Plan Overview

A Data Management Plan created using DMPonline

Title: Correlating LA Museum Visitors to Crimes in LA

Creator: Hasan Alkhatib

Principal Investigator: Hasan Alkhatib

Data Manager: Hasan Alkhatib

Affiliation: Universitaet Wien - University of Vienna (Austria)

Funder: European Commission

Template: Horizon 2020 DMP

ORCID iD: 0000-0002-1223-9620

Project abstract:

This data management plan provides assistance and clarification of the process of testing the correlation between numbers of museums visitors city and numbers of crimes in Los Angeles.

ID: 39556

Last modified: 20-04-2019

Correlating LA Museum Visitors to Crimes in LA - 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**

In this experiment, I wanted to test the correlation between a specific behaviour and the numbers of crimes in the same geographical area. Based on that, I found two datasets from Los Angeles city in the US:

1. LA crime data from 2010 to present
2. LA museum visitors

The first dataset was about crimes data in LA from 2010 to 2019, and the other one was about the museums' visitors in the same city from 2014 to 2019. Both datasets were in the format of comma separated values "csv" and contains only numbers and plain text.

LA crimes dataset is owned by Los Angeles Police Department, and it had 26 data attributes, the most important ones were the date of occurrence, the victim information (e.g. gender, age, race) and the crime description. Adding to that, there were 1.9 Million incidents recorded in the dataset.

LA museums visitors dataset is owned by Los Angeles Open Data, and it had 12 columns, the first one was the month in which the reading was taken in, and the other 11 columns represented the numbers of visitors for 11 different museums.

About the output data from this experiment, it was a combination of calculated numbers grouped by the year of occurrence. So, as the datasets intersect in the period of 2014 to 2019, the output data was a CSV file with 1 column for the reading type and 5 other columns for the years results.

Datasource	2014	2015	2016	2017	2018
Museum_Visitors	#	#	#	#	#
Crimes	#	#	#	#	#

The output data may be useful for other researchers if they were interested in aggregated data about the crime rate in LA or even about the numbers of museums visitors.

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**

All datasets that were used or generated by the experiment are stored on Zenodo repository, which is an open-access repository. The data was cited by a DOI citation (Digital Object Identifier), and this is applied to each version will be uploaded on the repository.

Adding to the open-access repository, the data was reserved on Bitbucket in order to allow alternative version control using git. The reserved data link was added to the cited metadata under "replace" tag which refers to a possible replacement resource.

The data were described using the metadata obtained by DCMI (Dublin Core Metadata Initiative).

About the generated data, I used a convention in which I described the indicated reading (LA crimes, LA museums visitors) in the first column, followed by 5 columns for each year starting from 2014.

In order to enhance the findability of data, I added keywords regarding to the data science experiment in the metadata file.

In order to keep track of dataset, I used git-versioning next to the Zenodo preservation. On Bitbucket the versioning happens

according to git, but on Zenodo I get a new DOI for each version I upload.

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 datasets used in the experiment are openly available from the moment they were published on Zenodo, which indexed it in OpenAIRE.

About the data itself, it was stored in CSV file format (Comma Separated Values), which can be opened and manipulated using any text editor or sheets software like Excel. Based on that, there will be no need for any documentation to display and access data.

As the data was published on Zenodo, it makes the data itself associated with its metadata and documentation available, DOI cited and indexed on OpenAIRE and stored on CERN Data Center.

There are no restrictions actually because OpenAIRE provides open access to research outputs as a project financed by public funding in Europe.

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?**

Generated data can be easily interoperated by other users because of the used format "CSV". Adding to the functional format, the data was described according to a standard metadata schema called DCMI.

All data used and stored in the experiment are either plain text or counting numbers, which are easy to understand and to interoperate.

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**

Generated data from the experiment were licenced under Creative Commons Attribution 4.0 International License. With the CC license, data was availble for re-use and sharing by the time it was published and cited on Zenodo.

Both input and output data are reusable by third parties as they are valid CSV files with well-described metadata.

Beside publishing the data, it will be available and reusable as long as it still reserved and indexed on Zenodo and stored on CERN Data Center.

A method I used for assuring the experiment was to work with a small sample file from the input data and check results. Other than that the process was going smoothly as all data attributes are primitive types.

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**

Regarding costs, there were no expenses as the publication repository provides its services for free. Adding to that, the process of managing data was managed by the researcher himself "Hasan Alkhatib", and it started by analyzing input raw data, then specify a model for the generated data, then export data and publish it with DOI citation.
Even for the long-term preservation, it is stated on Zenodo that even for future data migration, the DOIs will stay working just as fine.

4. Data security

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

As the data is preserved on Zenodo, which stores data in CERN Data Center, there are guarantees that there will not be any loss of data, because even if Zenodo ended its services, data and its metadata will be moved to other data centre and with the existence of DOIs the data and accessing it won't be affected.

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

Input data and the generated data content were completely anonymized, so with that being preserved on Zenodo, there will be no unethical act.

6. Other

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

There are no obligations regarding procedures for the management process as there was no funder for this experiment.