Documentation of the Dataset

**Introduction to the dataset**

The dataset that is used for this research is a dataset that is collected by the United Nations Framework Convention on Climate Change (UNFCCC) and the EU Greenhouse Gas Monitoring Mechanism. This dataset is used to make an index of the individual emissions of greenhouse gasses of the 28 European Union member states between the period of 1985 and 2014. The dataset displays the annual values of the sectors Energy, Industrial Processes and Product Use, Agriculture, Land-Use Change and Forestry and Waste.

The dataset can be found on the website of the European Environment Agency (EEA) through the following link:

<https://www.eea.europa.eu/data-and-maps/data/national-emissions-reported-to-the-unfccc-and-to-the-eu-greenhouse-gas-monitoring-mechanism-14#tab-european-data>.

The file can be downloaded as a zip-file that contains the actual dataset. The used format for the dataset is a CSV file. The file is about 87 MB, which makes it very hard to work with without a rather fast processor. It is possible to use the dataset for personal researches, however it is copyrighted by the EEA. This means that if you want to use the dataset, you have to comply with the rules of the reuse policy of the EEA. More information about this matter is provided on the website.[[1]](#footnote-0)

**Provenance of the data**

As stated before, the dataset is set up by the UNFCCC and the EU Greenhouse Gas Monitoring Mechanism. The information of the dataset is a collection of submissions of the 28 member states of the European Union. The 28 countries have to annually submit files with about 50 tables in which they report the emission values of the different types of industries of that year. To keep these submissions structured and standardized, every country has to conform to a strict set of rules. This is the reason why the dataset is very structured and very clean.

A problem that is specific to the provenance to this dataset is the fact that it is not possible to access the original reports that were handed in by the 28 countries on Eionet. Eionet is a governmental open big data source. In order to access these Eionet files, one needs a username and a password. We were not able to get one, which made us unable to look at the original reports. This could be seen as a problem, because this made us unable to compare the original reports with the information provided in the dataset. Therefore we couldn’t see if the author of the dataset changed certain elements of the original reports before putting it in the final dataset. And it is certain that changes are made because some countries re-submit the reports 2 or 3 times in a year. What they change is not available to the public. Another drawback of the unavailability of these original reports is the fact that we couldn’t see the annotations within the dataset. For instance, if there are no values about the emissions of a certain type of industry every country has to explain the reason why there is no information and submit these in a table. This information can unfortunately only be found in the original reports.

In conclusion it is possible to trace the provenance of the data, but only to a certain point. Because the original data is not authorized for the average user of Eionet, it is not clear what has been done to the data before it entered the dataset. Especially if there are several re-submissions annually. For this reason we have to assume that the given information per country is correct, however, we can not assume that the reported data is factual. So the data is public in name only but is still private in practice, which makes curating the dataset nearly impossible.

**Curation of the dataset**

As explained, there was no curation done to the dataset. It was pretty clean and because of the previously mentioned issue of provenance it was not possible to go through the values and see if these were correctly done. And if we go further, it is not even possible to see if the values are correct because we have no information about the actual data collection.

Nonetheless, we did go through the dataset with Openrefine. It was not only important to at least try to curate, but to also try and enrichen. These were not possible unfortunately. What we did find Openrefine useful for is to go through the dataset to make conclusions on our research question and data analysis. For the research we went through the relevant columns in Openrefine and to find the relevant values. Furthermore, for the data analysis we took an interest to see if there were eventual annotation made that would explain why the original data was not available. For instance, maybe it could be annotated in the dataset why countries resubmitted, why some greenhouse gasses have no values ect. To do this, we used OpenRefine to go through the columns. Annotations were available in the column “Notation”. To see all the types of notations available in the dataset, in openrefine we clicked the downwards arrow, facet, text facet to see all the text available in this column. We found the annotations “C”, “IE”, “NA”, “NO” and a few more. What these notations mean are not immediately clear in the dataset. We found the meaning of these in the IPCC reporting guidelines after some digging on the Eionet website. This goes to show that the dataset has a lot of information, is structured and nicely curated but there are loads of ambiguity still.

1. <https://www.eea.europa.eu/legal/copyright> [↑](#footnote-ref-0)