Horizon Europe

Data Management Plan

Version 1.0
05 May 2021

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
1.0	05.05.2021	▪ Initial version
		▪

Action Number: 345353454
DOI: 10.5281/zenodo.6511888

Action Acronym: MAE
Action title: Machine learning algorithm evaluation

Date: 01/05/2022

DMP version: 1.0

1. Data Summary

1.1. Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

Two datasets have been reused as part of this project; the goal is to evaluated machine learning algorithms on these datasets, requiring the data to be in a state suitable for machine learning. To achieve this, pre-processing must conduct, to ensure re-use both the original and all subsequent version of these datasets must be preserved. This data will provide a common benchmark for algorithms implemented in the future.

1.2. What types and formats of data will the project generate or re-use?

The generated data are 4 datasets, two original and two created, all in a standardised csv (comma-separated values) format

1.3. What is the expected size of the data that you intend to generate or re-use?

The original data used constitutes a total of 56KB, with processed data containing a total of 94KB

1.4. What is the origin/provenance of the data, either generated or re-used?

Produced data sets:

dataset ID	name	type	format	estimated volume	contains sensitive data
P1	Credit_preprocessed	standard office documents	CSV	< 100 MB	no
P2	congressional_voting_preprocessed_neg	standard office documents	CSV	< 100 MB	no

Reused data sets:

dataset ID	name	type	format	estimated volume	contains sensitive data
R1	Credit	standard office documents	CSV	< 100 MB	no
R2	Congressional_voting	standard office documents	CSV	< 100 MB	no

Data used in this project was collected and hosted by UCI (…), an open repository which host both examples under the "Open Data Commons license". To provide an overview of the data generation/collection process the following breakdown is given:

- Congressional Voting: dataset includes votes for each of the U.S. House of Representatives Congress(wo)men on the 16 key votes identified by the CQA. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition). A modified version was provided by TU Wien and is used as the foundation of this experiment. The dataset in its original form can be found here: https://www.kaggle.com/c/184702-tu-ml-ws-21-congressional-voting/data

- Credit: This datasets concern's itself credit card applications. All attribute names and values have been changed to meaningless symbols to protect the confidentiality of the data, removing the need for special precautions when sharing. The version of this dataset was provided at by "openml" https://www.openml.org/d/29

1.5. To whom might your data be useful ('data utility'), outside your project?

This will provide a common benchmark for all users who wish to compare machine learning algorithms. Additionally, it can be useful for investigating the underlying information found in the induvial datasets.

2. FAIR data

Level of distribution	cc BY	This DMP is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). DOI: [10.5281/zenodo.6511888]

2.1.1. Making data findable, including provisions for metadata

2.1.1.1. Will data be identified by a persistent identifier?

The used of GitHub provides an inbuilt hashing and persistence facility, this provides every file with a unique idea that tracks all changes over all versions. Secondly, each dataset has been provided with a DOI and published on Zenodo.

2.1.1.2. Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Yes, the key words **Congressional**, **Credit** and **Voting** will be used enable increased searchability.

2.1.1.3. Will metadata be offered in such a way that it can be harvested and indexed?

Each original dataset is packages with a standardised machine-readable metadata file, additionally we will provide a README file with an explanation of all values and terms used [at file level/at dataset level/at project level]. This will help others to identify, discover and reuse our data. As the data is openly accessible, search engines (e.g., Google) can index all file within the repository. For generated datasets, the DublinCore standard will be utilized. It is a widely used metadata standard and so provides a good based for future work.

2.1.2. Making data accessible

2.1.2.1. Will the data be deposited in a trusted repository?

All data ("Credit" and "Congressional voting") with their associated code and dependencies are stored in GitHub. It offers free, open access and is widely used across industry. The original data was edited as part of experimental processes and may changes over-time. Therefore, all data, both now and in the future is available on original repository from which it was taken, and a copy stored within the repository. All changes to the data sets will be tracked through a hash check sum and ledger.

2.1.2.2. Will all data be made openly available?

Yes, all data will be openly accessible via the internet and is hosted under the "Creative Commons Attribution 4.0 International License", the repository freely and openly hosted by GitHub and can be found at the following link: https://github.com/theShamrockCoder/ML_evaluation

2.1.2.3. Will the data be accessible through a free and standardized access protocol?

Yes, all data is accessible through standard HTTP/s and API conventions, the repository can be interacted with through a standard web-browser.

2.1.2.4. How will the identity of the person accessing the data be ascertained?

A "README.md" file which outlines the owner and maintainer of the data and supporting code and dependencies, this will be placed in the repository.

2.1.3. Metadata

2.1.3.1. Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

Metadata is made openly accessible along with all code and data under the "Creative Commons Attribution 4.0 International License".

2.1.3.2. How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

The data will remain in the repository indefinitely, however, this may be subject to change.

2.1.3.3. Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

The repository contains a "README.md" file which outlines all code used in data generation. The tools used to achieve all tasks are open source and freely available online. An exhaustive list of software packages and versions can be found in a "requirements.txt" file, found within the repository.

2.2. Making data interoperable

2.2.1. What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

All data is sorted within the standardised CSV format, this is a widely used format for tabular data within the data and machine learning communities. A list of packages which support its accessibility can be found with a "requirments.txt".

2.2.2. Will your data include qualified references1 to other data (e.g. other data from your project, or datasets from previous research)?

Yes, all data which is re-use from external source will be referenced within the "experiment" document, this is located along with all data used in the above-mentioned repository.

2.3. Increase data re-use

All data is documented with an experiment guide and "README.md, along with all code, graphs and diagrams developed to produce the desired output. These can be found both in the above-mentioned repository and Zenodo. All output produced will be published under the "Creative Commons Attribution 4.0 International License", this allow the free manipulation, conversion, or distribution of all output with the only requirement being that outputs are attributed to the above Author. Data quality checks will be done, e.g., checks of consistency of labels, logical errors in the data, data curation, and version control. This was done at the start of the pre-processing stage, where the data was investigated from issues. All re-sources and intermediate data outputs (code, graphs, etc.) are made available for all datasets to ensure reproducibility. All changes are also tracked, and the history of the repository can be evaluated by all.

3. Other research outputs

As stated, all intermediate outputs, code and graphs will be maintained along with both the original and generated data. All artifacts can be found both on Zenodo and GitHub, with the complete changes history,

[1] *A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)*

packages and code made freely available for re-use.

4. Allocation of resources

No monetary considerations need to be accounted for this project, the publication, hosting and management of data, artefacts and code is free and will remain so. The storage time for the data is a present unlimited, therefore the data will remain free and accessible to all indefinitely.

Project Coordinator Principal Investigator	Warren Purcell, warren.purcell@tuwien.ac.at, ORCID iD: 0000-0002-1886-2632
Contact person (responsible for data management and DMP)	Warren Purcell, warren.purcell@tuwien.ac.at, ORCID iD: 0000-0002-1886-2632
DMP contributors	Warren Purcell, warren.purcell@tuwien.ac.at, ORCID iD: 0000-0002-1886-2632, Data Manager
Start date	28/03/2022
End date	1/05/2022
Funder, funding programme, grant number	N/a
Internal project number TU Wien	194.045

5. Data security

All datasets are relatively small and can be found in both the UCI data repository and GitHub, backup facilities are provided without intervention. Another copy of all artefacts, including this document can be found on Zenodo. As the data involved are public record and collected within a set period, no changes are expected to take place.

6. Ethics

All sensitive data has been removed and no special considerations are needed when distributing this above data. In the case of the “Credit” dataset anonymization was utilized before distribution. All data is available under the “Open Data Commons (ODC)” and so requires that all work is attributed to the original publisher. At this stage, it is not foreseen to process any personal data in the project. If this changes, advice will be sought from the data protection specialist at TU Wien (Verena Dolovai), and the DMP will be updated.

7. Other issues

N/a