Plan Overview

A Data Management Plan created using DMPonline

Title: Analysis of the distribution of the population of Austria by altitude

Creator: David Wagner

Principal Investigator: David Wagner

Data Manager: David Wagner

Affiliation: Other

Funder: European Commission

Template: Horizon 2020 DMP

ORCID iD: 0000-0002-5142-1054

Project abstract:

In order to get a better understanding of the altitude on which the Austrians live certain data had to be collected and transformed. At this point in time the analysis was only done at the "Bezik" level of Austria. The final result is that ever Bezirk in Austria was connected with the number of inhabitants in this Bezirk and furthermore the altitue of the main city in the bezirk was added.

ID: 39628

Last modified: 22-04-2019

Analysis of the distribution of the population of Austria by altitude - 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**

Purpose
The purpose of the data generation is to enable analysis on which elevation Austrians typically live. As there is currently no data available that links the population of Austria on a granular level (for example Bezirk) to the altitude that the people live at.

Data collected and reused:

- Population Data of Austria per Bezirk
  - https://www.data.gv.at/katalog/dataset/3bfba412-7053-3a60-937a-8c3dd2c71294
    - Unique ID of the data on data.gv.at: 3bfba412-7053-3a60-937a-8c3dd2c71294
  - Three CSV files:
    1. Bezirk Name mapping to internal code
       1. Name: OGD_f0743_VZ_HIS_GEM_4_C-GRGEMAKT-0
       2. Size: 7 Kb
       3. Location: ./data/raw
    2. Year mapping to internal Code
       1. Name: OGD_f0743_VZ_HIS_GEM_4
       2. Size: 8 Kb
       3. Location: ./data/raw
    3. Actual Population data
       1. Name: OGD_f0743_VZ_HIS_GEM_4
       2. Size: 40Kb
       3. Location: ./data/raw
- Altitude Data
  - from DBpedia
    - Accessed directly and stored in a pandas DataFrame
  - Missing values that were not found on DBpedia
    - Was looked up manually on Wikipedia
    - A csv file was created with the data
      - Name: MissingHM
      - Location: ./data/raw
      - Generated manually
      - Format: CSV
        - Non proprietary
      - Size: <2 Kb

Data created:

- Comined table of Population of Austria in 2011 per Bezirk plus the altitude of the main city in the Bezirk.
  - Name: Population2011_altitude
  - Location: ./data
  - Format: CSV
    - Non proprietary
  - Size: <9 Kb
  - DOI: https://doi.org/10.5281/zenodo.2648511

The data will be useful to anybody that wants to work with the elevation data from Austria.

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**

Metadata that instructs on how to run the code is included, as well as metadata that defines the software requirements. Lastly, there is also metadata on the software architecture and a description of it.
The output data will be stored in the following repository:

- Zenodo
  - DOI: https://doi.org/10.5281/zenodo.2648511
  - License: Creative Commons Attribution 4.0 International
  - Availabe since: 22nd of April 2019

The Jupyter Notebook, as well as all the documentation, input files and output were pubished on:

- Github with the integration to Zenodo
  - DOI: https://doi.org/10.5281/zenodo.2648637
  - License: MIT License
  - Availabe since: 22nd of April 2019

For the data files the **naming conventions** that were found at the data source were kept.
Keywords that relate to Austria, Population and Altitude were choosen.
Versioning will only occur every few years, as only then new popualtion data will be available. Then changes to the input data will still not break the Jupyter notebooks, as the old data will still be included.
Versioning will be taken care of on Zenodo
The metadata that will be created is the following:

- README.md
  - explanation of the project and how to reproduce it/ run the code.
- requirements.txt
  - SW requirements that need to be installed in order to run the code
- metadata.xml
  - Metadata describing the tables and the origin of the project
- architecure.png
  - Picture displaying the architecture for preprocessing the data.
- description.txt
  - Short description of the architecture for preprocessing the data.

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**

All of the data will be made openly available, as there is are neither license issues nor is any personal data included.
The input data will be made available with the code on Github. However, it is also availabel on the data.gv.at homepage.
The output data will be made available on Zenodo.
The data will be uploaded to Github and Zenodo and DOI will be assigned.
The only software that is required is a browser, that allows to access Zenodo and Github. The data can even be read by a simple texteditor, as it is just in csv format.

The output data will be stored in the following repository:

- Zenodo
  - DOI: https://doi.org/10.5281/zenodo.2648511
  - License: Creative Commons Attribution 4.0 International
  - Availabe since: 22nd of April 2019

The Jupyter Notebook, as well as all the documentation, input files and output were pubished on:

- Github with the integration to Zenodo
  - DOI: https://doi.org/10.5281/zenodo.2648637
  - License: MIT License
  - Availabe since: 22nd of April 2019

There are no restrictions on the data.

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?**

All of the data is stored as csv files, which are non proprietary.
Furthermore metadata that instructs on how to run the code is included, as well as metadata that defines the software requirements. Lastly, there is also metadata on the software architecture and a description of it.

No vocabulary is used.

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 output data will have the following **license**:

- License: Creative Commons Attribution 4.0 International

The Jupyter Notebook, as well as all the documentation, input files and output are published with the following license:

- License: MIT License

There is no data embargo period.
Both the code and the data are **available since:** 22nd of April 2019
The data can be **reused** by third parties from the moment it is availbale, as defined by the licensing of it.
Quality assurance is possible, by checking agains the input data and testing if any rows were lost. Furthermore, since we are dealing with the population of Austria it is also possible to test if the sums still add up to the total population of Austria.
The data will remain reusable of an indefinit timeframe.

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**

- The main costs for making the data FAIR occur in the steps of making it interoperable, as here metadata,.. has to be created. This is a timeconsuming task that has to be covered. However, since I am a student nobody is paying for my time, therefore this part is also very affordable.
- The project is centered around data management from the beginning. The key contact for this is: David Wagner
- As long as the used services stay free there are very little direct costs for long term preservation. The main cost here also occurs due to the time that is required to keep documents and data up to date.
    - The potential value is that the future analysis is still possible.

4. Data security

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

Storage:
As the total size of all files is less than 1 Mb storage is not an issue.
Backup:
The data is stored locally for ease of use, but it is also saved on University Servers of TU Vienna.

- Backups are only made when the data changes, which does not happen often as we are dealing with altitude data of cities - both of with do not really change a lot over time. The only attribute that is prone to change is the number of inhabitants, but also this only happens every few years.

Recovery:
In case of fatal issues where the data is lost locally it will still be possible to access it from the original sources (data.gv.at and DBpedia).
Risks to data security:

- As we are not dealing with any personal data there is little risk of personal data being published.
- However, there is still the risk that data could be changed or deleted. However this is also of little concern since all the data that was used is publically available so it can easily be checked it anythin was changed.
    - Chrosschecking with the original data sources is also a good option to continously check the health of the data.

Access to the data:

- Access to the data (and the backup) will only be granted for the people that also conducted the experiment.

Safe transfer of data:

- Will be granted by using the VPN network of the Technical University of Vienna.

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

Ethical
No personal data is used for the experiment, so there are no real ethical issues that could arise.

- This also means that there is not sensible data that needs to be anonymised.

Legal
Data from data.gv.at:

- This is the main data source.
- Everything is under: Creative Commons Attribution License 3.0

Data from DBpedia:

- This is the second main data source where the altitude of cities is queried from.
- Everything is licensed under: Creative Commons Attribution-ShareAlike 3.0 License and the GNU Free Documentation License

Therefore there are not issues with reusing the raw data as well as the data that is produced by the experiment.

6. Other

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

Question not answered.