MoSD Exam Data Management Plan

Table of Contents

general process

1. get an overview of all available material
2. analyse the task for the exam
3. create repository to store all metadata for the exam itself and to additionally add history to the project
4. select one of the given scenarios
    a. I chose the covid scenario because it is still very relevant to all of us and it pairs well with the topic of our ReproHack Review
5. come up with a reasonable research question
    a. i looked at the two datasets and took one of the more straightforware questions
    b. "is there connection between cases/deaths and testing volume?"
6. look detailed at the datasets on the website cases/deaths dataset, testing volume dataset (time frame from start 2020 till end 2023)
    a. pro & cons cases/deaths dataset (ca. 12600 entries)
        i. available in many different formats
        ii. pretty self explainatory csv file structure
        iii. not that much metadata on the website itself
        iv. link to github repo
    b. pro & cons testing volume dataset (ca. 6100 entries)
        i. same as for the first dataset

        ii. but much more metadata on the website

    c. when downloading from the website, both files get saved as *data.csv* so renaming is necessary

7. select and fill out one DMP template
    a. i chose the Horizon Europe template found here
    b. reason: i felt that DMP are very elusive when not working with a konkrete example, so i chose one DMP that resembled those in the exercise and was not to large in size for the project
8. add slides for the exam
    a. development of slides in microsoft office (not ideal)
    b. after finishing the slides they will also be added to the repository as *pptx* and *pdf*
9. go through the whole data lifecycle and add informations to all slides based on the informations learned in the lecture and in the corresponding moodle tiles

data lifecycle

Plan

- mostly covered by the data management plan

Collect

1. Tools/Source für Daten
    a. reported by EU/EEA Member States to the European Surveillance System (TESSy)
    b. when not available, ECDC compiles data from public online sources
2. Automatisches Sammeln von Daten?
    a. no need, data provided by ECDC (European Centre for Disease Prevention and Control)
    b. they automatically or manually retrieved (web-scraped) on a daily basis
3. Define a purpose of data collection
    a. collect data to form a comprehensive answer to the research question
4. Define amount of data
    a. too much data from too many different sources not good either, data quality may get worse
    b. the two datasets have probably enough data for the analysis
5. Data representativeness
    a. data is biased towards europe, no other countries included (could be interesting to compare)
    b. quality or noise of the data can’t be evaluated
6. Cost of collection process

a. collection cost by ECDC is not transparent

b. collection process for this scenario was done by me alone, so no additional costs

7. Cost of storage
    a. data is placed in free github repository so no storage costs
8. Define data collection strategy
    a. no strategy required, maybe a little bit of preprocessing, e.g. select time frames, countries etc.
9. Data is
    a. Quantitative data
    b. structured data
    c. generally trustworthy because of the ECDC

Assure

1. Completeness
    a. see *src/data_quality.ipynb*
    b. only 7.63% of all rows in the deats/cases dataset have values of NA in important rows
    c. but 18.86% of all rows in the testing dataset have NA values
2. Uniqueness
    a. is garantueed because of how the data is structured (one entry for every week for every country)
3. Timeliness
    a. is fairly representative
    b. ofc in a pandemic there is no guarantee that everything is 100% accurate
4. Validity
    a. all columns are valid and concise
    b. if a value is not present, it’s value is NA
5. Accuracy
    a. the data entries are for each week, so there is no date format problem
    b. in general: all columns have values that make sense and can be expected
6. Consistency
    a. good consistency
    b. a minor flaw is that there are entries for countries but also for the EU as a whole. this could lead to some minor missunderstandings
    c. also some country codes don’t match between the two datasets, e.g. Austria (AUT/AT)
    d. the column *year_week* has different format, in one it has a leading "W" for the week-number

In general there are a few multi source problems on the instance level (inconsistent data) but apart from that, there are no major flaws. maybe a single source problem at the schema level with the EU/Country mixup

Describe

1. More metadata from the website itself
2. for the testing dataset there is much more information available
3. github repo doesn’t offer more metadata then the website itself
4. in general more metadata everwhere would help
5. BUT: the data itself is pretty self explainatory even as a non-medical person

Preserve

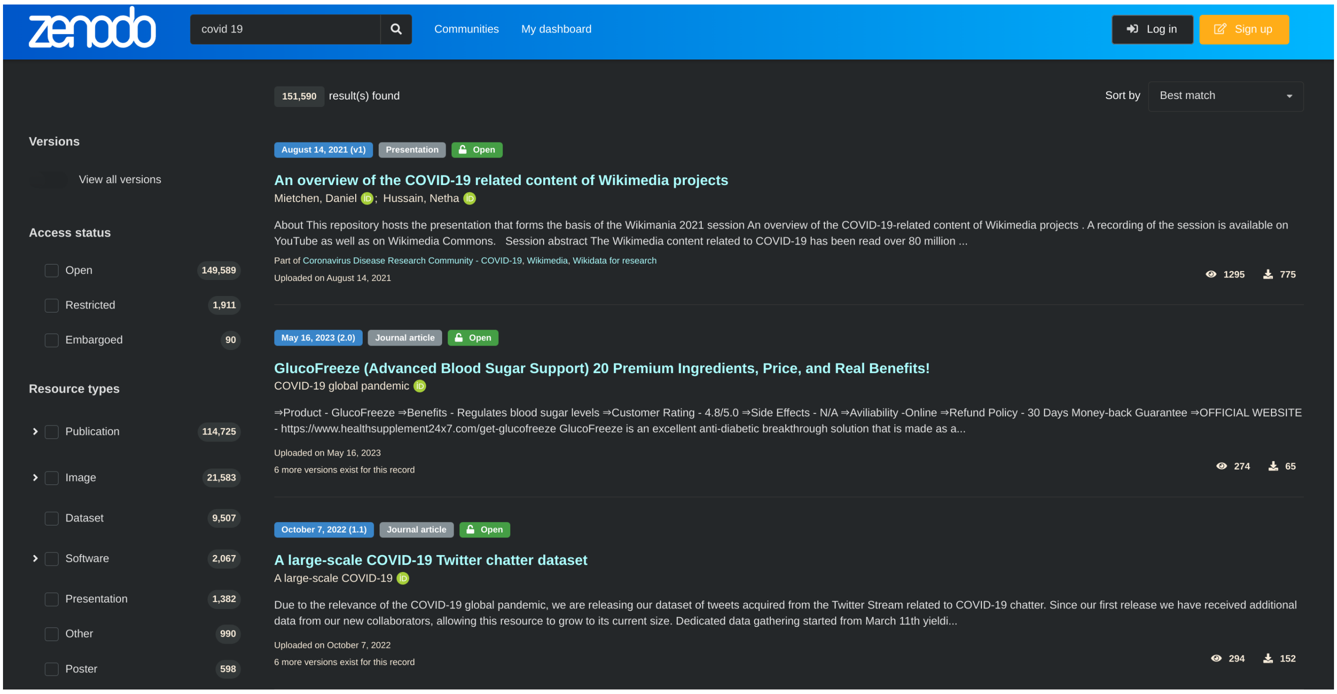
1. the data website is from the ECDC, so that is pretty reliable with backups on github, so there is no single point of failure
2. they could however upload the data to research repository like zenodo
3. there is no indicator that they published a paper with the datasets provided, but they could explicit tell if they "only" preserved or also published the data
4. on the website itself there are no quality features
5. DOI or other PID are not found on the website or the repository
6. authors are not named on the website at all, only the members of the repository given informations about that
7. metadata is present but to lesser extend on informations about preservation/publishing etc.
8. download options are fully supported
9. basic description and documentation is there, but it’s by no means comprehensive
10. the data is freely accessable for everyone
11. there is no explicit licence in the repository but in the website is a link to the ECDC Copyright policy
    a. ECDC has to be acknowledged as original author
    b. The copyright policy of ECDC is compatible with CC BY 4.0 license
12. there is no overview of the data at all, not even the column names
13. archives are present for one time frame 20.06.2022 with additional script for R
14. the repository has 42 commits but the initial commits are on the 01.12.2023, so at the end of the record time frame
15. data is indexed by google and can be found pretty consistent @ 10

FAIR

Most of the FAIR criterias are met by the ECDC datasets. The given informations are easy to find, even though there could be more. Also the Website is indexed on Google and other search engines, so machines/humans can find the data pretty easy. the structure of the data is very good and one can get started pretty fast (considering pre-exisiting knowledge in data science) The data is free accessible without any kind of paywall or required login, but the files could have better naming conventions internally. The data itself is pretty reusable but the repository lack some of the informations available on the website, including the licence.

Discover

1. covid datasets are pretty common, also because of the recency of the topic itself
2. example i found covid-19 vaccines
    a. uses also a 7-day-period, so can presumably work very well with the ECDC datasets
3. also our ReporHack data can be used (topic: impact of fake news on vaccination) but probably harder to integration without extensive preprocessing
4. when searching for covid-19 on zenodo, one gets this overview (or see picture below)
    a. even though there are many articles found, when applying some filter criteria, zenodo only finds 1.297 results, many of wich are for very niche regions or countries
    b. i would've liked the filter for a region to search datasets in, but this seems currently not supported
    c. the dataset itself doesn't have any keywords or terms that may produce search results for similar data.
    d. zenodo probably offers more than only one sorting options, but the standard is *bestmatch* what makes me believe, that it uses some kind of BM25 retrieval system under the hood (but thats just speculation on my side)
    e. searching for *covid-19* or *covid 19* doesn't change the number of results so they probably remove special characters from the query terms
    f. stopwords seem to get indexed by zenodo and remain in there query, so they defenetly make a difference
    g. there also is no stemming happening on the query terms (buy vs. buys)



151,590 result(s) found

August 14, 2021 (v1) Presentation Open

An overview of the COVID-19 related content of Wikimedia projects

Mietchen, Daniel; Hussain, Netha

About This repository hosts the presentation that forms the basis of the Wikimania 2021 session An overview of the COVID-19-related content of Wikimedia projects . A recording of the session is available on YouTube as well as on Wikimedia Commons.   Session abstract The Wikimedia content related to COVID-19 has been read over 80 million ...

Part of Coronavirus Disease Research Community - COVID-19, Wikimedia, Wikidata for research

Uploaded on August 14, 2021

1295 775

May 16, 2023 (2.0) Journal article Open

GlucoFreeze (Advanced Blood Sugar Support) 20 Premium Ingredients, Price, and Real Benefits!

COVID-19 global pandemic

⇒Product - GlucoFreeze ⇒Benefits - Regulates blood sugar levels ⇒Customer Rating - 4.8/5.0 ⇒Side Effects - N/A ⇒Aviliability -Online ⇒Refund Policy - 30 Days Money-back Guarantee ⇒OFFICIAL WEBSITE - https://www.healthsupplement24x7.com/get-glucofreeze GlucoFreeze is an excellent anti-diabetic breakthrough solution that is made as a...

Uploaded on May 16, 2023

6 more versions exist for this record

274 65

October 7, 2022 (1.1) Journal article Open

A large-scale COVID-19 Twitter chatter dataset

A large-scale COVID-19

Due to the relevance of the COVID-19 global pandemic, we are releasing our dataset of tweets acquired from the Twitter Stream related to COVID-19 chatter. Since our first release we have received additional data from our new collaborators, allowing this resource to grow to its current size. Dedicated data gathering started from March 11th yieldi...

Uploaded on October 7, 2022

6 more versions exist for this record

294 152

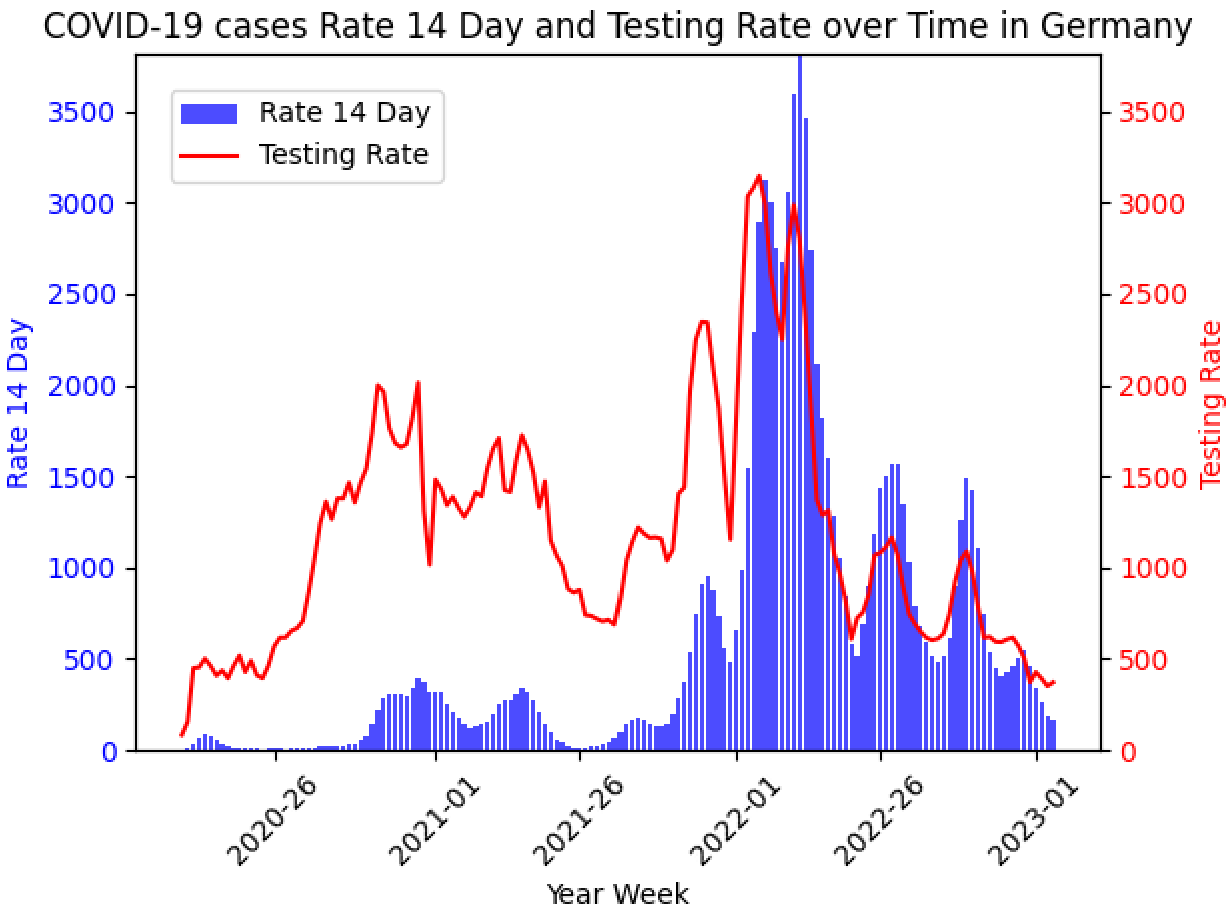
Integrate

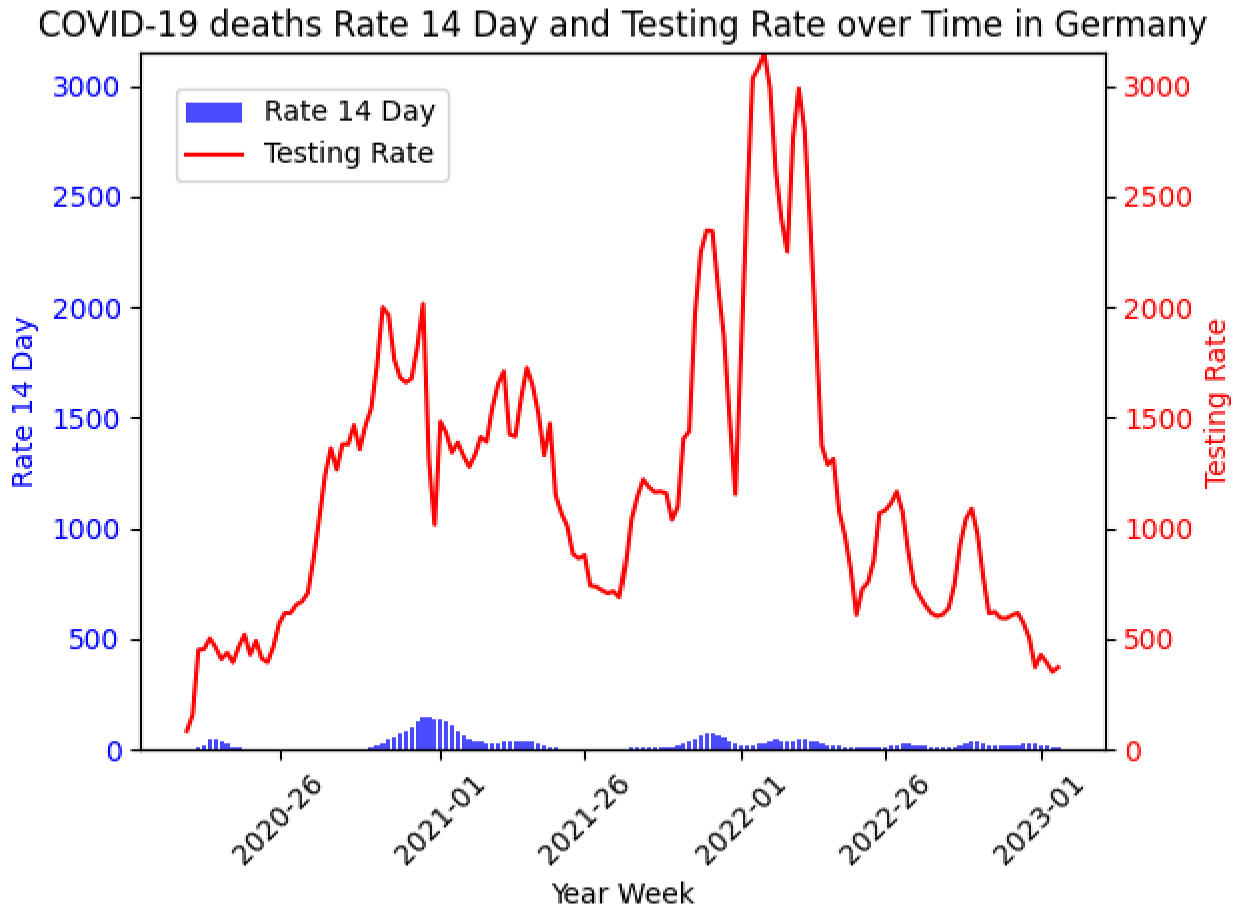
1. in this step, the datasets are processed in a way that makes it easier to analyse them and to find a solution for the research question
2. because the goal is to see if there’s a connection between cases/deaths and testing, the idea is to:
   a. remove the leading "W" in the testing dataset *year_week* column, so the format for both is the same
   b. merge/join the datasets together based on the date
   c. remove the for this task not needed columns
   d. remove NaN rows
   e. export the dataset for every country in the directory `./data/per_country/` (so we have even an additional thing to analyse besides the main research question)
3. this should be achieved with another jupyter notebook *data_processing.ipynb*
4. when doing this part, i worked together with an artifical intelligence assistent to help me write the code, what shortened the developing part drastically

Analysis

1. now comes the analysis itself, which will be demonstrated on one country dataset
   a. but development should be so, that running the analysis for other countries should not take that much time.
2. all developing happends now in the *data_analysis.ipynb* notebook and with help of matplotlib
3. it can be seen, that with rising test cases, also the covid-19 14 day rate rises, so there is a connection between those two factors
   a. this shows, that it is beneficial to test, because many more positive cases will be revealed.

b. this helps to fight the disease





Tools

1. Ubuntu Oracular Oriole (development branch)
2. Lenovo Yoga Slim 7 Pro 14ACH5 O
3. AMD Ryzen™ 7 5800H with Radeon™ Graphics × 16 CPU
4. 16,0 GiB RAM
5. Linux 6.8.0-35-generic Kernel
6. git
7. VS Code 1.88.1
8. Office 365 Web
9. Python 3.12.3
10. Python venv, pandas 2.2.2, ipykernel 6.29.5, matplotlib 3.9.1