**Datasheet: Emergency Events Database (EM-DAT)**

Original authors/creators: *EM-DAT, CRED / UCLouvain, Brussels, Belgium*

Organization: *Centre for Research on the Epidemiology of Disasters (CRED)*

Source: [*www.emdat.be*](http://www.emdat.be)

This datasheet was edited by: *Eric Banzuzi, Lena Feiler, Marie Picquet*

**Motivation**

*The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.*

1. **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

*To create a database on natural and technical disasters around the world from 1900 to present in order to provide valuable data for research and decision making for international organizations.*

1. **Who created this dataset (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)**?

*The dataset was created by the Centre for Research on the Epidemiology of Disasters (CRED) at the School of Public Health of the Université catholique de Louvain located in Brussels, Belgium.*

1. **What support was needed to make this dataset?** (e.g. who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)

*The dataset was created with the support and funding of the World Health Organisation (WHO) and the Belgian Government.*

1. **Any other comments?**

*No.*

**Composition**

*Dataset creators should read through the questions in this section prior to any data collection and then provide answers once collection is complete. Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU’s General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.*

1. **What do the instances that comprise the dataset represent (e.g. documents, photos, people, countries)?** Are there multiple types of instances (e.g. movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

*The instances in the complete dataset are international disasters of large impact in the years 1900 to present. The subset of the dataset that will be analyzed in this project only contains all Climatological, Geophysical, Hydrological, and Meteorological natural disasters from 1960 to the present day. From here on we will analyze this subset. For natural disasters, there exist 10 types: Drought, Earthquake, Extreme temperature, Flood, Glacial lake outburst, Landslide, Mass movement, Storm, Volcanic activity and Wildfire.*

1. **How many instances are there in total (of each type, if appropriate)?**

There are 14089 instances of disasters in the data set.

1. **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g. geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g. to cover a more diverse range of instances, because instances were withheld or unavailable).

*The dataset contains all instances that are from the subgroups that the group has selected to investigate for this project.*

1. **What data does each instance consist of?** "Raw" data (e.g. unprocessed text or images) or features? In either case, please provide a description.

*Each instance consists of 50 different features: These describe the type of disaster, location where it occurred, the time during which it occurred, its impact (economic damages, Death toll, number of affected/ injured people), CPI, and administrative codes.*

1. **Is there a label or target associated with each instance?** If so, please provide a description.

*No.*

1. **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g. because it was unavailable). This does not include intentionally removed information, but might include, e.g. redacted text.

*Yes, some of the information is missing due to it not being recorded or relevant for a specific disaster instance.*

1. **Are relationships between individual instances made explicit (e.g. users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

*No, there is no specific relationship between the different events.*

1. **Are there recommended data splits (e.g. training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

*No, there are no recommended data splits.*

1. **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

*This will require further investigation.*

1. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g. websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g. licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

*The dataset is not linked and only relies on updates from official sources. Thus only new instances change the dataset, the old instances will remain constant.*

1. **Does the dataset contain data that might be considered confidential (e.g. data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.

*No, the dataset is available to the public as it does not contain information about private individuals. However, it is limited for commercial use.*

1. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

*No, it does not contain such data.*

1. **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

*No, the dataset does not relate to people.*

1. **Does the dataset identify any subpopulations (e.g. by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

*Skipped*

1. **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.

*Skipped*

1. **Does the dataset contain data that might be considered sensitive in any way (e.g. data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

*Skipped*

1. **Any other comments?**

*No*

**Collection**

*As with the previous section, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals of the prior section, the answers to questions here may provide information that allow others to reconstruct the dataset without access to it.*

1. **How was the data associated with each instance acquired?** Was the data directly observable (e.g. raw text, movie ratings), reported by subjects (e.g. survey responses), or indirectly inferred/derived from other data (e.g. part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

*The data is gathered from various sources, including UN agencies, non-governmental organizations, insurance companies, research institutes and press agencies, which collect the data from records generated during the disaster period (manual human curation). EM-DAT also uses automated processes to collect the data. In fact, satellite imagery is used to identify the extent of flooding or other natural disasters. Furthermore, relevant information from news articles and social media posts are also identified using natural language processing.*

*To validate the accuracy of the data collection, EM-DAT proceeds quality control checks (cross checking information from different resources and comparing historical data), peer review by experts in the field and user feedback.*

1. **What mechanisms or procedures were used to collect the data (e.g. hardware apparatus or sensor, manual human curation, software program, software API)?** How were these mechanisms or procedures validated?

*The data was collected through human curation and automated processes. The validation occurs through quality control checks as mentioned above.*

1. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g. deterministic, probabilistic with specific sampling probabilities)?**

*The data is not a sample.*

1. **Who was involved in the data collection process (e.g. students, crowdworkers, contractors) and how were they compensated (e.g. how much were crowdworkers paid)?**

*UN agencies, non-governmental organizations, insurance companies, research institutes and press agencies were involved in the collection process, compensated by their respective agencies.*

1. **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g. recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.

*The dataset that will be investigated contains data from 1960 to the present. The complete dataset contains data from 1900 to the present, and thus is of larger size.*

1. **Were any ethical review processes conducted (e.g. by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

*We did not find any officially conducted ethical review of this dataset.*

1. **Does the dataset relate to people?** If not, you may skip the remainder of the questions in this section.

*No, the dataset does not relate to people.*

1. **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g. websites)?**

*Skipped*

1. **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

*Skipped*

1. **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

*Skipped*

1. **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

*Skipped*

1. **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g. a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

*Skipped*

1. **Any other comments?**

*No*

**Preprocessing / Cleaning / Labeling**

*Dataset creators should read through these questions prior to any pre-processing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this section are intended to provide dataset consumers with the information they need to determine whether the “raw” data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a “bag-of-words” is not suitable for tasks involving word order.*

1. **Was any preprocessing/cleaning/labeling of the data done (e.g. discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remainder of the questions in this section.

*Yes, the dataset is only a subset of the original data from the em-dat website, from which certain instances have been removed. This mainly concerns disaster types or groups that are not relevant to the questions that will be investigated within the parameters of this project.*

1. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g. to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.

*The original data is still accessible on the website that the data set was taken from and can thus be used in the future if the need arises.*

1. **Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.

*No.*

1. **Any other comments?**

*No.*

**Uses**

*These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.*

1. **Has the dataset been used for any tasks already?** If so, please provide a description.

*Yes, the data set has been widely used for private and commercial purposes. It has been used in research (disaster, climate change) and for journaling and news reporting purposes.*

1. **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

Yes: [*https://www.emdat.be/publications*](https://www.emdat.be/publications)

1. **What (other) tasks could the dataset be used for?**

*We want to use the data for predicting future natural disasters and identifying regions with the most catastrophic outcomes in the next few years, as well as predicting the magnitude of the damages.*

1. **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g. stereotyping, quality of service issues) or other undesirable harms (e.g. financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

*No.*

1. **Are there tasks for which the dataset should not be used?** If so, please provide a description.

*It is important that the data is not taken out of context and only used for private purposes (in accordance with the em-dat database).*

1. **Any other comments?**

*No*

**Distribution**

*Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.*

1. **Will the dataset be distributed to third parties outside of the entity (e.g. company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

*No. Other entities may query the website directly, if they require to use the data.*

1. **How will the dataset be distributed (e.g. tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

*It will not be distributed.*

1. **When will the dataset be distributed?**

*No.*

1. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

*The distribution of the dataset is done under a commercial license agreement. It is free for public use, but restricted for commercial purposes. If it is used for such purposes, the user has to submit a request which is then processed, and a fee must be payed before usage. The Guidelines and license agreement can be viewed under this link:*

[*https://public.emdat.be/about*](https://public.emdat.be/about)

1. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

*No, not to our knowledge.*

1. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

*Yes, before accessing the data, a user must first create an account and state the purpose for which the data will be used. Depending on this, the access is granted or the case will be placed under closer review before the data can be accessed for usage.*

1. **Any other comments?**

*No.*

**Maintenance**

*As with the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.*

1. **Who is supporting/hosting/maintaining the dataset?**

*The Centre for Research on the Epidemiology of Disasters (CRED) at the School of Public Health of the Université catholique de Louvain located in Brussels, Belgium.*

1. **How can the owner/curator/manager of the dataset be contacted (e.g. email address)?**

*They provide two email addresses for making contact. The database manager (*[*regina.below@uclouvain.be*](mailto:regina.below@uclouvain.be)*) and a general contact email (*[*contact@cred.be*](mailto:contact@cred.be)*). If the user wants to use the dataset for commercial use, they need to contact the manager.*

1. **Is there an erratum?** If so, please provide a link or other access point.

*We did not find such a document for this dataset.*

1. **Will the dataset be updated (e.g. to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to users (e.g. mailing list, GitHub)?

*The em-dat data is internally updated on a daily basis, and the publicly accessible records are updated every 3 months (after validation and cross-checking of different sources).*

1. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g. were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

*The dataset does not relate to people.*

1. **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

*There are no older versions of this dataset, but the current one is continuously updated with new records.*

1. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

*We could not find information on personal contribution to the dataset. Thus, if one wants to contribute, a good strategy would be to support one of the verifiable organisations that em-dat cooperates with. Any contribution form these organisations is verified and cross-checked before publication.*

1. **Any other comments?**

*No*