CGA DIRT

Damage Information Reporting Tool

Digging into DIRT Data:

How CGA Analyzes Multiple Reports on the Same Damage Event

March 2021

To download or to access additional analysis,visit CommonGroundAlliance.com/DIRT.

Abstract

Since the early days of DIRT, these important questions have been raised: Are multiple stakeholders submitting reports on the same events, and if so, how is that accounted for in the data analysis? What if an excavator and a locating company (or facility operator) both enter DIRT reports for the same event? How often does it occur? If they enter different root causes, which one is used in annual DIRT Reports?

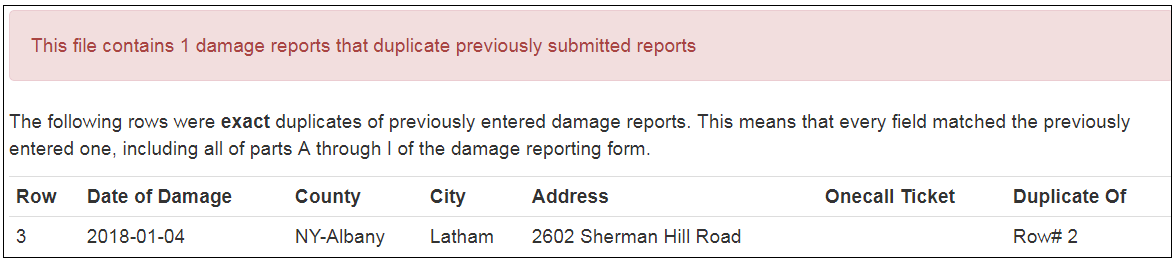
In this paper, we review the method to identify instances of multiple reports of the same event in the DIRT data, describe how they are handled in the annual DIRT Report, and then compare and contrast root causes reported by excavators versus other event sources.

Contents

"Multiple Reports of the Same Event" Defined

- **This report details the methods used by CGA's Data Committee to match and weight "multiple reports of the same event," defined as situations where two or more stakeholders submit separate DIRT reports for the same damage or near miss.**
- **Importantly, the DIRT system will give an error for the attempted entry of "duplicate reports," or reports in which every DIRT field exactly matches another report – so annual DIRT Reports do not overrepresent the number of actual or estimated damages or near misses.**

Stakeholders often refer to situations where two (or more) parties submit separate DIRT reports for the same damage or near miss as "duplicate reporting." For this report, we are referring to these situations more precisely as "multiple reports of the same event." When users enter data, DIRT produces an error message if a report **exactly** duplicates (every question is answered the same and every free-text field is filled in exactly the same) a report previously entered in the system. For a bulk upload (CSV Excel file) the error message would be produced if the file contains one or more sets of duplicate rows in which every column is filled in **exactly** the same.



This file contains 1 damage reports that duplicate previously submitted reports

The following rows were **exact** duplicates of previously entered damage reports. This means that every field matched the previously entered one, including all of parts A through I of the damage reporting form.

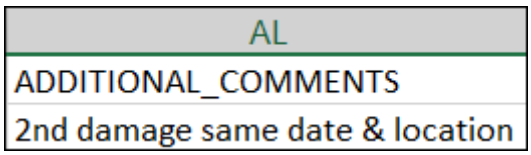
Row	Date of Damage	County	City	Address	Onecall Ticket	Duplicate Of
3	2018-01-04	NY-Albany	Latham	2602 Sherman Hill Road		Row# 2

The DIRT user then needs to determine if the duplication error(s) refer to the same or separate events. If it is the same event, they must delete duplicate row(s) and reload the file. If they are separate events, they must change at least one answer or free-text field in one of the reports so that the reports are no longer identical. If there really were two damages on the same date, at the same address, city, and intersection, with the same type of excavator, work performed, equipment used, same root cause, etc., the user could enter "2nd damage same day" in the free-text "Additional Comments" field of one of the reports, and leave it blank in the other, so that they no longer are exact duplicates.

Part J: Additional Comments

Additional Comments: 2nd damage same date and location

(Character Limit: 4000)



AL
ADDITIONAL_COMMENTS
2nd damage same date & location

If two different companies entered reports on the same event, presumably the original sources of the event ("event source")[1] would be different (e.g., an excavator and a locator). This is sufficiently different to allow the second report to be entered in DIRT, even if every other field is filled in exactly the same.

When the event source is the same for a report already in the system (e.g., entered by the same person or a different person from the same company, where they would presumably choose the same event source), a slight difference in any other field would permit the subsequent report to be entered. This

[1] Prior to the revisions to DIRT effective Jan. 1, 2018, this was referred to as Reporting Stakeholder.

could be a difference in how the address is entered (e.g., St. vs. Street, Ave. vs. Avenue, E. vs. East, a report with vs. without a house number), or in the dollar value of downtime cost, or something as innocuous as an extra space or comma or other stray character in the free text DAMAGE_OTHER_DESC field.

Because of this strict definition of "duplicates" for DIRT purposes, there technically are no duplicates in the dataset when it is extracted for the annual DIRT Report. By design, duplicates are prohibited from being entered in DIRT in the first place. What *can* get into the DIRT database are "multiple reports of the same event," which are two (or more) reports that are not exact duplicates but have enough similarities in the date, location, and facility affected fields to indicate that the reports very likely are based on the same event.

The majority of "multiple reports of the same event" that end up in DIRT are combinations of different event sources, such as facility operator, locator, excavator, one call center, regulator, etc. However, a small percentage do come from the same company accidently entering reports with slight differences. For brevity, in this report we may sometimes refer to these as "matching" reports.

Key Takeaways from This Analysis

1. When known root causes[2] are provided in a pair of matching reports from an excavator and locator, they point to each other as the responsible party slightly more often than they agree on the root cause. For pairs involving excavators and natural gas, it is a virtual tie.
2. A very low percentage of excavator-as-event-source reports have a known root cause, making it inconclusive as to which party they perceive as responsible.
3. When excavators do provide a known root cause, they identify a locating issue the majority of the time.
4. The fact that such a small percentage of excavator reports provide a known root cause while a high percentage of locators/facility operator reports do so—with most pointing to the excavator as the responsible party—shapes the root cause graphs depicted in the annual DIRT reports and online dashboards.
5. The majority of reports attributed to excavators as the event source are actually entered in DIRT by one call centers and are the source of a majority of reports with an unknown root cause. Reports entered in DIRT directly by excavators have a much higher percentage of known root causes but are much smaller in quantity than those funneled through one call centers.
6. Increased and higher-quality reporting from excavators is needed for DIRT to reflect the excavator point of view. One way to achieve this would be to encourage more excavators to report directly to DIRT.
7. No stakeholder should hesitate to submit DIRT reports over concerns that some other party may already be doing so. This report demonstrates that the Data Committee has a method to handle multiple reports of the same event. In fact, the Data Committee welcomes reports from multiple organizations as they help identify the concerns and perceptions of the various stakeholders.
8. Third party app/software developers have recently shown interest in adding features that allow users to collect and submit data into DIRT directly through their applications. The Data Committee's DIRT Product Task Team is working on outlining requirements and formalizing a process to evaluate and certify these programs. Hopefully, this will make DIRT reporting easier for excavators and thereby lead to increased direct reporting.

[2] A root cause aside from "Other" or "Data Not Collected."

How Are Matching Reports Handled in the DIRT Report?

- **CGA's Data Committee worked with data science consultants to create an algorithm that searches key fields in DIRT reports (such as city, address, and intersection) to identify matching reports and assign them a unique set number (see Figures 1 and 2). This process has been applied to the annual DIRT Reports from the 2015 edition onward.**
- **Matching damages' attributes are then individually weighted to add up to 1 (a single damage). For example, if two reports were entered for a single damage and one listed the excavator type as a municipality and the other as a utility, each excavator type would be weighted as 0.5 (together equaling 1; see Figures 3 and 4). This methodology ensures that all perspectives on a given damage are included and appropriately weighted while not over-counting the number of actual damages.**

Since the early days of DIRT, a frequently asked question has been: If an excavator and facility operator or locator both report on the same event, which report is used and becomes the *official* report, and who decides? The answer is that the Data Committee's role is to collect and analyze damage and near miss data from all stakeholders, but not to be the arbiter of which report is correct in a set of two or more. In the very early years of DIRT reporting,[3] the Committee was focused on growing DIRT reporting and felt that any matching reporting that might be occurring was minimal and would not affect data analysis. By the mid- 2010s, the Data Committee realized that matching reporting had likely grown to the point where it should be accounted for in annual DIRT reports. In 2016, the committee sought and hired a data science consultant that specializes in this area. A volunteer task team from the Data Committee worked closely with the consultant to develop and test a method to identify potential matching reports.

At the same time, the Data Committee considered how apply the results of such a method to the DIRT Report. One option considered was to use the report with the best Data Quality Index (DQI) score out of a set of matching reports and discard the other(s). For example, if one report had a known root cause, work performed, equipment used, etc., and the other had UNKNOWNS for those fields, the report with more known data would be used for the DIRT Report. Although this would improve the DQI of the overall database, the Committee was concerned that this approach would exclude some stakeholders from having their voices heard and discourage DIRT reporting rather than encourage reporting. Instead, the Committee decided to accept all data (including UNKNOWN fields) and use a weighting method based on the number of reports in a set of matching reports.

If one were to manually search for matching reports in an Excel spreadsheet, the logical approach would be to sort by Date, State/Province, County, City, Address, Intersection, and Facility Damaged. In fact, this was the approach of the Data Committee's DIRT Report consultants, who advised the DIRT Report team that there was minimal matching reporting, and most of it was the same companies accidently entering matching reports. However, based on the knowledge of which organizations were reporting to DIRT and from where, the Data Committee suspected there was more undetected overlapping reporting in the DIRT data.

In computer science, fuzzy string matching is the technique of finding strings that match a pattern approximately (rather than exactly). In other words, fuzzy string matching is a type of search that will find

[3] DIRT reporting began in 2003.

matches even when users misspell words or enter only partial words for the search. It is also known as approximate string matching.[4]

At a very high level, the data science consultant used fuzzy string matching techniques to develop algorithms that search free-text DIRT fields such as city, address, and intersection, looking at the number and order of matching characters, how far apart they are, and the number of deletions, insertions, or substitutions required to have them match.[5] This technique can identify matching reports that may be easily missed with a manual approach because they may not sort consecutively, or even closely. In hindsight, it is apparent the DIRT Report consultants were mostly finding reports from the same companies because individuals tend to have their own consistent habits when entering free-text information, so matching reports will sort adjacently or close and be easier to spot manually. But different people have different habits, which led the report consultants to underestimate the extent of matching reports from different companies.

Figure 1 shows examples of matching reports from assorted event sources, but with differences in the city and address fields, demonstrating how such reports get past the DIRT exact-duplicate test but are captured by the algorithms.[6]

Throughout this report we use the same abbreviations for original event sources and root causes as found in the DIRT Upload Specification for entering data with Excel CSV files. Please see the Appendix at end of this report for the full descriptions of those abbreviations.

ORIG_EVENT_SOURCE	CITY	ADDRESS
TELC	JAX	14436 LAKE JESSUP DRIVE
LOCA	JACKSONVILLE	14436 LAKE JESSUP DRIVE
TELC	JACKSONVILLE	6957 WEST LA MESA DRIVE
LOCA	JCVL	6957 WEST LA MESA DRIVE
TELC	ORLANDO	349 CANOE TRAIL LANE ORL FL 32825 3336 USA
LOCA	ORL	349 CANOE TRAIL LANE
TELC	ORLANDO	FRONT OF 722 PARK MANOR DRIVE
LOCA	ORLANDO	722 PARK MANOR DRIVE
ELEC	Madison	BRADFIELD ROAD
LOCA	MADISON	145 BRADFIELD ROAD
EXCV	SAINT	MOECKEL PL
REGU	Saint	305 MOECKEL PLACE
LOCA	SAINT	MOECKEL PL
LOCA	ST	305 MOECKEL PLACE

Figure 1–Sample Data Output with Mismatching City and Address

[4] Source: https://towardsdatascience.com/natural-language-processing-for-fuzzy-string-matching-with-python-6632b7824c49.

[5] For additional background, find "Levenshtein Distance" in an internet search engine.

[6] DIRT allows the city and address fields to be blank. This is because not all events occur in an incorporated town, city, or village, and not always at or near a building or property with an address. However, this practice can cause false positive matches if the same facility operation is affected on the same day in the same county, because the algorithm compares blank against blank and treats it as a match. This practice could also cause the algorithm to NOT identify a match if one party enters the city and address data and another party leaves it blank. DIRT users are encouraged to enter city and address data whenever possible – do not leave it blank just because it is allowed.

Beyond date and location information, the facility damaged must match in a set of matching reports, except that cable TV and telecommunications are considered a match. Unknown facilities damaged are treated as "wild cards" (see Figure 2 for examples).

ORIG_EVENT_SOURCE	CITY	ADDRESS	FAC_DAMAGED
EXCV	GACHER	245 HAMES ROAD	TELECOM
LOCA	GACHER	245 HAMES ROAD	CATV
EXCV	NCWAKE	129 CHINABROOK COURT	TELECOM
LOCA	NCWAKE	129 CHINABROOK COURT	UNKNOWN

Figure 2–Sample Data Output with Mismatched Facility Damaged

DIRT has always assigned a unique damage report ID to each report entered. The data consultant's output uses these ID numbers to group matching reports into "sets" with a linking identification number (called a "set number"). Each report in the set is given a "weight" equal to one divided by the number of reports in that set. See Figure 3 for examples.

Set number	Weight	DAMAGE_REPORT_ID
2	0.5	3807275
2	0.5	3879131
13	0.3333333	3807270
13	0.3333333	3879147
13	0.3333333	3879148
34	1	3879125
35	1	3879126
36	0.5	3645171
36	0.5	3879277
37	1	3879127
38	1	2820817
705	0.25	3658394
705	0.25	3700765
705	0.25	3800002
705	0.25	3879659

Figure 3–Sample Data Output with Set Number, Weight and Damage Report ID

Once the matching reports are sorted into sets, the appropriate weights are applied to the other DIRT fields within those reports. Using the examples in Figure 4 for illustration, from Set 21 we would count 0.5 municipality and 0.5 utility as the excavator type, and 0.5 "excavator failed to maintain clearance" (EXCLEARANCE) and 0.5 "insufficient excavation process not listed above" (INSUFEX) as the damage cause. From Set 447 we count 0.666666 backhoe and 0.333333 boring as the excavation type and 0.3333333 "root cause not listed above (NOTCOL), 0.333333 "facility was not located or marked" (INSUFMARKING), and 0.333333 "insufficient excavation process not listed above" (INSUFEX) as the damage cause. (Pleases see Appendix for full abbreviation definitions.)

Set number	Weight	EXCAVATOR_TYPE	EXCAVATION_TYPE	WORK_PERFORMED	DAMAGE_CAUSE
21	0.5	MUNICIPALITY	BACKHOE	WATER	EXCLEARANCE
21	0.5	UTILITY	NOTCOLLECTED	NOTCOLLECTED	INSUFEX
32	0.5	UNKNOWN	NOTCOLLECTED	NOTCOLLECTED	EXTESTHOLE
32	0.5	UNKNOWN	NOTCOLLECTED	NOTCOLLECTED	INSUFEX
83	0.5	UNKNOWN	UNKNOWN	NOTCOLLECTED	INSUFCALL
83	0.5	CONTRACTOR	BACKHOE	WATER	INSUFEX
447	0.333333	NOTCOLLECTED	BACKHOE	NOTCOLLECTED	NOTCOL
447	0.333333	NOTCOLLECTED	BACKHOE	NATGAS	INSUFMARKING
447	0.333333	MUNICIPALITY	BORING	UNKNOWN	INSUFEX
712	0.333333	CONTRACTOR	NOTCOLLECTED	WATER	NOTCOL
712	0.333333	CONTRACTOR	UNKNOWN	WATER	NOTCOL
712	0.333333	UNKNOWN	TRENCHER	SEWER	INSUFEX

Figure 4–Sample Data Output Showing Weighting of Various DIRT Fields

Once the cut-off date for entering annual data has passed (March 31 of the next year), the data is extracted and run through the matching/weighting program.

Appendix B of the 2019 DIRT Report describes the **Damage Report Path—Entry to DIRT Report.** The first two (out of four) steps are:

1. DIRT users entered 530,945 underground damage reports and 3,206 near miss reports from the United States and Canada for 2019.
2. A program was run to match and weight reports of the same event. This compressed the totals to 453,766 unique underground damages and 2,524 unique near misses. Near misses are set aside for separate analysis. The online DIRT dashboard is based on the number of unique damages (453,766 with no filters applied), as are all figures and tables in this report except those associated with Data Quality Index (DQI).

For all annual DIRT Reports up to and including 2014, data was based only on the overall count of damage report IDs (Step 1 above). Starting with the 2015 DIRT Report, this method has been applied, and data presented in the annual report has typically been compressed by 15-20% following the matching/weighting process. This “Digging into DIRT Data” report is essentially a deep dive into Step 2 (above).

It should be emphasized that the matching/weighting process applies only to the extracted data for the annual DIRT Reports and dashboards. All original data entered and stored in DIRT is unaffected. When the DIRT Query Wizard feature is used, the data displayed is based on the reports as originally entered. Organizations that host statewide Virtual Private DIRT (VPD) applications, whereby other companies share their data with the host, have questioned why what they see in Query Wizard differs from what they see on the DIRT Dashboards. This is one of the leading reasons. CGA Staff can assist in reconciling differences upon request.

Root Causes Reported by Excavators Versus Other Event Sources

- **This section provides detailed examinations of the top three event source combinations for matching events (excavator-locator, excavator-natural gas, and excavator-excavator), following a process that examines grouped root causes and then analyzes the combinations by sorting according to whether the matching reports agree on the root cause, are in conflict, or are inconclusive. For conflicting reports, we look at whether each party points to the other or to itself as the one responsible.**
- **When excavators and locators submit reports on the same event, they are slightly more likely to point to each other as the responsible party than they are to agree on the root cause. For pairs involving excavators and natural gas, it is essentially a tie.**
- **Very few excavator-as-event-source reports have a known root cause, providing no information as to which party they perceive as responsible. When excavators do provide a known root cause, they mostly identify locating issues.**

For any DIRT field, the number of theoretical combinations of answers in a set of matching reports equals the number of DIRT options for that question, raised to the power equal to the number of matching report IDs in the set.

DIRT Field	# of DIRT Options	Combinations in a Set of 2	Combinations in a Set of 3
Event Source	14	$14^2 = 196$	$14^3 = 2{,}744$
Root Cause	26	$26^2 = 676$	$26^3 = 17{,}576$
Work Performed	31	$31^2 = 961$	$31^3 = 29{,}791$

For example, a set of three matching reports could include excavator-locator-regulator as the event sources, or road builder-natural gas-natural gas. We may not actually find every possible combination in the data set, but we see that it can become complicated very quickly.

Figure 5 below comes from the 2019 DIRT Report and shows the major root cause groupings by event source. The numbers at the top are the weighted events from that event source, with unknown event source and root cause data filtered out. (An explanation of the root cause groupings starts on page 12 of the 2019 DIRT Report.)

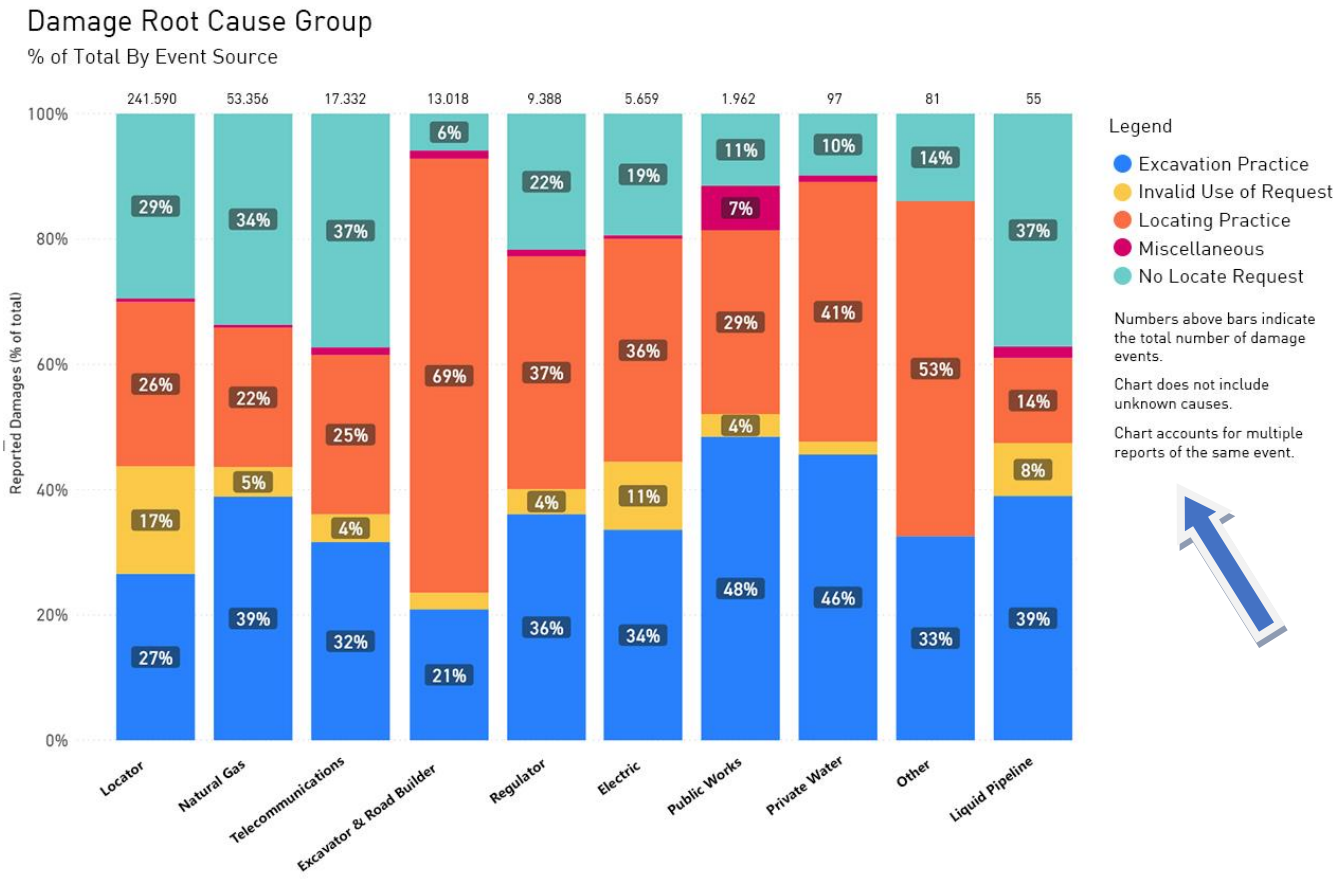


Figure 5–2019 DIRT Data, Root Cause Groups by Event Source

Figure 5 shows that excavators and road builders cite the highest percentages of locating practices as damage root causes, the lowest percentages of excavating practices, and no locate request root causes (keeping in mind that unknown data is filtered out). This has been a consistent trend in past DIRT Reports. Since the question of matching reports most often revolves around root causes reported by excavators versus locators or facility operators, this analysis will focus on root cause choices from sets of two matching reports, where at least one of the event sources is excavators.

Matching sets from the 2015 through 2019 data with a weight of 0.5 (i.e. pair of reports), and with one (or both) reports having excavator (EXCV) as the event source, were extracted from the DIRT dataset. Road builders are counted as excavators. Figure 6 shows the number of paired reports for each combination of event sources, sorted high-to-low.

Event Source 1	Event Source 2	Event Source Abbrev	Pairs of Reports
Excavator	Locator	EXCV LOCA	71,153
Excavator	Natural Gas	EXCV NATGAS	5,306
Excavator	Excavator	EXCV EXCA	4,464
Excavator	Regulator	EXCV REGU	2,351
Excavator	Telecommunications	EXCV TELC	791
Excavator	One Call Center	EXCV 1CAL	679
Excavator	Public Works	EXCV PUBW	296
Excavator	Electric	EXCV ELEC	246
Excavator	Unknown	EXCV UNKNOWN	223
Excavator	Private Water	EXCV PRVW	9
Excavator	Liquid Pipe	EXCV LIQPIPE	3
Excavator	Engineering	EXCV ENGR	2
		Sum of Pairs	85,523

Figure 6—Number of Paired Reports by Event Source Involving Excavator (EXCV)

The root causes from each pair of reports were also combined. To minimize the number of combinations of event sources and root causes, the following adjustments are made:

- Root causes are sorted into their major groupings: excavation practices, locating practices and miscellaneous. No locate request and invalid use of request are combined under “notification” (see the Appendix for abbreviation definitions and groupings). These, combined with unknown, give us 5 root cause groups with $5^2 = 25$ possible combinations.
- Excavators and road builders are combined under excavator (EXCV).
- Data not collected (DNC) and other are combined as “unknown.” Note that starting in 2018, DNC was eliminated as an option for all fields, and one call (1CAL) was eliminated as an option for event source.
- Abandoned is grouped with locating rather than miscellaneous, for all years covered here, to be consistent with the revised root cause groups, effective Jan. 1, 2018. The rationale is that an abandoned facility may cause an active facility to be mislocated or not located at all. An excavator may know only that a facility was mis- or un-marked and report it as a locator error,

while a locator or facility operator would be better able to determine that it was an abandoned facility issue (where it would have been categorized as miscellaneous in years 2015–2017).

Figure 7 shows the results of this process for the top three event source combinations. The top row is the pair of event sources, with excavator (EXCV) always first. The first column lists the paired root causes, with the excavator's root cause group choice always indicated first (aside from EXCV-EXCV). For example, there were 462 pairs of excavator-locator (EXCV-LOCA) reports where both indicated excavating root causes, and 3,952 pairs where the excavator chose a locating root cause, and the locator chose an excavating root cause.

Root Cause Group	EXCV LOCA	EXCV NATGAS	EXCV EXCV
EXCAVATING EXCAVATING	462	316	266
EXCAVATING LOCATING	170	71	59
EXCAVATING MISC	14		
EXCAVATING NOTIFICATION	230	110	2
EXCAVATING UNKNOWN	127	190	25
LOCATING EXCAVATING	3,952	453	17
LOCATING LOCATING	3,222	495	2,082
LOCATING MISC	19	5	3
LOCATING NOTIFICATION	1,466	471	30
LOCATING UNKNOWN	848	716	87
MISC EXCAVATING	24	14	
MISC LOCATING	6	5	1
MISC MISC			15
MISC NOTIFICATION	14	8	
MISC UNKNOWN	13	22	2
NOTIFICATION EXCAVATING	727	55	4
NOTIFICATION LOCATING	294	35	32
NOTIFICATION MISC	6		
NOTIFICATION NOTIFICATION	772	125	187
NOTIFICATION UNKNOWN	110	90	7
UNKNOWN EXCAVATING	26,414	830	21
UNKNOWN LOCATING	14,451	552	75
UNKNOWN MISC	121	2	4
UNKNOWN NOTIFICATION	12,539	333	5
UNKNOWN UNKNOWN	5,152	408	1,540
Grand Total	**71,153**	**5,306**	**4,464**

Figure 7—Number of Sets by Concatenated Event Source and Root Cause

The EXCV-LOCA combination has the largest number of sets by far. Next, we analyze these combinations by sorting according to whether the matching reports agree on the root cause, are in conflict, or are inconclusive. For conflicting reports, we look at whether each party points to the other or to itself as the one responsible. For purposes of this analysis the following guidelines apply:

- Excavators are the responsible party for notification and excavation.
- Locators and facility operators are the responsible party for locating.

- Agreement means both parties agree on the root cause group, and by implication the responsible party. Some examples of why this makes sense follow below.
- "Other" means a party points to the other as responsible.
- "Self" means a party points to itself as responsible. For example, "Self>Other" means both parties point to the excavator, but for different (root cause) reasons.
- "Inconclusive" means the responsible party is not identified (unknown root cause) or it is not an excavator or locator responsibility (e.g., a miscellaneous root cause). "Unknown>Unknown," even though the same root cause group, is considered inconclusive rather than agreement because we cannot tell which party the event sources are pointing to as responsible.
- Root causes for abandoned facilities are grouped with locating rather than miscellaneous root causes. That leaves one call center error, previous damage, and deteriorated facility as the miscellaneous root causes, which are the responsibility of neither excavators nor locators. "Misc-Misc" will be considered agreement. If only one party chooses miscellaneous, it will be considered inconclusive. The miscellaneous group makes up a very small percentage of the total.

Here are several examples of how different event sources might choose different individual root causes, but each have validity if from the same root cause group:

- An excavator chooses *excavator dug after valid ticket expired*. If a long time has passed, the locator or facility operator may not search back far enough to find the ticket and report it as *no notification made to one call center / 811.* (**Notification Group)**

- An excavator chooses *excavator provided incorrect notification information* or *excavator dug outside area described on ticket*. A locator or facility operator may only know that they have no ticket for the damage location and report it as *no notification made to one call center / 811.* **(Notification Group)**

- A locator or facility operator is better able to determine if a mismark is due to bad maps, tracer wire issue, or an unlocatable or abandoned facility. All the excavator knows is that the marks were inaccurate and reports it as *locator error*. **(Locating Practices Group)**

- An excavator chooses *excavator dug prior to verifying marks by test-hole (pothole)* or *excavator failed to maintain clearance after verifying marks*. A locator or facility operator knows only that a one call notification was made, and the site was marked accurately and on time. By the time the facility operator arrives on site to investigate the damage, the hole or trench has been expanded such that it is difficult to determine if potholing occurred, so they report it as *improper excavation practice not listed above (INSUFEX)*. **(Excavation Practices Group)**

The "# SETS" column in Figure 8 takes the numbers from the "EXCV-LOCA" column in Figure 7 and sorts the rows according to whether the two reports agree on the root cause group, each point to the other as the responsible party, or some other combination as described above. Beyond agreement and each party pointing to the other, there are several other combinations involving "inconclusive" and parties pointing to themselves as responsible ("self").

Opinion EXCV - LOCA	Root Cause Group EXCV – LOCA	# Sets	Subtotal #	Subtotal %
Agreement	NOTIFICATION NOTIFICATION	772		
Agreement	EXCAVATING EXCAVATING	462		
Agreement	LOCATING LOCATING	3,222	4,456	6.26%
Other - Other	LOCATING EXCAVATING	3,952		
Other - Other	LOCATING NOTIFICATION	1,466	5,418	7.61%
Self- Self	EXCAVATING LOCATING	170		
Self - Self	NOTIFICATION LOCATING	294	464	0.65%
Self - Other	EXCAVATING NOTIFICATION	230		
Self- Other	NOTIFICATION EXCAVATING	727	957	1.34%
Self - Inconclusive	NOTIFICATION UNKNOWN	110		
Self- Inconclusive	EXCAVATING MISC	14		
Self - Inconclusive	EXCAVATING UNKNOWN	127		
Self - Inconclusive	NOTIFICATION MISC	6	257	0.36%
Inconclusive - Other	UNKNOWN EXCAVATING	26,414		
Inconclusive - Other	UNKNOWN NOTIFICATION	12,539		
Inconclusive - Other	MISC EXCAVATING	24		
Inconclusive - Other	MISC NOTIFICATION	14	38,991	54.80%
Other - Inconclusive	LOCATING MISC	19		
Other - Inconclusive	LOCATING UNKNOWN	848	867	1.22%
Inconclusive - Self	MISC LOCATING	6		
Inconclusive - Self	UNKNOWN LOCATING	14,451	14,457	20.32%
Inconclusive	MISC UNKNOWN	13		
Inconclusive	UNKNOWN MISC	121		
Inconclusive	UNKNOWN UNKNOWN	5,152	5,286	7.43%
	Total	**71,153**	**71,153**	**100.00%**

Figure 8—Excavator vs. Locator Root Cause Groups

Figure 8 shows that excavators and locators point to each other (7.61%) slightly more than they agree (6.26%) on root cause groups. However, "Agreement" and "Other – Other" make up only about 14% of the total.

"Unknown – Locating" (14,451) means the locator points to themselves and the excavator report has an unknown root cause. Assuming the excavator would agree that 100% of these involved locating root causes, those could be added to "Locating - Locating," bringing the "Agreement" subtotal up to 26.57%.

"Unknown – Excavating" (26,414) and "Unknown – Notification" (12,539) make up approximately 54% of the total. For these pairings, the locator points to the excavator but the excavator report does not

provide their side of the story. The excavator may not agree with the locator's selection 100% of the time, but we have no way of determining where else to categorize these pairs of reports.

The same analysis was performed for the next largest event source combination, excavator-natural gas ("EXCV-NATGAS"), to see if there are any significant differences from the excavator-locator ("EXCV-LOCA") combination. The results are shown in Figure 9. There are higher percentages of "Agreement" and "Other – Other" than in the excavator-locator combinations, with "Agreement" slightly exceeding "Other - Other," and both combinations making up about 35% of the total. "Inconclusive-Other" is again the largest combination, but not as high as in the "EXCV-LOCA" reports. Switching "Unknown Locating" to "Locating Locating" would bring "Agreement" to 28%.

Opinion EXCV - NATGAS	Root Cause Group EXCV – NATGAS	# Sets	Subtotal #	Subtotal %
Agreement	NOTIFICATION NOTIFICATION	125		
Agreement	EXCAVATING EXCAVATING	316		
Agreement	LOCATING LOCATING	495	936	17.64%
Other - Other	LOCATING EXCAVATING	453		
Other- Other	LOCATING NOTIFICATION	471	924	17.41%
Self - Self	EXCAVATING LOCATING	71		
Self - Self	NOTIFICATION LOCATING	35	106	2.00%
Self - Other	EXCAVATING NOTIFICATION	110		
Self - Other	NOTIFICATION EXCAVATING	55	165	3.11%
Self - Inconclusive	NOTIFICATION UNKNOWN	90		
Self - Inconclusive	EXCAVATING UNKNOWN	190	280	5.28%
Inconclusive - Other	UNKNOWN EXCAVATING	830		
Inconclusive - Other	UNKNOWN NOTIFICATION	333		
Inconclusive - Other	MISC EXCAVATING	14		
Inconclusive - Other	MISC NOTIFICATION	8	1,185	22.33%
Other - Inconclusive	LOCATING MISC	5		
Other - Inconclusive	LOCATING UNKNOWN	716	721	13.59%
Inconclusive - Self	MISC LOCATING	5		
Inconclusive - Self	UNKNOWN LOCATING	552	557	10.50%
Inconclusive	MISC UNKNOWN	22		
Inconclusive	UNKNOWN MISC	2		
Inconclusive	UNKNOWN UNKNOWN	408	432	8.14%
	Total	**5,306**	**5,306**	**100.00%**

Figure 9—Excavator vs. Natural Gas Root Cause Groups

The next largest combination of matching event sources is excavator-excavator ("EXCV-EXCV"), where more agreement on root causes is to be expected. Some of these pairs are the same company accidently reporting the same damage twice, probably with differences in the address field or some other fields aside from root cause, such as type of work, equipment, downtime, etc. Figure 10 bears out this expectation, with the highest level of agreement at 57.12%. The single largest pairing is Locating–Locating. Note that since here both reports are from excavators, "notification" and "excavating" now always point to "self." Pairs that mirror each other, such as "Locating-Unknown" and "Unknown-Locating" can be added together, thereby reducing the number of rows in the table.

All the combinations including the locating group add to 2,386, or 53% of the total. Pairs involving "Inconclusive" are less than in excavator-locator and excavator-natural gas reports, but still sizable at about 40%.

Opinion EXCV - EXCV	Root Cause Group EXCV – EXCV	# Sets	Subtotal #	Subtotal %
Agreement	NOTIFICATION NOTIFICATION	187		
Agreement	EXCAVATING EXCAVATING	266		
Agreement	LOCATING LOCATING	2,082		
Agreement	MISC MISC	15	2,550	57.12%
Other - Self	LOCATING EXCAVATING	76		
Other-Self	LOCATING NOTIFICATION	62	138	3.09%
Self - Self	NOTIFICATION EXCAVATING	6	6	0.13%
Self - Inconclusive	NOTIFICATION UNKNOWN	12		
Self - Inconclusive	EXCAVATING UNKNOWN	46	58	1.30%
Other - Inconclusive	LOCATING MISC	4		
Other - Inconclusive	LOCATING UNKNOWN	162	166	3.72%
Inconclusive	MISC UNKNOWN	6		
Inconclusive	UNKNOWN UNKNOWN	1,540	1,546	34.63%
	Total	**4,464**	**4,464**	**100.00%**

Figure 10—EXCV-EXCV Root Cause Groups

Why Do So Few Excavator Event Source Reports Include a Root Cause?

- **Most events with excavator as the event source are funneled through one call centers and do not list a root cause, which leaves holes in our understanding of the overall picture of damage root causes.**
- **In contrast, excavator companies that register in DIRT and submit damage data into DIRT directly are much more likely to provide a root cause. CGA is working to make DIRT submissions more seamless for excavators through third-party API authorizations in an effort to increase quality submissions from this stakeholder group.**
- **CGA's Data Committee has the ability to identify and account for duplicate reports and excavators need not hesitate to enter their own DIRT reports for worry that another stakeholder has already done so.**

There are actually two pieces of information in the DIRT data output relating to report sources. One is the "original source of event information," or "ORIG_EVENT_SOURCE." This is entered for each individual DIRT report and is the basis for the analysis above.

The other piece of information is "company stakeholder group." This is not entered for each individual DIRT report, but rather is based on what the submitting company selected as its stakeholder group when first registering in DIRT. The "original event source" on individual DIRT reports can sometimes differ from the company stakeholder group that enters the report. Figure 11 shows a sampling of several possible combinations.

COMP_STAKEHOLDER_GROUP	ORIG_EVENT_SOURCE
1CAL	EXCV
EXCV	EXCV
LOCA	LOCA
NATGAS	NATGAS
REGU	EXCV
1CAL	NATGAS

Figure 11—Company Stakeholder/Event Source Combinations

Prior to 2018, the "original event source" field allowed one call as an option. Several one call centers take "damage tickets" from excavators (and other sources) and use those as a source of DIRT reporting. Some were using one call as the event source on each individual report, even though the DIRT Users Guide gave the following guidance:

> One call centers and insurance companies that compile member or customer data for submission, please select the stakeholder group of the original source of the information.

The reason for this guidance is that when one call centers use one call as the event source, it masks the true source of the report. The Data Committee made efforts to educate the one call centers on the proper way to enter this data, and usage of one call as the event source was declining but not entirely eliminated prior to 2018. Since the 2018 DIRT revisions took effect, one call centers have improved at identifying excavators as the event source, but some appear to have defaulted to "unknown" now that one call is not an option. Note that Figure 6 above cited 679 excavator-one call event source combinations. Those were all from 2015 to 2017.

In this dataset, there are 84,364 individual reports with excavators as the event source, but the company stakeholder group listed as one call (i.e., DIRT report was funneled through a one call center). Of those, 66,249 (78.5%) had an unknown root cause. There are 572 reports where both event source and company stakeholder group are excavator (i.e., excavator submitted report through their own DIRT registration). Of those, 46 (8%) had an unknown root cause. In other words, DIRT reports entered directly by excavators through their own DIRT registrations have much higher quality, but much lower quantity.

Performing the same comparison of excavator-locator pairs as above, but slicing the data to isolate reports where both company stakeholder group and original event source are "excavator," leads to Figure 12. "Other – Other" is still higher than "Agreement," but together they total approximately 75% of the total dataset.

Opinion EXCV - LOCA	Root Cause Group EXCV - LOCA	# Sets	Subtotal #	Subtotal %
Agreement	EXCAVATING EXCAVATING	18		
Agreement	LOCATING LOCATING	28	46	35.94%
Other – Other	LOCATING EXCAVATING	37		
Other – Other	LOCATING NOTIFICATION	14	51	39.84%
Self- Self	EXCAVATING LOCATING	2	2	1.56%
Self - Other	EXCAVATING NOTIFICATION	12	12	9.38%
Self - Inconclusive	NOTIFICATION UNKNOWN	1		
Self - Inconclusive	EXCAVATING UNKNOWN	1	2	1.56%
Inconclusive - Other	UNKNOWN EXCAVATING	8		
Inconclusive - Other	UNKNOWN NOTIFICATION	1		
Inconclusive - Other	MISC EXCAVATING	1	10	7.81%
Other - Inconclusive	LOCATING UNKNOWN	4	4	3.13%
Inconclusive - Self	MISC LOCATING	1	1	0.78%
	Total	**128**	**128**	**100.00%**

Figure 12—Excavator vs. Locator Root Cause Groups with Excavator Based on Company Stakeholder Group

Performing the same data-slicing for the excavator-natural gas pairs results in Figure 13. Again, the data quality improves but the dataset is smaller. The parties agree nearly 2.5 times as much as they point to each other.

Opinion EXCV - NATGAS	Root Cause Group EXCV - NATGAS	# Sets	Subtotal #	Subtotal %
Agreement	NOTIFICATION NOTIFICATION	4		
Agreement	EXCAVATING EXCAVATING	51		
Agreement	LOCATING LOCATING	39	94	54.02%
Other - Other	LOCATING EXCAVATING	27		
Other- Other	LOCATING NOTIFICATION	11	38	21.84%
Self - Self	EXCAVATING LOCATING	2	2	1.15%
Self - Other	EXCAVATING NOTIFICATION	9		
Self - Other	NOTIFICATION EXCAVATING	1	10	5.75%
Self - Inconclusive	EXCAVATING UNKNOWN	5	5	2.87%
Inconclusive - Other	UNKNOWN EXCAVATING	11		
Inconclusive - Other	UNKNOWN NOTIFICATION	3		
Inconclusive - Other	MISC EXCAVATING	1	15	8.62%
Other - Inconclusive	LOCATING MISC	1		
Other - Inconclusive	LOCATING UNKNOWN	5	6	3.45%
Inconclusive	MISC UNKNOWN	1		
Inconclusive	UNKNOWN UNKNOWN	3	4	2.30%
	Total	**174**	**174**	**100.00%**

Figure 13—Excavator vs. Natural Gas Root Cause Groups with Excavator Based on Company Stakeholder Group

Paired reports entered in DIRT by one call centers and/or excavators as the company stakeholder,[7] with excavators as the original event source, can be from the following combinations of company stakeholder group (numbers in parentheses reflect what is actually in this dataset):

- Two reports entered by the same one call center (4,026)
- Two reports entered by the same excavating company (50)
- One report from the one call center and one from the excavating company (26)

[7] REGU-EXCV is another possible COMP_STAKEHOLDER/ORIG_EVENT_SOURCE combination and is actually higher in quantity than EXCV-EXCV, but still much less than 1CAL-EXCV. ROAD-ROAD and ROAD-EXCV are also possible combinations, but there are very few in the dataset. In this analysis ROAD is combined with EXCV.

The results of each of these company stakeholder group combinations are shown in Figures 14 to 16:

Opinion EXCV - EXCV	Root Cause Group EXCV – EXCV	# Sets	Subtotal #	Subtotal %
Agreement	NOTIFICATION NOTIFICATION	185		
Agreement	EXCAVATING EXCAVATING	225		
Agreement	LOCATING LOCATING	1,855		
Agreement	MISC MISC	14	2,279	56.61%
Other - Self	LOCATING EXCAVATING	46		
Other-Self	LOCATING NOTIFICATION	53	99	2.46%
Self - Self	NOTIFICATION EXCAVATING	5	5	0.12%
Self - Inconclusive	NOTIFICATION UNKNOWN	12		
Self - Inconclusive	EXCAVATING UNKNOWN	18	30	0.75%
Other - Inconclusive	LOCATING MISC	1		
Other - Inconclusive	LOCATING UNKNOWN	87	88	2.19%
Inconclusive	MISC UNKNOWN	3		
Inconclusive	UNKNOWN UNKNOWN	1,522	1,525	37.88%
	Total	**4,026**	**4,026**	**100.00%**

Figure 14—Two Excavator Event Source Reports Entered by the Same One Call Center as Company Stakeholder

Opinion EXCV - EXCV	Root Cause Group EXCV – EXCV	# Sets	Subtotal #	Subtotal %
Agreement	NOTIFICATION NOTIFICATION	1		
Agreement	EXCAVATING EXCAVATING	3		
Agreement	LOCATING LOCATING	34	38	76.00%
Other - Self	LOCATING EXCAVATING	5	5	10.00%
Self - Inconclusive	EXCAVATING UNKNOWN	2	2	4.00%
Other - Inconclusive	LOCATING UNKNOWN	1	1	2.00%
Inconclusive	MISC UNKNOWN	1		
Inconclusive	UNKNOWN UNKNOWN	3	4	8.00%
	Total	**50**	**50**	**100.00%**

Figure 15—Two Excavator Event Source Reports Entered by the Same Excavating Company as Company Stakeholder

The sets represented in Figure 15 all appear to be the same company accidently entering a report based on the same event twice. It could be different individual users within the same company. It could be an attempt to correct a report previously entered and not deleting the original. Not surprisingly, we find the best agreement here (76%) with "LOCATING – LOCATING" the leading combination. In these cases, the pairs in agreement will count as two reports at 0.5 weight of the same root cause group, adding to 1.0 for that set in the DIRT Dashboard. In other words, it is as if they only entered one report. The remaining 24% is miniscule in the context of the entire dataset. Remember that this is 50 sets over a five-year period.

Opinion EXCV - EXCV	Root Cause Group EXCV - EXCV	# Sets	Subtotal #	Subtotal %
Self - Inconclusive	EXCAVATING UNKNOWN	12	12	46.15%
Other - Inconclusive	LOCATING UNKNOWN	14	14	53.85%
	Total	**26**	**26**	**100.00%**

Figure 16—One Report from a One Call Center Registration and One from an Excavating Company Registration

This analysis demonstrates that excavator reports funneled through one call centers (company stakeholder group) have a high percentage of unknown root causes, while excavator companies that register in DIRT (i.e., COMP_STAKEHODER = EXCV) do much better at providing a known root cause. However, in actual practice the vast majority of reports with excavator as the event source are funneled through one call centers. This influences the root cause analyses found in annual DIRT Reports and online dashboards. Figure 17 shows the 2019 Dashboard DIRT Explorer page filtered on Event Source = Excavator + Road Builder. The figure reflects the entire 2019 dataset with the matching and weighting method applied as described in this report, but a similar dynamic occurs every year.

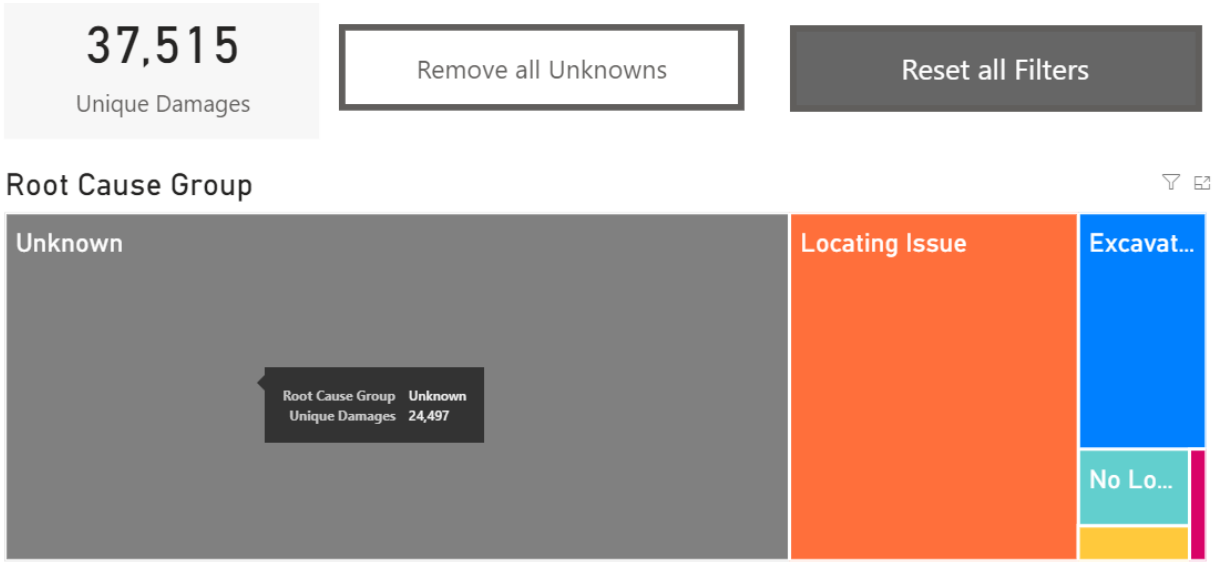


Figure 17—2019 DIRT Dashboard, Root Cause Groups Filtered on Excavators and Road Builders as Event Source

Of course, one way to improve this situation would be for one call centers to collect better root cause data. However, for a variety of reasons there has been little improvement in that area over the years.[8] Another way to improve the situation would be for excavating companies to increase their own direct DIRT reporting. There may be a perception among excavators that if they are reporting "damage tickets" to a one call center, they should not also submit their own DIRT report out of concern for "duplicate" reporting. Hopefully, this report will demonstrate that the Data Committee has the ability to identify and account for such *duplicate* reports and excavators need not hesitate to enter their own DIRT reports. This would be the best way for excavators to have their voices heard and their experiences reflected in the DIRT data.

In fact, no stakeholder should hesitate to report to DIRT due to concern that another entity may be reporting on the same events. First, they might be mistaken, in which case nobody reports the events to DIRT. Second, the Data Committee welcomes multiple reports of the same event as it would enable analysis beyond the annual DIRT Report, such as this supplemental report, if data quantity and quality made it feasible.

In furtherance of this, in July 2020, the CGA announced an initiative to expand options for stakeholders to submit their damage incident and near miss data into the DIRT by formalizing a process for data submission through third-party app integration.[9] The effort will address two goals of CGA's Data Reporting & Evaluation Committee:

- Expand data submitted into DIRT across all stakeholder groups
- Increase the "completeness" of data submitted into DIRT as measured by DIRT's Data Quality Index (DQI)

As part of this initiative, the Data Committee's DIRT Product Task Team is working on outlining requirements and formalizing a process to evaluate and certify apps/software developed by third-parties that allow users to collect and submit data into DIRT directly through their application. The team will prioritize the security and integrity of the data and adherence to current DIRT values/requirements.

[8] Some one call centers do not ask for a root cause when taking these damage tickets because their purpose is to immediately relay the information to affected facility owners for emergency response. The excavator may also not know or want to share a root cause at the time of the call. The rationale for one call centers to submit to DIRT despite the poor-quality data is that some data is better than none at all. At the very least, the one call data is useful for the estimates of total U.S. damages in the annual DIRT Reports. Recent changes to state damage prevention laws are trending to require direct DIRT reporting. As this continues, some one call centers may phase out submitting DIRT reports based on damage tickets, which will improve the overall quality of DIRT data.

[9] https://commongroundalliance.com/Publications-Media/Monthly-Updates/2020/Monthly-Update-July-2020

Should We Examine Sets of Three or More Matching Reports?

- **Due to the complexity of analysis and poor data quality, examining sets of three and more reports would yield little benefit at this time, but the Data Committee may revisit this idea in the future if data quality improves.**

The analysis thus far has focused on excavator reports pulled from matching pairs of reports, i.e., sets of two at 0.5 weight each. The analysis increases in complexity each time the set size increases by one. For example, to extract excavator and locator reports from sets of three, we could potentially have the following combinations:

- EXCV EXCV EXCV
- EXCV EXCV LOCA (*which EXCV Root Cause would we use?*)
- EXCV LOCA LOCA (*which LOCA Root Cause would we use?)*
- EXCV LOCA NATGAS …1CAL, REGU, TELC, CATV, ELEC, …. etc.

The above analysis of sets of two reports, with five root cause groups, lead to 25 (5 × 5) potential root cause combinations. For the "EXCV-LOCA" event source combination, we actually had 22 combinations, all except for "NOTIFICATION-MISC," "MISC-MISC" and "MISC-UNKNOWN." Each time one more report is added to a set, the number of possible root cause combinations increases exponentially. For sets of three matching reports, with five root cause groups, there could theoretically be 125 (5 x 5 x 5) root cause combinations.

Due to the added complexity, and little potential benefit due to the poor data quality, examining sets of three and more reports does not appear worth the effort at this time and would not be likely to significantly alter the key takeaways. This could be revisited in the future if data quality improves.

That said, the Data Committee welcomes suggestions on other ways to slice and analyze the data.

Appendix – Abbreviations and Root Cause Groupings

Original Source of Event

NOTE: 1CAL / One Call Center eliminated starting in 2018

ORIG_EVENT_SOURCE_ABBR	ORIG_EVENT_SOURCE_DESCRIPTION
1CAL	One Call Center
ELEC	Electric
ENGR	Engineer/Design
MFGR	Equipment Manufacturer
EXCV	Excavator
LIQPIPE	Liquid Pipeline
LOCA	Locator
NATGAS	Natural Gas
PRVW	Private Water
PUBW	Public Works
RAIL	Railroad
ROAD	Road Builders
REGU	Federal/State Regulator
TELC	Telecommunications
UNKNOWN	Unknown/Other

Root Causes (up to and including 2017)

ROOT CAUSE GROUP	DAMAGE_CAUSE_ABBR	DAMAGE_CAUSE_DESCRIPTION
	NOLOCATEREQ	No notification made to one call center/811
NOTIFICATION	INSUFCALL	Notification to one call center made but not sufficient
	WRONGINFO	Wrong information provided
	ABANDONED	Abandoned facility
	NOTLOCATED	Facility was not located or marked
LOCATING	INSUFMARKING	Facility marking or location not sufficient
	BADMAP	Not marked due to incorrect facility records/maps
	NOTFOUND	Unlocatable facility
	EXTESTHOLE	Excavator dug prior to verifying marks by test hole (pot hole)
	EXCLEARANCE	Excavator failed to maintain clearance after verifying marks
	EXMARKS	Marks faded, lost, or not maintained
EXCAVATING	EXSUPPORT	Excavator failed to shore excavation/support facilities
	EXHANDTOOL	Failure to use hand tools where required
	EXBACKFILL	Improper backfilling
	INSUFEX	Improper excavation practice not listed above
	CALLCENTER	One call center error
MISCELLANEOUS	DETERIORATED	Deteriorated facility
	PREVDAMAGE	Previous damage

Root Causes (Since 2018)

ROOT CAUSE GROUP	DAMAGE_CAUSE_ABBR	DAMAGE_CAUSE_DESCRIPTION
	NOLOCATEREQ	No notification made to one call center/811
	EXDUGOUT	Excavator dug outside area described on ticket
NOTIFICATION	EXDUGBEFORE	Excavator dug prior to valid start date/time
	EXDUGAFTER	Excavator dug after valid ticket expired
	EXBADINFO	Excavtor provided incorrect notificaiton information
	NORESPLOC	No response from operator/contract locator
	NOMARKABAND	Not marked due to abandoned facility
	BADMAP	Facility marking or location not sufficient
	LOCERROR	Not marked due to incorrect facility records/maps
	TRACEWIRE	Not marked due to tracer wire issue
LOCATING	INCOMPLETE	Site marked but incomplete at damage location
	NOTFOUND	Unlocatable facility
	INACCABAND	Marked inaccurately due to abandoned facility
	INACCBADMAP	Marked inaccurately due to incorrect facility records/maps
	INACCLOCERR	Marked in accurately due to locator error
	INACCTRACEW	Marked inaccurately due to tracer wire issue
	EXTESTHOLE	Excavator dug prior to verifying marks by test hole (pot hole)
	EXCLEARANCE	Excavator failed to maintain clearance after verifying marks
	EXMARKS	Marks faded, lost, or not maintained
EXCAVATING	EXSUPPORT	Excavator failed to shore excavation/support facilities
	EXBACKFILL	Improper backfilling
	INSUFEX	Improper excavation practice not listed above
	CALLCENTER	One call center error
MISCELLANEOUS	DETERIORATED	Deteriorated facility
	PREVDAMAGE	Previous damage
UNKNOWN	NOTCOL	Root Cause not listed above (comment required)