It's Dirty Work: Cleaning Data

Goal: Share insight into cleaning data such as common mistakes, common tools that make data cleaning faster, standard operating procedures, etc. Ask questions about best practices. Determine if there are processes we can all agree on to normalize statistical reports released across states or departments.

Prompt Questions

- What fields do you always check for accuracy?
- What fields are normally accurate?
- Do you correct invalid reports, send them back for correction, or not worry about them?
- What tools do you use to cleanup data?
- What's your process for cleaning data?

Discussion opened by asking what fields are usually the messiest and require correction. Responses included:

- Response times from typos in dates/times on the Basic module
- Street names inconsistently entered with different spellings and abbreviations
- Misusing the parceled address fields and including all information in one line. Example: 1620 N Main St in the Street Name field instead of each portion being in the correct field
- Extra zeros in the Property/Contents Loss and value
- Extra zeros in acres
- Using the unit number in Resources instead of the total number of apparatus. Example: if unit 241 responded, a fire department might enter 241 under suppression apparatus in the Basic module Resources box, instead of just 1 to show one apparatus responded. This is not a problem for departments that use the Apparatus/Resources module
- Future dates being used
- Zip codes keyed incorrectly

Discussion included tools that help to clean data. Mentioned were:

- USPS API that cleans up bad addresses (https://www.usps.com/business/web-tools-apis/address-information.htm)
- Only valid reports can be counted. Currently there is an issue that will be addressed by USFA on the "No Flame Spread" checkbox validation error currently occurring in structure fires
- Creating a table of zip codes assigned to cities. These tables already exist and could easily be implemented into a software. The user would enter the zip code first and then choose from a list of appropriate cities assigned to that zip code. This would insure the zip code is correct and the city is also correct
- Export your invalid incidents using the Bulk Export Tool from USFA and then reimport the incidents. Many times the revalidation process will fix the critical errors
- Google Refine is an online tool that can clean up spreadsheets, especially useful for addresses with many different spellings and abbreviations. Refine will locate all iterations and combine them at your command

A long discussion centered around correcting reports and the legal implications of changing reports. Points included:

- It should be well advertised that reports should be correct and updated with the most current information
- Correcting reports cannot typically be done at the State level. In the event a report at the State level does not match the local level there could be legal problems for both
- Fire Departments should have a policy regarding who can make changes and how to document them
    - One tactic can include a "return to writer" policy sending the mistakes back to the original author. While this takes more time on the front with making notes and sending emails, the original author cannot fix (and normally is unaware) any of the mistakes. Return to writer helps personnel enter better reports and lowers errors in the long run
    - No matter who makes the changes at the FD, consider adding information to the narrative including the date changes were made, the person making the changes, and what changes were made. Example: "January 1, 2015 Captain Doe changed the Property Loss from $100,000 to $120,000 after receiving updated report from insurance."

- If reports are changed at the State level, permission should always be obtained from the original FD and communication on any changes should occur first. The FD should also confirm that they will update their local report to reflect the same
- "Cleaning Data" in separate databases, outside of the NFIRS system is perfectly fine because the fire department's report is not technically changed
- When importing reports to NFIRS, always check the log files. The log file ending in .err will include the warnings and critical errors. This file will include the report number, report date, error description, error location (name of field and name of module)
- Make sure to send updated reports after the errors have been corrected at the fire department. Cleaning up bad data does no good if it doesn't go anywhere. Remember that importing a report overwrites the old report entirely
- For problems with dollar estimation, pre-incident value can be determined using Kelly Blue Book for vehicles and county appraiser websites. Property loss is often missing when investigations don't return their results to the report writers. Opening the doors of communication must occur at all levels in the FD to get information where it needs to go

During the discussion on cleaning reports, Number of Acres burned was brought up. Many reports have utilized the checkbox "Less than one acre burned" and State program managers typically assign that checkbox an actual value. It was discovered that states use very different values from .001 acre to .5 acres. There was no consensus on what to use. Instead, the problem can be corrected by removing the checkbox entirely and forcing a FD to report an actual number using decimals, like the Wildland Fire module.

The Fire Suppression Factors field was also mentioned. It is believed that most departments are not utilizing that field. One suggestion was to move it from the Fire/Wildland Fire module to the Basic module and change it to a "Mitigation Factor" field to better reflect barriers that a FD may have had for all types of calls. The example was given that if you have to get out in 2 feet of snow, no matter where you are going or what you are doing that is going to affect your call. Extreme weather, hostile residents, hoarder houses, etc all can affect a call beyond just fires. While a hoarder house can make fighting fire extremely deadly, it can also increase the injury risk to firefighters during EMS calls while attempting to navigate a patient out. This would also help FD's who look at their "outliers" to see quickly what affected their calls and can lead to better discussion and policy evolution.

Everyone, at all levels of involvement, must realize what they can control. We cannot stop people who want to do poorly, who refuse our help, or who simply want to whine. We can offer assistance, moderate our attitude, develop better training, and institute policies. We cannot change someone who will refuse to improve but we can educate someone who doesn't know that they need to improve.
Fire departments and personnel working hard at good data must be recognized and thanked for their efforts. NFIRS cannot succeed without the data champs and respect must be given.

State program managers have to be helpful, use the data, and show that it matters. Stagnant datasets are difficult to champion. Using the data at the Federal/National level is too far removed from Fire Departments for them to see their hard work paying off. A positive State program manager can increase care and quality while a negative State PM quickly impacts the participation at the FD level.