Suggestions preprocessing data

By Hielke Muizelaar

# Media dataset

**Fields**

* **id-article-new**
* **id-article**
* **Event-id-country**
* **Event-id-no-country**

These fields signify general ID’s to differentiate between rows and the country of the respective articles. In my opinion these are not needed for us as we have other ways to find out the country and we can differentiate between the rows ourselves.

* **Event Date 1**
* **Event Date 2**

In my opinion these fields should be kept as they can be used to find out when the events the article describes happened. We could use this to show diagrams/charts by date.

* **Event-Year**
* **datetime precision**

We can obtain the year from the event date ourselves so that one is unnecessary. Regarding the datetime precision, I don’t think we need that either as the Event Date should give a clear enough indication.

* **Event-Type 1 - 4**
* **Event-Type-health 1 - 4**
* **Event-Type-Social 1 - 7**
* **Event-Type-Movements 1 - 6**
* **Event-Type Security 1 - 4**

Very handy so we can group data by type and show different statistics per type. Within the datasets there are a lot of fields that signify type. As the articles typically only have 1-2 types there are a lot of empty values within these columns. In order to properly obtain the articles per type we could split the data such that every type of every article gets its own row. For example, if an article has the types “Health” and “Social” we could split the row into two rows with the same information but making one row have “Health” in a “Type” column and the other “Social”.

* **Number of infected migrants**
* **Number of migrant in quarantine**
* **Number of deaths at sea**
* **Number of deaths when crossing borders**
* **Number of rescued people when crossing borders**
* **Number of arrived people by sea**
* **Number of arrived people through boarders**
* **Number of departed people by sea**
* **Number of departed people through boarders**
* **Number of migrants deaths by COVID -19**
* **Number of migrants healed of COVID-19**
* **Number of people deported on sea**
* **Number of people deported through borders**

Could also be very handy for visualization but a potential problem is that a lot of these fields are rarely not empty. Another potential problem is that the data is not strictly numeric, some examples include “80 per week”, “1268(2016) / 2125(2017) / 2127(2018) / 1739(2019)” and “3800”. We would need to be careful parsing this data. My suggestion would be that if we choose to include these data points we would need to figure out how to make all the values numeric in order to be able to visualize them properly.

* **Actors 1-3**
* **Sub-actors 1-7**

Shows what kind of people are the subject of the article. As these can be very interesting for visualization (such as visualizing the amount of articles per day concerning migrants) I think it would be good to keep them. Again, there are a lot of fields in this category. Perhaps we could take a similar approach as I proposed with the types, where we turn one row into multiple rows where the only difference is the name of the actor/subactor.

* **City 1**
* **Latitude-city1\_degrees**
* **Latitude-city1\_minutes**
* **Latitude-city1\_seconds**
* **Latitude-city1\_direction**
* **Longitude-city1\_degrees**
* **Longitude-city1\_minutes**
* **Longitude-city1\_seconds**
* **Longitude-city1\_direction**
* **City 2**
* **etc.**
* **Geo-precision**

A lot of information on the cities that are prevalent in the articles. The city indicates the city the article is about and the latitude and longitude of that city are signified in great detail. I found that if two articles occur in the same city they have the same coordinates in these fields, so the coordinates do not specify specific areas in a city but just the city in general. In my opinion it would be interesting to create map visualizations of the articles, so we could explore which countries/regions tend to report on what types of events. Similarly, we could explore for example which regions report on migrants the most.

* **Source-Country**
* **Source-date**
* **Source-Type**
* **Language**
* **Source-Name**

These fields show information on the source of the article. I think these fields would be interesting to keep for the visualizations, as we can look at potential bias in reporting regarding the area of publication. I don’t think we need to include the name of the source but the language and type could be indicators of bias. It could be for example that Tunisian radio reports more on migrants than Libyan radio and that might be interesting to visualize.

* **Notes**

Not important in my opinion.

So in my opinion the fields to keep are:

* **Event Date 1**
* **Event Date 2**
* **Event-Type 1 - 4**
* **Event-Type-health 1 - 4**
* **Event-Type-Social 1 - 7**
* **Event-Type-Movements 1 - 6**
* **Event-Type Security 1 - 4**
* **Number of infected migrants**
* **Number of migrant in quarantine**
* **Number of deaths at sea**
* **Number of deaths when crossing borders**
* **Number of rescued people when crossing borders**
* **Number of arrived people by sea**
* **Number of arrived people through boarders**
* **Number of departed people by sea**
* **Number of departed people through boarders**
* **Number of migrants deaths by COVID -19**
* **Number of migrants healed of COVID-19**
* **Number of people deported on sea**
* **Number of people deported through borders**
* **City 1**
* **Latitude-city1\_degrees**
* **Latitude-city1\_minutes**
* **Latitude-city1\_seconds**
* **Latitude-city1\_direction**
* **Longitude-city1\_degrees**
* **Longitude-city1\_minutes**
* **Longitude-city1\_seconds**
* **Longitude-city1\_direction**
* **City 2**
* **etc.**
* **Source-Country**
* **Source-date**
* **Source-Type**
* **Language**

# Interview dataset

**Fields**

* **Number of interview**
* **Number of interviewee**

We could use these fields to differentiate between interviews and people. Some rows belong to the same interviewees and some rows to the same interview. We could combine data that concerns the same person/interview.

* **How many people interviewe**
* **interviewee (migrant/source person)**
* **precision migrant (migrants, asylum seeker, refugee)**

Data on the people that were interviewed. The second field indicates the type of person that was interviewed, such as a migrant or a source person. The precision helps to further divide the interviewees into groups. We can use these types similarly to types in the media dataset where we group data by type and show statistics that way.

* **date**
* **place (country)**
* **place (city)**
* **shelter precision (personal home, home rented by UNHCR, UNHCR foyer, UNHCR transit center ghetto, IOM center)**
* **latitude degrees**
* **latitude minutes**
* **latitude seconds**
* **latitude direction**
* **longitude degrees**
* **longitude minutes**
* **longitude seconds**
* **longitude direction**

In contrast to the media dataset the coordinates here do differ even though the cities are the same. So there could be two rows of interviews held in Tunis and they would have different coordinates. This means we could most likely create more accurate maps if we choose to do so. We could use the shelter precision as a means of differentiation between rows.

* **Ill (yes/no)**
* **How many people known with Covid-19 symptoms**
* **How many people known who couldn't access to hospital because of Covid**
* **Access to healthcare (good/bad)**
* **fear of Covid (yes/no)**
* **difficult economic situation due to lockdown (yes/no)**
* **received help (yes economical, yes sanitary, yes but not enough, no)**
* **received help (yes economical, yes sanitary, yes but not enough, no)**
* **received help (yes economical, yes sanitary, yes but not enough, no)**
* **Name of NGO or other who helped**
* **Name of NGO who helped 2**
* **Name of NGO who helped 3**
* **willing to depart (no, legally to Europe, illegaly through Libya, illegaly through the sea in Tunisia, illegaly through Algeria, rerturn home)**
* **reported having left after the interview (no, yes to Libya, yes to Europe, tried to cross the sea but didn't succeed, yes to another city)**
* **precision**

A lot of fields on questions that were asked during the interview. We confirmed that the questions were not asked to every interviewee, causing some fields to be empty sometimes. The “How many people known with Covid-19 symptoms” field is always 0 or empty. The “How many people known who couldn't access to hospital because of Covid” field is only filled twice. The rest of the data is useful for visualization I think. There are three fields named “received help (yes economical, yes sanitary, yes but not enough, no)”, I don’t know why this is the case or what the differences between these fields are if there are any. The “precision” field contains a lot of information per row, sometimes in English and sometimes in French. We might need to create more fields based on the contents of this field.

In my opinion we should keep these fields:

* **Number of interview**
* **Number of interviewee**
* **How many people interviewe**
* **interviewee (migrant/source person)**
* **precision migrant (migrants, asylum seeker, refugee)**
* **date**
* **place (country)**
* **place (city)**
* **shelter precision (personal home, home rented by UNHCR, UNHCR foyer, UNHCR transit center ghetto, IOM center)**
* **latitude degrees**
* **latitude minutes**
* **latitude seconds**
* **latitude direction**
* **longitude degrees**
* **longitude minutes**
* **longitude seconds**
* **longitude direction**
* **Ill (yes/no)**
* **fear of Covid (yes/no)**
* **difficult economic situation due to lockdown (yes/no)**
* **received help (yes economical, yes sanitary, yes but not enough, no)**
* **received help (yes economical, yes sanitary, yes but not enough, no)**
* **received help (yes economical, yes sanitary, yes but not enough, no)**
* **Name of NGO or other who helped**
* **Name of NGO who helped 2**
* **Name of NGO who helped 3**
* **willing to depart (no, legally to Europe, illegaly through Libya, illegaly through the sea in Tunisia, illegaly through Algeria, rerturn home)**
* **reported having left after the interview (no, yes to Libya, yes to Europe, tried to cross the sea but didn't succeed, yes to another city)**
* **precision**