "rodents" include "vermin", "insects" and "pests."



Rhapsode Search Prototype, v0.21

BASIC SEARCH | CONCORDANCE SEARCH | CO-OCCURRENCES | VARIANT COUNTER | ANALYZER TEST

Concordance Co-Occurrence Counter

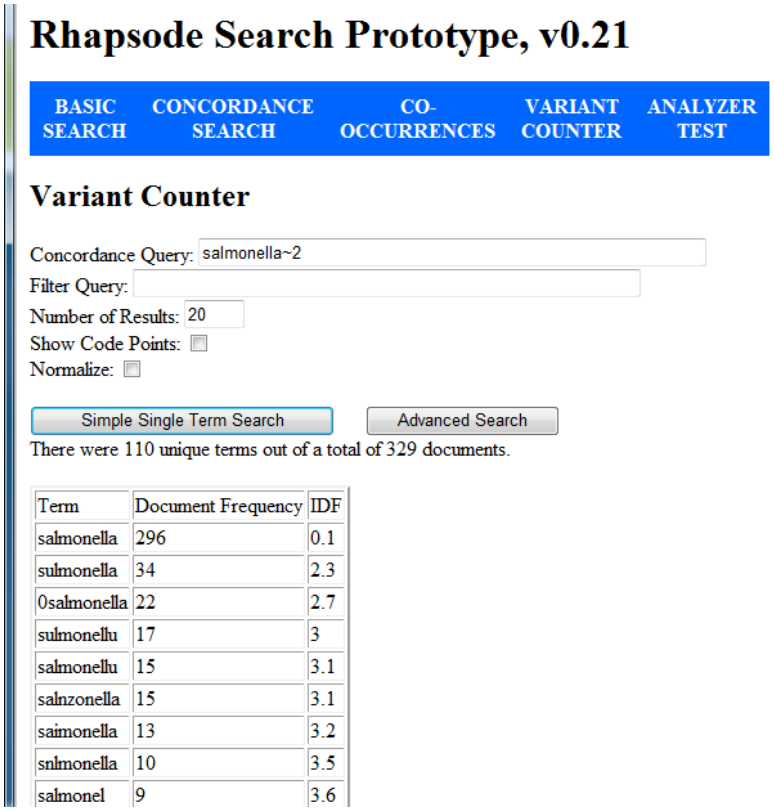
Concordance Query: rodents
Filter Query:
Number of Words Before: 10 Number of Words After: 10
Search

There were 130 windows in 59 documents out of a total of 329 documents.

Rank	Term	Term Frequency	IDF	TF*IDF
1.	vermin	67	1.9	130.6
2.	sealed	64	1.3	85.3
3.	insects	29	2.1	61.2
4.	pests	14	2.4	33.6
5.	ordered	15	1.4	21.6
6.	correction	18	1.1	19.1
7.	protected	5	2.8	13.8
8.	doors	9	1.5	13.2
9.	416.2	8	1.5	12.3

Figure 3: Example of Co-Occurrence Counter in Rhapsode

There is also a "Variant Counter" (Figure 4). This allows analysts to experiment with different query parameters to understand spelling variation and how to compose a query precisely to return the required documents. Figure 4 shows the results for searching for every term that is within two keystrokes difference of "salmonella" in a collection of foreign audit reports.



Rhapsode Search Prototype, v0.21

BASIC SEARCH | CONCORDANCE SEARCH | CO-OCCURRENCES | VARIANT COUNTER | ANALYZER TEST

Variant Counter

Concordance Query: salmonella~2
Filter Query:
Number of Results: 20
Show Code Points:
Normalize:

Simple Single Term Search | Advanced Search

There were 110 unique terms out of a total of 329 documents.

Term	Document Frequency	IDF
salmonella	296	0.1
sulmonella	34	2.3
0salmonella	22	2.7
sulmonellu	17	3
salmonellu	15	3.1
salnzonella	15	3.1
saimonella	13	3.2
snlmonella	10	3.5
salmonel	9	3.6

Figure 4: Variant Counter Example in Rhapsode

For geo-tagged data, Rhapsode allows users to export selected records as points in Keyhole Markup Language (KML) that can be visualized in Google Earth. Figure 5 shows the results for a query that included a free text component and a geographical search.

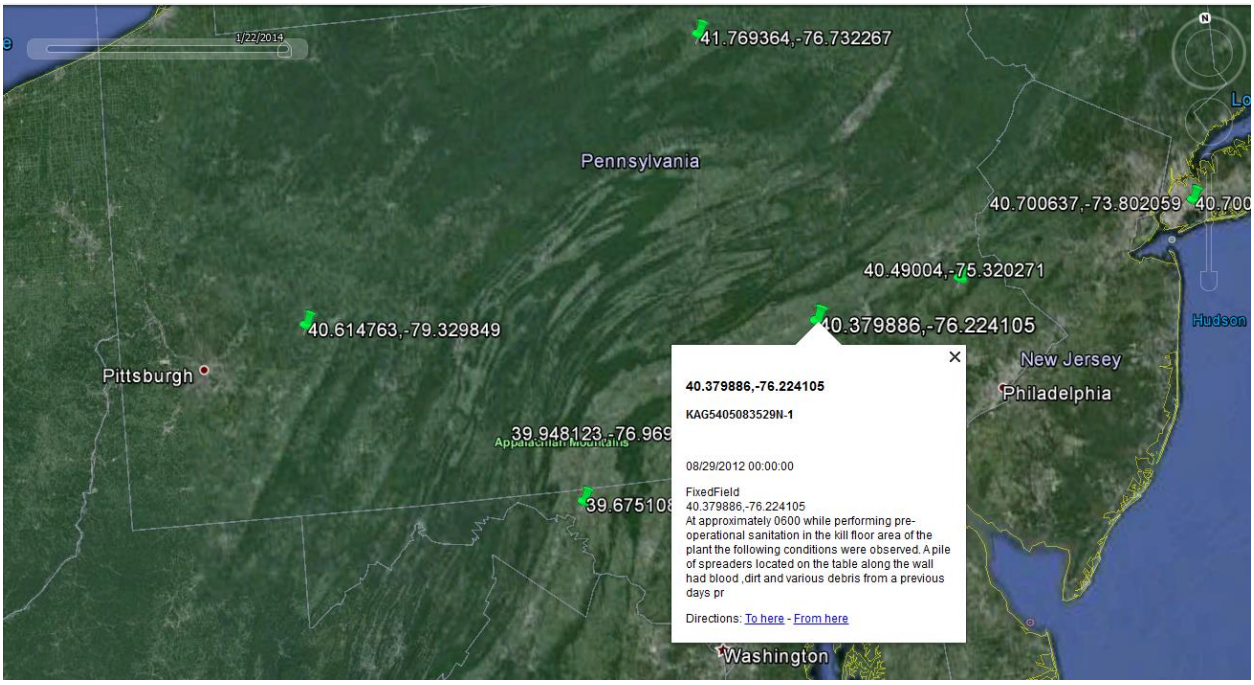


Figure 5: Example of Export NRs to KML for Google Earth

An alternate format represents the number of documents that were returned by a query as columns; the higher the column in a given geographical reason, the more documents were returned for that query. Figure 6 shows the results for searching for "rodents" in NRs. One can use these visualizations to detect geographic patterns.

Figure 6: Example of Export NRs to Columns in Google Earth

MITRE delivered a user's guide for Rhapsode along with the original executable code delivery, and we encourage interested readers to consult that document for further information on Rhapsode.

4 Results from Discussions

MITRE interviewed five analysts after they had had a chance to experiment with the prototype. One of the analysts did not use the prototype for a task, but she had important feedback on desired features of search. The staffs represented by these analysts include:

1) Data Analysis and Integration Staff, Office of Data Integration and Food Protection
2) Food Defense Assessment Staff, Office of Data Integration and Food Protection
3) Policy Analysis Staff, Office of Policy and Program Development
4) Risk, Innovations, and Management Staff, Office of Policy and Program Development
5) International Audit Staff, Office of Investigation, Enforcement, and Audit

The conversations focused on five areas – the first three questions align with the three items identified in sections 2.1 to 2.3.

1) What data sources do you use?
2) What query capabilities do you need?
3) What types of output or modifications to the presentation of results would be useful?
4) Which of the tools within Rhapsode did you use/need?
5) Did free text search help in your tasks vs. current practice?

4.1 Data Sources

As described in section 2.1, there are two primary types of data sources used by analysts at FSIS. The first included common data sources that analysts across several teams or staffs might need to search. These included:

1) Non-Compliance Reports
2) Policy documents, (active and inactive/historical) including:
    a. Notices
    b. Directives
    c. Guidelines
    d. Federal Register Notices
3) FSAs (historical full text and question-by-question as is now possible in PHIS).
4) AskFSIS questions and answers

The second type of data sources included batches of documents that only a single team or a few people might need to search. These included common office formats: PDF and Microsoft Office formats (Word, PowerPoint, Excel, Outlook). In one use case, an analyst took over responsibility for determining equivalence for a foreign country, and he received 3,400 documents that his colleague had gathered for this country. In another use case, a team began gathering all documents that they cited in their reports so that they can easily search these as they perform research for future reports.

Rhapsode has two basic indexing modes: the primary mode indexes a directory of documents, and the secondary mode indexes a table file (e.g. comma-delimited (CSV) file) and treats each row as an individual document. This secondary use case can be useful for search data that is exported from a database (NRs, for example). Most users had used Rhapsode in the primary mode, and two users had used it to index comma-delimited files. One user recommended improving the user interface for the table file indexer.

In the training sessions, a few analysts expressed frustration with the complexity and burden of gathering the documents and having to build an index before searching it. For search to be useful for this group of users, a dedicated expert or team would need to manage the input documents and update the index in a timely fashion.

4.2 Query Capabilities

For some of the users, exact term match and exact phrase match was sufficient. Some users, however, reported making use of the fuzzy term search and the more advanced phrasal searches. For data streams that have obvious dates (NRs, AskFSIS), analysts reported that they often would like to subset their queries by specific date ranges; one analyst had configured Rhapsode to allow for date searches of NRs. No analyst had yet used the geo-query capability, but several thought this might be useful.

At least two analysts recommended allowing users to search more fields than just the date field, the geo-coordinate field(s) and the text field. For example, one analyst wanted to be able to search for specific terms within NRs for a specific district during a specific date range. This functionality could be accomplished via a complex set of query boxes or via faceted search. This type of complex search interface would likely need to be tailored for each data source (for example, Notices do not normally have a "district" field), but this type of capability would allow for very powerful views into FSIS's data.

4.3 Format of Output/Display of Results

Analysts at other agencies have expressed a rich set of desires for the types of presentation of results and the output of search systems. Some of these include: the ability to see summaries of the data in a dashboard format or via facets; the ability to visualize the results geo-spatially or temporally; the ability to download a zip file of all relevant documents.

The FSIS analysts interviewed for this project were content with the capabilities available within Rhapsode.

4.4 Use of Advanced Search Tools

Several users found the BasicSearch tool sufficient. Some users found the Concordance and Co-Occurrence tools to be necessary to carry out their tasks. Overall, analysts thought the added features beyond basic search were useful.

One analyst recommended adding functionality to help identify trends in data. Specifically, for analysts of AskFSIS questions, it would be helpful to have a tool that recommends terms that are trending in questions. This would help the analysts add new questions (and answers) to the AskFSIS Frequently Asked Questions list. This may also help indicate areas for extra training or more guidance. This functionality would also allow analysts to find trends in non-compliance reports.

4.5 Current Methodology vs. Search-Enabled Analysis

Analysts identified four methods they currently use to navigate through batches of documents and reports:

1) Analysts open each document of interest and either perform a search ("Find" or "Control-f") within the document or read the document.
2) For documents that have been published by FSIS and indexed by Google, many users use the main Google website, not FSIS's internal enterprise search application.
3) For AskFSIS questions, there is a custom application that allows users some rudimentary search capabilities.
4) For advanced users who are SQL-savvy, the use of SQL syntax to search database records has been the primary method for search.

One analyst observed that Microsoft Windows has a built-in desktop search feature, but he did not think that the Microsoft Windows indexer actually worked on his documents. The source of this failure was not clear: indexer not indexing documents in a timely fashion; indexer prevented from indexing documents because of security configuration; indexer failing in another way.

The analyst who works with the AskFSIS search tool noted that it did not meet users' needs because it was not clear why the tool returned the documents that it did given users' queries.

Analysts who use Google to search FSIS documents that are in the public domain noted that it was difficult to ensure that Google was returning the most recent version of a policy document.

With the above points defining the state of the practice, users found search to be quite helpful for their tasks. Most users did not need search on a daily basis, but they found it to be useful for specific projects. On one end of the spectrum, one user noted that search of some large, multiply embedded documents helped him accomplish his task somewhat more swiftly – he did not have to open up attachments within attachments and then use "Find" within the target document. Another user estimated that a search capability saved days or weeks for her task. On the other end of the spectrum, a user carried out a novel type of analysis that he said he would not have tried without advanced search; his analysis of free text (enabled by search) helped shed light on the effectiveness of a recent policy decision.

Three of the five analysts found Rhapsode useful enough to recommend it to colleagues, and they passed Rhapsode to at least five other colleagues. As of this writing, MITRE has not yet had a chance to follow up with these other analysts.

5 Summary and Conclusions

For many analysts, free text search can be an important enabling tool. Analysts estimated modest time savings to saving days or weeks of effort to accomplish their tasks. Other analysts stated that they could carry out innovative types of analysis that they would not have considered without an advanced free text search capability. While this report represents the views of a handful of analysts across three different services and the results were not intended to be generalized, one can obtain a glimpse of the need for free text search capabilities through these analysts.

MITRE identified a need for both enterprise search systems for common data sets and desktop search tools for team or individual batches of documents. MITRE found that some analysts were unwilling or not interested in using a desktop search tool because of the extra burden placed on the user to gather the documents and keep the index up to date; these users may well be served by enterprise search. MITRE found that basic search capabilities were enough for some users and that other users needed more advanced search features including a complex query capability and the ability to search by facets or by dates. MITRE found that current search offerings hosted by FSIS – desktop indexing via the built-in Windows search functionality, internal enterprise search via Bing, and the AskFSIS search component – were not meeting users' needs.

MITRE recommends that FSIS consider a pilot search system to make access to common data sources within PHIS easier for users who lack SQL skills or who need a more advanced text search capability. At the same time, MITRE recommends that FSIS consider identifying and procuring desktop search tools to allow teams to search their local documents. Any acquisition or development choice has to be ranked among other priorities, but it appears that robust search tools could greatly help analysts accomplish their missions.

Appendix A – NR Search Prototype

In addition to delivering the Rhapsode desktop search prototype, MITRE developed a specific prototype for an NR search tool. While this was a working prototype, it did not reach a state of maturity sufficient for transfer to FSIS in fiscal year 2014. The goal of this was to demonstrate to users the state of the possible and obtain feedback based on the demonstration. Depending on FSIS's level of interest, MITRE may continue work on the NR search prototype or at least transfer feedback to FSIS's developers so that FSIS can deploy one on their own.

Unlike Rhapsode, which offers very few query options, the NR Search Prototype is tailored to NR data and offers a much more complex method of querying data. The drawback to data-specific user-interfaces is that they cannot easily be applied to new data streams. However, if there is a need, data-specific user-interfaces can be extremely useful.

Figures 7 through 9 include a facets or dashboard view of search results to give a quick overview of where and when NRs of interest are occurring, and Figure 10 shows a "documents" level view of hits. Both interfaces are driven by the same search engine and the same query window.

Figure 7 shows the results for all NRs that were loaded into the prototype. The pane in the bottom right shows that these NRs cover a period from some of 2011 to the beginning of 2014.

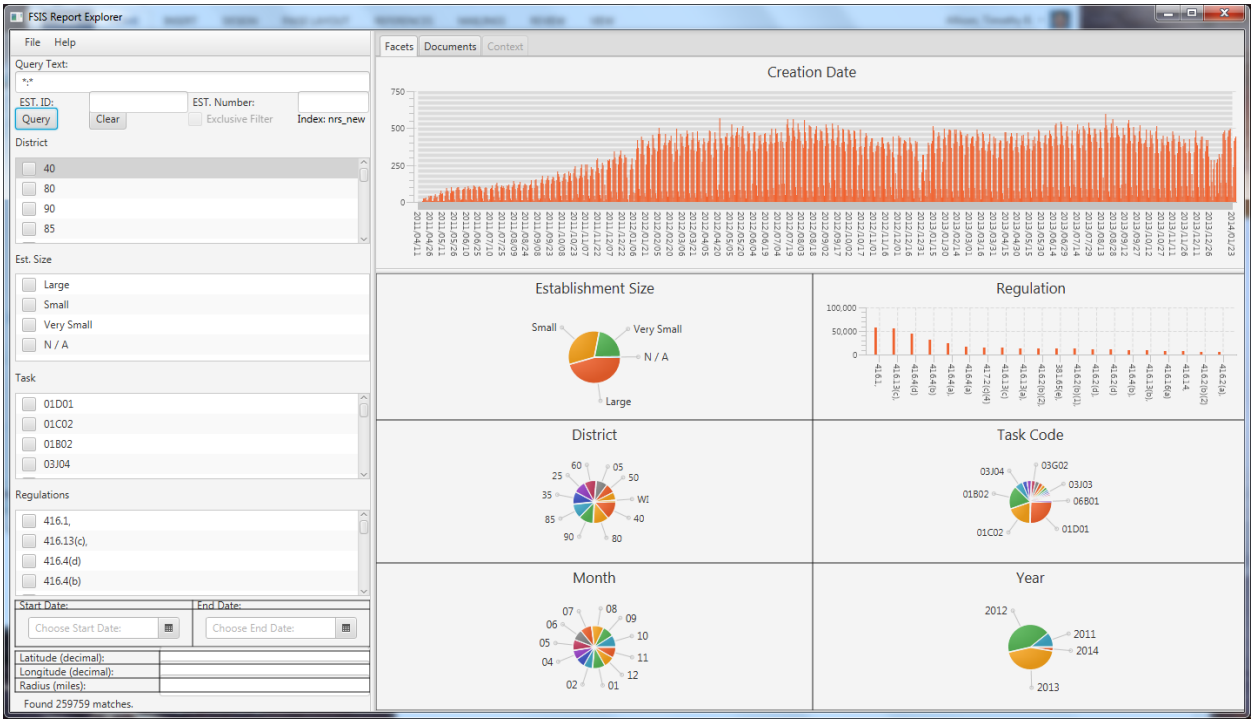


Figure 7: NR Search Prototype, Show All Results

Figure 8 shows the results for the search term "veal" in NRs within this data set. Compared with Figure 7, one can see that "veal" returns far more records for small and very small establishments and that "veal" is much more common in NRs in district 60 than in the overall set of NRs.

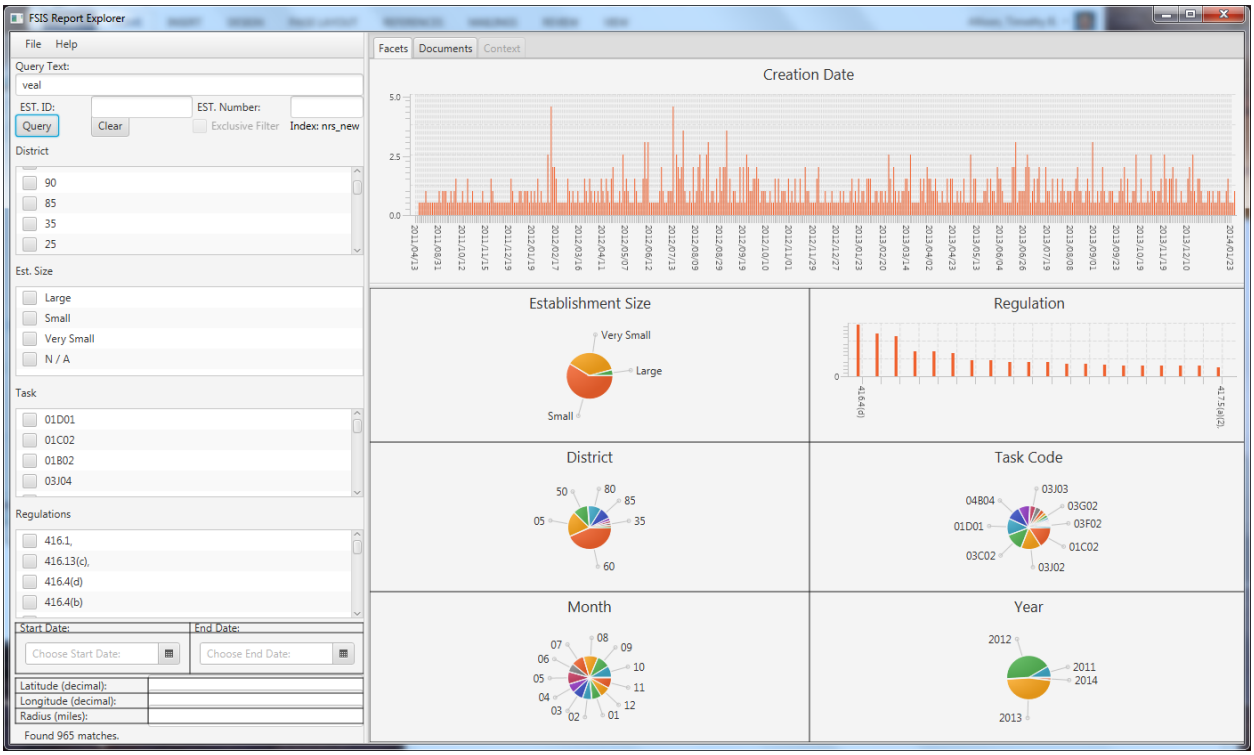


Figure 8: NR Search Prototype -- Show Results for "veal"

Figure 9 shows the results for a query of "veal" in NRs from small establishments in district 60 that occurred between January 1, 2012 and December 31, 2012.

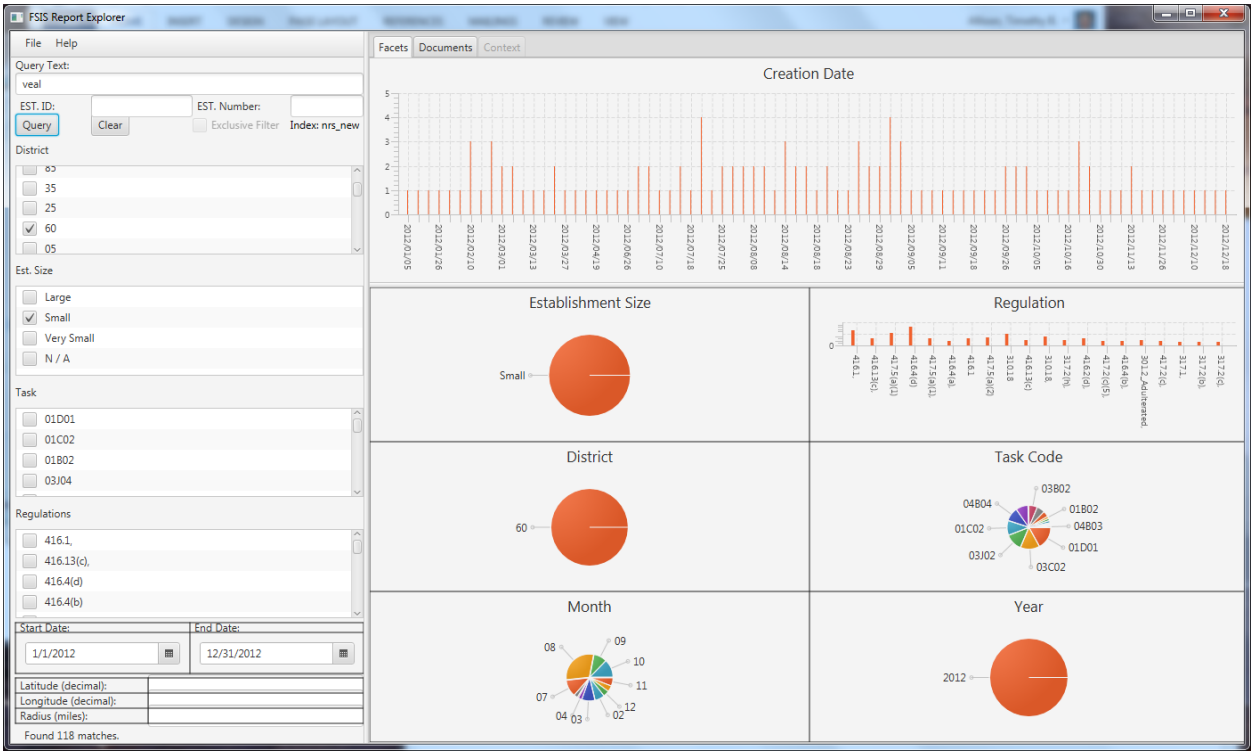


Figure 9: NR Search Prototype --Show Results for "veal" in 2012 NRs in Small Establishments in District 60

Figure 10 shows the "Documents" view of the results. With the same query window, analysts could view the document level view to see the contents. Ideally, we would also like to integrate the Rhapsode tools into other views: concordance, co-occurrence and variant finder.

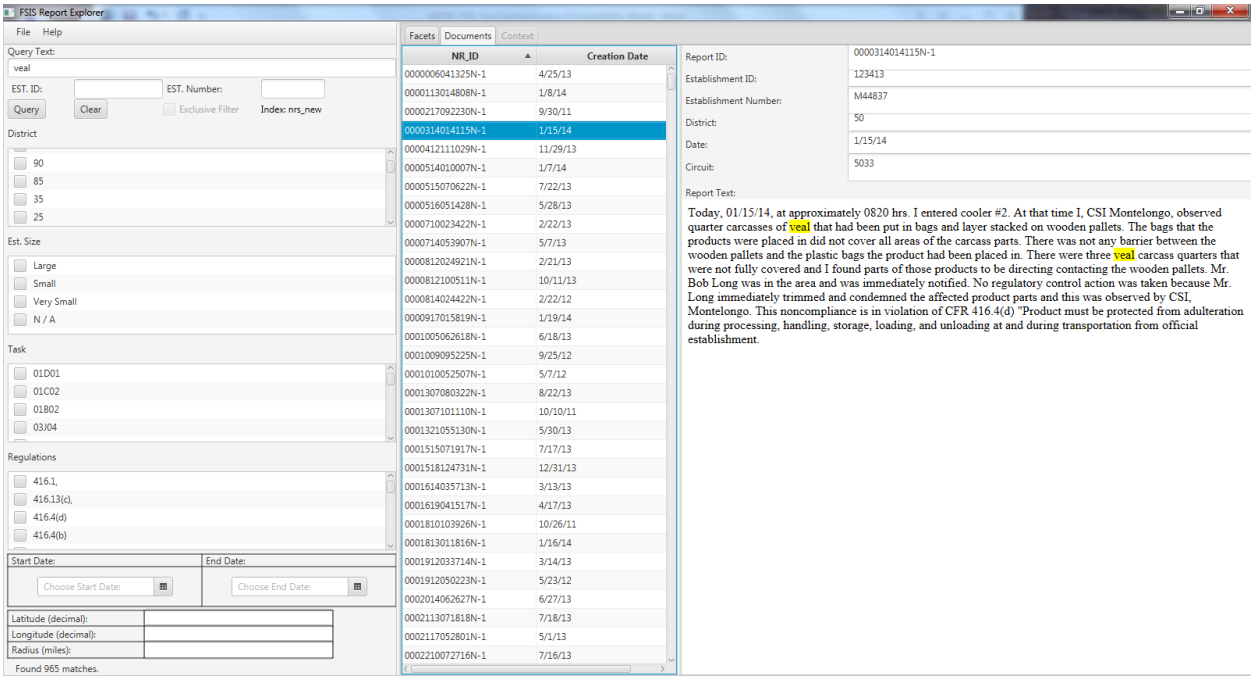


Figure 10: NR Prototype -- Document View of NRs that Contain "veal"