Correlation between tobacco smoker and liver cancer incidence 2000-2010 - Initial DMP

1. Data summary

Provide a summary of the data addressing the following issues:

- State the purpose of the data collection/generation
- Explain the relation to the objectives of the project
- Specify the types and formats of data generated/collected
- Specify if existing data is being re-used (if any)
- Specify the origin of the data
- State the expected size of the data (if known)
- Outline the data utility: to whom will it be useful

the purpose of the data collection is to describe the effect of increasing number of smoker in increasing / raising number of liver cancer incidance in alabama state from 2000 to 2010.

the relation to the objectives of the project is that all the data between the period of 2000-2010 , in the alabama state and the cancer site is the liver .

This project produces aggregated dataset in CSV format (Filesize ~371 bytes) that contains data points that combine tobacco usage data with the liver cancer incidance data and a correlation plot of these in PNG format (Filesize (~100K)).

the data will be usefull to the tobacco users to help him to get rid of this bad habit.

Project accesses two external CSV datasets. Both datasets have been downloaded and saved along with the source code in the folder data/raw.

1. Tobacco Usage

2. Liver cancer incidance

(indicator). DOI: 10.1234/khaled.maher File Location: data/transformed/transformed.csv File Size: 371 byte

The experiment has been conducted with byCharm Python Editor. The notebooks contain the experiment's code, accompanying documentation, tables and plots

2. FAIR data

2.1 Making data findable, including provisions for metadata:

- Outline the discoverability of data (metadata provision)
- Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?
- Outline naming conventions used
- Outline the approach towards search keyword
- Outline the approach for clear versioning
- Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

The metadata file can be found inside the project folder (/documentation/metadata.xml).

the DOI identifier is : 10.1234/khaled.maher

the metadata contain the following information about (title , rights , creator , subject , description , date , type , format , source , language , coverage)

there is some search keyword and they all positive keyword "estimated by google keyword tool " like : (1.tobacco usage 2.liver cancer 3.cancer 4.tobacco 5.USA states 6.Alabama 7.smoking 8.smoker 9. 2000-2010 10. tpbacco)

2.2 Making data openly accessible:

- Specify which data will be made openly available? If some data is kept closed provide rationale for doing so
- Specify how the data will be made available
- Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?
- Specify where the data and associated metadata, documentation and code are deposited
- Specify how access will be provided in case there are any restrictions

the resulted data will contain the following data (state , year , number of smoker , number of livercancer incidance) in csv format

the data will be available by sharing / uploading it to the network on github and zenedo website

to access the data all you need a csv viewer , the data is in the folder /SW_Exercise/Raw datasets/Tobacco Usage.csv & /SW_Exercise/Raw datasets/Liver Cancer

Incidance.csv , the resulted file is in the folder /SW_Exercise/Transformed datasets/transformed.csv

and the metadata is in the folder /SW_Exercise/Documentation/metadata.xml

the access to the file is open with no restriction .

2.3 Making data interoperable:

- Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.
- Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

the final documentation is collected and created to describe the effect of increasing number of tobacco smoker in one of USA states on the increasing number of people who is suffering from liver cancer , the data collected from 2 datasets from different websites , the first one is from kaggle and its about number of tobacco usage in USA state from 1995 - 2015 and the other one was collected from data.gov website and its about liver cancer incidance in USA states from 1999 to 2010 , i made a set of procedure on the original datasets to prepare and analyze the data to be useful and can be used the 2 datasets has 2 common column (Year , State) and the resulted file is grouped by these common columns

2.4 Increase data re-use (through clarifying licenses):

- Specify how the data will be licenced to permit the widest reuse possible
- Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed
- Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why
- Describe data quality assurance processes
- Specify the length of time for which the data will remain re-usable

the data's license is : Creative Commons Attribution-NonCommercial-ShareAlike (CC-BY-NC-SA)

there is no data embargo on the resulted data , the produced data can be used by students or another researcher in thier researchs , the produced data has no personal data or sensitive data so there is no restricted data.

the data will be available to be re-used from ending the research untill unknown time .

3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- Estimate the costs for making your data FAIR. Describe how you intend to cover these costs
- Clearly identify responsibilities for data management in your project
- Describe costs and potential value of long term preservation

there is no economical costs , but making data FAIR costs us the time and physical storage space

i dont expect that there is a cost for long term preservation

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

the data is stored in safe repository in github website and also uploaded in Zenedo and i put a username and password so can nobody can access it untill ending the research except the author

but after first publication anyone can use the data and share it and edit it

in other aspects , we prevent data loss by weekly data backup that is free for 28 days and after that it will cost money

also we created a DOI for the resulted data that identify the dataset from other dataset and its :

10.1234/khaled.maher

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

in the research and collecting data we put in account that there isnt any ethically questionable material included just in Tobacco Usage dataset there is a coulmn about sex but i think it will not be a big deal or it will not harm any person , there is no limitation on the image size or resolution just that to be enough to include all person in USA no other country , the dataset is open access and with no limitation

The resulted dataset is open access to all users to access , maintain , use , share and so on , the original datasets thave license of ODBL but the other one has a license but its unknown

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

_