

Last update: 28.08.2024

Guidelines for Validation Data Table Creation

General guidelines

1. For information about an event to be entered into the data table, it must have a location, time (at the least should have the Start_Date_Year) and impacts information for at least one impact category.
2. We adopt 3 levels information about the event, Level 1(L1), Level 2 (L2), and Level (3) in the database as schema shown in Figure 1, and below we give a detailed description of them.
 - L1 presents the overall impact in the event with the basic information that includes time, location (only at country_level), main_event and hazards, which covers all numeric impacts (deaths, injuries, homeless, displaced, affected, and buildings damaged) and monetary impacts (damage and insured damage). In the validation table, if the article doesn't mention the total impacts, please fill with "NULL", don't sum up numbers in the validation table to present total in L1. The main_event and hazards relationship please refer the Table 1. And the definition of all impact categories, please refer Table 2.
 - L2 presents the overall impact in countries, differ from L1, most of the case, L2 contains overall impact information affecting an individual country. However, some cases, the impact may cross several countries, which needs to record in L2. In the validation table, don't sum up numbers from the impact cross different cities in one country to present in the L2, "NULL" should be used. Hazards is not included in this level.
 - L3 presents the impact in specific locations within one country, Hazards is not included in this level.
 - Special notes for L2/L3
 - If the location of two L2/L3 records is the same, but the time of occurrence is different, these will count as two L2/L3 records. For example, 3 people dead at Brussels on 2023/12/15, and later 5 people were found dead at brussels on 2023/12/30. In the table, should record: [2023/12/15, Brussels, 3], [2023/12/30, Brussels, 5] (this is not the same format as data table, just for reference)
- If any information is missing, please use "NULL"

Table 1 main event and hazards relationship

Main_Event	Hazards
Flood	Flood
Extratropical Storm/Cyclone	Wind; Flood; Blizzard; Hail
Tropical Storm/Cyclone	Wind; Flood; Lightning
Extreme Temperature	Heatwave; Cold Spell
Drought	Drought
Wildfire	Wildfire
Tornado	Wind

Table 2 definition of all impact categories


Variable	Definition
Deaths	The number of deaths in the event is determined by focusing on words in the text including "die", "dead", "killed", "fatality", "lost lives", "perished", and "passed away". Missing people are not included as deaths.
Injuries	The number of non-fatal injuries in the event is determined by focusing on words in the text including "injure", "hurt", "wound", and "hospitalized".
Homeless	The number of homeless people in the event is determined by focusing on words in the text including "lost home", "homeless", "household damage", "household destroy", "house damage", "home destroy", "unhoused", "without shelter", "houseless", and "shelterless".
Displaced	The number of displaced people in the event is determined by focusing on words in the text including "evacuated", "displace", "transfer/move to shelter", "relocated", and "flee".
Affected	The number of affected people in the event is determined by focusing on words in the text including "affect", "impact", and "influence".
Insured Damage	The damage refers to the physical harm or loss to property, assets, or individuals covered under an insurance policy in the event, with the terms "insurance" or "insured" explicitly mentioned in the text.
Damage	The economic damage caused by the event.
Damaged buildings	The number of buildings damaged by the event is determined by identifying words in the text including "home", "house", "household", "building", "apartment", "apartment block", "school", "church", "office buildings", "retail stores", "hotels", "hospitals", "dwellings", and "structures".


- Any date information entered into the data table MUST be represented as three columns with suffixes *_Date_Day, *_Date_Month, and *_Date_Year. Representing the date in an excel cell in a specific date format is thus not needed (neither is it recommended) since we want to allow for *_Date_Day and *_Date_Month to be optional fields. The date format must be DD, MM, and YYYY (respectively) where days and months only have two digits, and years always have 4 digits. The start year is always mandatory. All other date fields are optional.
- When date information is provided as text, such as "12th October" for the date a storm impacted Europe, then convert that textual information into the three date columns with suffixes *_Date_Day, *_Date_Month, and *_Date_Year (the missing date information, the year in this example, can be presumably inferred from or given elsewhere in the article's text). If the year can be inferred from an article, then enter that into the data table, otherwise if the year or missing date information cannot be inferred, enter NULL into the field. For the sub-event, the year maybe missing, and it can be inferred from the main event.
- For ALL numeric data table columns, such as amounts of insured or total losses, people killed, buildings damaged, number injured, etc., enter only numeric information into the

data table. If the field is numeric, the “amount text” given in the article should be converted into its number. For example, an article may provide total loss information as “estimated losses from the storm are approximately thirty billion USD” or “estimated losses from the storm are approximately 30 billion USD” – in both instances, you would enter the total loss information into the appropriate data table field as 30000000000 rather than “30 billion” or “thirty billion.”. Do not use commas or periods or other separators (such as writing “30 000 000 000” or “30,000,000,000”) if the number is a natural number. For decimals, use a dot.

6. The information that should be included in the source column Source should be the URL(s) of the articles from which the information comes. A source column follows most columns within the data table, as different information may have been found in different articles. If the information for a category was found in two or more articles, then list the URLs in that category’s source column, separated by the vertical bar separator. (|). The URL for an article can be found in its metadata that’s listed to the right of the article that you see when you open an individual article on Doccano, for Wikipedia and Artemis articles.
7. To avoid ambiguity, if no information for a field in the data table is available for an event, then fill in that field in the data table with NULL rather than NaN or a numeral 0 (do NOT use NaN or NA or “not available” or anything like this at all in the table). A field should be filled with the numeral 0 only when we know explicitly from the article at 0 people were killed/injured/losses incurred, etc. For example, if an article says something like “...; however, 0 people died as a result of the flooding...” or “...; however, no people died as a result of the flooding...”, then you should enter the numeral 0 into the Deaths field. However, if the article mentions nothing at all about the number of deaths (in which case we can’t know if the number of people killed was actually 0 or not), then enter NULL into the Deaths field.
8. For populating the data table with the impacts of the extreme events identified in the articles, focus on the numeric data, such as the amounts of insured loss or total loss for the time being. Do NOT populate the data table with any information from the “hazards” or “other impacts” categories, as these are being left out of the data table for the time being.
9. When entering location information, all location information found in the article where the event occurred should be included, and should be ordered from most specific location information to least specific location information, again separated by the vertical bar separator for the national locations for the main event, and, by an ampersand for sub-national events as in, for example, "Stockholm&Sweden" or "Brittany&France".
10. Fill in the start and end dates in their respective columns with their article source URL for each event if given; otherwise, fill in these columns with NULL as you would for the

other fields in the data table. It is totally fine if all two date fields (Start_Date_Day, Start_Date_Month, Start_Date_Year, End_Date_Day, End_Date_Month, and End_Date_Year) are filled with the same date based on the information that was gleaned from the articles, as this can be the case for some events that lasted only a single day, for instance.

11. For the Affected category, enter numeric information only if the article explicitly gives this information. The total number affected presumably includes the previous impact categories (deaths, people made homeless, etc.), but the total number of people affected (if given by the article) could include people affected in ways we have not considered here; a sum of the previous impact categories therefore may not give an accurate total affected number and NULL should be entered instead.
12. For the monetary loss columns (Insured_Damage and Damage), must also fill out the columns following these on the units used for the monetary amounts, and whether they are inflation-adjusted or not.
 - Insured_Damage_Units and Damage_Units refer to the currency in which the loss information is provided. These currency units are standardized by using the 3-letter currency codes in ISO 4217 (such as EUR for euro or SEK for kronor or USD for American dollars, etc.) and entered in their correspondent columns.
 - The information on inflation is split into two columns
 - The*_Inflation_Adjusted column is a simple yes/no column – if the article indicates in any way that the loss amount is adjusted for inflation. When End_Date_Year of the event is 2012 and in the article {damage:“30 million (USD 2012)” }, then this column should be filled in with “No” , because the damage is not adjusted to another year. If the same event, and the article says {damage:“30 million (USD 2020)” }, then fill in this column with “Yes” as it’s adjusted to 2020 with inflation. And once the article only mentions {damage:“30 million (USD) } without year, then it can be assumed that the currency value is for the End_Date_Year of the event, and this field should be filled with “No” .
 - In the *_Inflation_Adjusted_Date_Year column, fill in the year to which the loss amount was adjusted (in the above example, that would be the year 2020). If the units are given just as USD, EUR, or whichever currency is used, but no year information is associated with the currency nor any other inflation indication is given, then it is assumed that the loss amount has not been adjusted and this column should also be filled with NULL.
 - If loss information is provided in multiple currencies, such as loss given in both USD and EUR, the information entered into each field can then be separated with the vertical bar separator (|).
 - Useful documents on how inflation adjustment could be represented:
 - 1) How EMDAT does economic adjustment:
<https://doc.emdat.be/docs/protocols/economic-adjustment/>
 - 2) Wikipedia guidelines for inflation adjustment:
 <https://en.wikipedia.org/wiki/Template:Inflation>

-  <https://en.wikipedia.org/wiki/Template:Inflation/year>
-  A template specifically made for one extreme weather event:
https://en.wikipedia.org/wiki/Template:Infobox_tropical_cyclone_season

13. Follow the same guidelines for Buildings_Damaged as for the other numeric categories in the data table above.
14. In some cases, the numeric information may be expressed in comparative adjectives or quantifiers. The following rules only suit for the case where a single number is presented, if the text already mentions two numbers, it will automatically put in a range without applying the rules below. In these cases, we first, define the scale of a number:

Proposition 0:

The scale of a number **N** is the power of ten corresponding to the last non-zero digits, where:

N = $a \times 10^n$, where

- **a** : **a** is a decimal with all the non-zero digits to the left of the decimal point
- **n** : **n** is an integer
- **scale** : **scale** = 10^n

Note: This is a slight modification of the scientific notation of a number, in which case **a** is a decimal with a non-zero digit to the left of the decimal point.

According to our definition, the scale of

- 500 = 5×10^2 corresponds to $10^2 = 100$

- 560 = 56×10^1 corresponds to $10^1 = 10$

- 543 = 543×10^0 corresponds to $10^0 = 1$

This scale is then used to convert expressions such as “at least 560”, “over 560”, “less than 560” into ranges, according to the following proposition.

Proposition 1:

As a general rule, the length of an interval estimated to the left or to the right of a given number must be equal to the scale of the number. In other words, all departures or shifted departures to each side of the number must be at a distance equal to the scale of the number.

Therefore, the length of a range estimated (the absolute difference between the values in the boundaries) is equal to:

- The scale of the number, in the case of “more than”, “less than” and “at least” and all the synonyms of these expressions. For a given number **N**, and its scale:
 - i. At least $N \rightarrow N \in [N, N+scale]$, # over, inclusive
 - ii. Over $N \rightarrow N \in [N+1, N+1+scale]$ # over, exclusive
 - iii. Less than $N \rightarrow N \in [N-1-scale, N-1]$. # under, exclusive
 - iv. At most $N \rightarrow N \in [N-scale, N]$ # under, inclusive

- Twice the scale, in the case of “about” and all the synonyms of this word
 - About $N \rightarrow N \in [N\text{-scale}, N\text{+scale}]$
- In the case of “about N” and “less than N”, the lower boundary of the estimated interval does not include zero, but stops at 1. If an impact is reported, it is not zero, even if the exact value is uncertain.

Examples, range estimations:

	500	560	543	3.54e3
At least	[500,600]	[560,570]	[543,544]	[3540,3550]
Over	[501,601]	[561,571]	[544,545]	[3541,3551]
Less than	[399,499]	[549,559]	[541,542]	[3529,3539]
At most	[400,500]	[550,560]	[542,543]	[3530,3540]
about	[400,600]	[550,570]	[542,544]	[3530,3550]

Rules for equivalent expressions and additional expressions are indicated below, through some examples:

- I. Greater than/more than/exceed/over/+ 700: 701-801 ($700 + 1$, $700+1\text{+scale}$), scale = 100
- II. Less/lower/fewer than 700: 599-699 ($700-1\text{-scale}$, $700 - 1$), scale = 100
- III. At least/a minimum of 630: 630-640 (630 , $630 + \text{scale}$), scale = 10
- IV. Up to 270: 260-270 (270-scale , 270)
- V. Approximately/around/nearly/about/almost/roughly/~ / estimated 700: 600-800 (700-scale , 700+scale), scale = 100
- VI. Dozens of, tens of, hundreds of, thousands of, etc: $2*\text{scale} - 9*\text{scale}$; so, “thousands of injuries” becomes 2000-9000 injuries.
 - Tens: scale = 10
 - Dozens: scale = 12
 - Hundreds: scale = 100
 - Thousands: scale = 1000
- VII. A number of/a group of/a few/several/numerous: 2-6
- VIII. A few dozen: 24-72 ($12*2$, $12*6$)
- IX. A dozen hundreds (if it ever appears!): 2400-7200 ($12*2*\text{scale}$, $12*6*\text{scale}$)
- X. A few/several hundred/thousand/million, etc: $2*\text{scale}-6*\text{scale}$; so “several millions” becomes 2000000-6000000.
- XI. Many: 20-60
- XII. SPECIAL CASE: A couple/a couple hundred/thousand, etc...: $2*\text{scale} - 3*\text{scale}$; so “a couple of deaths” becomes 2-3 deaths
- XIII. SPECIAL CASE: If the number of human victims is reported as “family/families”, multiply by 5 to determine the number of human victims. For “families”, use $2*5-9*5$ if there is no number before families. For example: “5 families are displaced” = $5*5 = 25$ people displaced, “dozens of families were evacuated” = $2*12*5-9*12*5 = [120,540]$ people displaced.

- XIV. SPECIAL CASE: "minimal", "inconsequential", "negligible", "minor", and "limited", put "NULL" in the validation table.
- XV. SPECIAL CASE: no causalities, no fatalities, no injuries, none, none reported, and similar expressions must be annotated as 0. If the information is missing, the annotator must enter NULL instead.

Additional Notes on the vertical bar separator (|):

The vertical bar separator is used in several fields in addition to the Location field:

- Event_Names (for Wikipedia, this is the article title, so you will have to use the vertical bar separator with non-Wikipedia articles)
 - Sources (for non-Wikipedia articles, you may find several sources for the same Event or per impact season (such as the "Hurricane Season")
 - Hazards (for cases with more than one hazard)
 - Insured_Damage
 - Insured_Damage_Units
 - Insured_Damage_Inflation_Adjusted
 - Insured_Damaged_Inflation_Adjusted_Date_Year
 - Damage
 - Damage_Units
 - Damage_Inflation_Adjusted
 - Damage_Inflation_Adjusted_Date_Year
1. In the case of Event_Name, the vertical bar separator (|) denotes alternative names for the same event
 2. In the cases of Insured_Damage and Damage (and all the associated fields for the currency units and inflation adjustment information), the vertical bar separator (|) denotes that the damage amounts were provided in more than one currency for the same event.
 3. Sometimes articles provide the insured and/or total damage amounts as a range, so the damage information is entered into the data table as a range (with format ###-###). You may enter multiple ranges with multiple currencies; an example with two currencies and two amounts that are ranges: "3000000-4000000|20000-25000" where the corresponding *_Units, *_Inflation_Adjusted, and *_Inflation_Adjusted_Date_Year would follow the same format: "TRY|WST", "Yes|Yes", and "2012|2013" (respectively). Any data separated by the vertical bar separator (|) must be in order. In the previous example, the first range represents data in the "TRY" (Turkish lira) currency and is inflation adjusted for the year 2012.

Schema

the L1-3 schema is shown below, and the please review this link for a clear version of the schema, and example: <https://drive.google.com/file/d/112dvJnSa8rGT2ZrM-gvimo83DNMRtHK0/view?usp=sharing>

SCHEMA / prototype

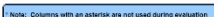


Figure 1 Prototype Schema of L1-3

Guideline used in the NLP2024 paper

In the paper, we have Time, Location (only country), Total_Deaths, Total_Damage and Event_Category(Main_Event) evaluated. For numerical data such as deaths and damage, we primarily obtain the total impact information from the Wikipedia infobox. If the infobox does not contain this information, or when working with Artemis articles, we sum the individual impact numbers to present the total in the gold data. We adopt the normalization rules for unclear numerical description as follows:

Single number:

“None reported” : 0

“At least 1,152” : 1152

“EUR54 billion” : 54000000000

Ranges:

If two numbers appear in the text, we assume that they represent a range, for example, 2-5 people dead: [2,5]

Special cases:

Dozens of, tens of, hundreds of, thousands of, etc: $2 * \text{scale} - 9 * \text{scale}$; so, “thousands of injuries” becomes 2000-9000 injuries.

- Tens: scale = 10
- Dozens: scale = 12
- Hundreds: scale = 100
- Thousands: scale = 1000

The definition of deaths and damage in the paper:

- i) Death: die, dead, killed, fatality, lost lives, perished, passed away. We do not include missing people as deaths, but it is important to note that EM-DAT does.
- ii) Damage: total economic damage

Keywords for definition of the impact categories, freeze on 20240828

The following keywords must be followed in the data table annotation, especially when determining how to add up numerical data. Below is a list of keywords for each impact category:

- Death: die, dead, killed, fatality, lost lives, perished, passed away. We do not include missing people as deaths, but it is important to note that EM-DAT does.
- Homeless: lost home, homeless, household damage, household destroy, house damage, home destroy, unhoused, without shelter, houseless, shelter less
- Displace: Evacuated, displace, transfer /move to shelter, relocated, flee
- Injure (non-fatalities): injure, hurt, wound, hospitalized
- Affected: affect, impact, influence
- Insured damage: refers to the physical harm or loss to property, assets, or individuals that is covered under an insurance policy. When such damage occurs, the insurance company is obligated to compensate the policyholder according to the terms and conditions of the policy. Annotators must ensure that the word “insurance/insured” is included in this damage information to qualify for this category. Otherwise, it belongs to “total damage”.
- Total damage: total economic damage
- Building damage: home, house, household, building, apartment, apartment block, school, church, office buildings, retail stores, hotels, hospitals, dwellings, structures.

Old main-sub event schema

An "Event" (referred to as “main event” in this document) is an extreme climate event that occurs at least one location with at least time information available for the start and end year (which can be identical), and at least one impact. For the impact like total_deaths in the event, once the article only mentions one instance of death but not explicitly mentions as total, which will not use for presenting the total, and if the article contains several instances, the gold data should not sum them up as the total (for evaluation purpose)

-An "Event" can affect several different locations at different times. We must annotate each specific impact instance of these “sub-events” found in the article. There are several types of sub-events impacts:

- Deaths
- Injuries
- Displaced
- Affected
- Homeless
- Buildings_Damaged
- Insured_Damage
- Damage

- Each main event can have zero or more sub-events
- For any category where data is missing in the source article, the annotators must write NULL.
- In the flat-format data table, sub-event refers to the event with a confirmed location, and with one or more impacts. If the location of two sub-events is the same, but the time of occurrence is different, these will count as two sub-events, In the structured database (see the schema shown at the end of the document): for each main event,
 - Total_*_Per_Country columns contain the summary of one impact in the specific country, notice that this requires the information directly extracted from the article instead of adding numbers up, distinguish it from the Specific_Instance_Per_Country_* columns as below.
 - the Specific_Instance_Per_Country_* columns contain only one impact per row which occurs in a specific location (finer than country, if possible) and only for that specific type of impact
- When entering data into the data table, main events and sub-events all share a single sheet with a flat column format.
- Row entries in the data table should be split up both by data source and by national vs sub-national locations (so information for the same locations have been taken from two different data sources and are thus given in different rows), There is no limit on the number of sub-events that an event can have, it is just important that the information within each row for an event/sub-event come from the same data source and apply to the same location information.

Post-processing

The diagram above shows an example of how the post-processing will split “main events” from “sub-events”, especially since that all main and sub-events flat-format data table share the same column and are only distinguishable by their Event_ID. Notice how there is no entry for sub-event 5.01 in the Specific_Instance_Per_Country_Deaths table since no fatalities occurred for that sub-event. Also notice how sub-event 5.01 ends up in Specific Instance tables for both Deaths and Injuries.

After annotation, the data in the flat-format data table is post-processed and split across the Events table and several Specific_Instance_Per_Country_* tables (as shown in the schema below and the diagram above) to make the evaluation process possible. In the flat-format data table, a single row can represent several sub-event impact categories. When the data is post-processed, each sub-event impact category is separated and inserted into its corresponding table. Since all numerical fields (like Deaths or Injuries) is a range, it is split by a dash (-) into two columns: *_Min and *_Max during the post-processing.

The schema diagram represents what the excel data table sheet is transformed to. A main event and sub-event share some columns (such as Event_ID which is shared across all sub-event types). In the schema diagram, these shared columns have identical names.

It’s good to know that this SQL schema can always be “rejoined” to re-create the flat format data table.

Schema (click here for [online version, interactive with notes](#))

