Data Stewardship Experimental Use Case - Detailed DMP

1. Data summary

State the purpose of the data collection/generation

Data is collected to analyze the time series of energy balancing as well as traffic statistics. This is done for an experimental research project for a couse at the university.

Explain the relation to the objectives of the project

The collected data is necessary to get recommendations for actions, e.g. if there is constantly a high traffic value the operator asfinag had to think about expanding the traffic route.

Specify the types and formats of data generated/collected

The first dataset is provided as a zip file per year and consists of a csv file for each month. The second dataset from the asfinag is provided with the xlsx format.

Specify if existing data is being re-used (if any)

The data is provided as tabular data is provided by sensor measurements. The traffic volume is measured by infrared sensors and the energy balancing statistics of the import / export balance is calculated out of actual electricity flow and might be done by computers.

Specify the origin of the data

The data comes from the two named companies, the APG AG and the Asfinag AG. They use various sensors to collect the data.

State the expected size of the data (if known)

It is difficult to estimate the real raw size of the data, because a lot of preprocessing might have been already done by the companies, but at least we can tell the size of the provided datasets. The datasets of the traffic statistics are around half a megabyte per month, which means, we will have around 6MB per year for the provided datasets. Considering storage costs this might not be a problem now and in future. The dataset with the energy balancing statistics are a bit less big, because they have been preprocessed by APG and they already provide a well suited human-readable format. We have around the same size, 6MB, per year.

Outline the data utility: to whom will it be useful

The data are provided by both companies by their own, which means all rights are by APG AG or by Asfinag AG. Depositing them in a data repository might fit well and in case of type, size, complexity and sensitivity of the data there might not be any problem. They also might have to decide which copyright they place on them.

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

The datasets can be as already mentioned found on the public available company website (asfinag.at and apg.at). Metadata are not provided and this is indeed a very problematic point.

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?

Both datasets do not make use of persistent and unique identifiers.

Outline naming conventions used

The datasets do follow their own systematic naming covention mainly consiting of the date and in some cases of the name. Making the datasets available in a data repository might need revision work.

Outline the approach towards search keyword

A useful approach towards search keywords would be pretty simple for both datasets. It is pretty clear in which area both of them are taking place.

Outline the approach for clear versioning

Unfortunately both datasets do not provide an approach for clear versioning. As it seems, they are just overwritten, what makes traceability unfeasible.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

Thestandard

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

The aggregated and summarized data will be made open. This means, that the raw data is preprocessed before, which makes sense, because aggregating is necessary the create some useful output.

Specify how the data will be made available

The data can be pushed to various data repositories. It might be useful to use data repos in relation to the energy or traffic sector. It would be also possible to create a own repository because especially the energy sector has a lot of statistics.

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)?

For downloading and accessing the data a browser or http compatible software is needed. After downloading the data had to be unzipped, which means there is also some archiving software necessary. After that any excel or csv compatible program can be used.

Specify where the data and associated metadata, documentation and code are deposited

The data and the associated metadata are provided on the homepage of the companies. Documention and code might be published at github.

Specify how access will be provided in case there are any restrictions

If the rights are available the data will be published taking account to the owner in a repository.

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

For this datasets they follow a few standards. Datasets are provided as csv-utf8 encoded, which is per se a standard in data formats and easy readable by many

programs. The dataset with the traffic statistics had some weak points regarding the machine readability of the date format, but this can be avoided by further pre-processing.

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?

?

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

Question not answered.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

The summarized data will be made available for re-use. Therefore, no 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

The data will be provided in a data repository and available for further working on the data.

Describe data quality assurance processes

The data will be collected with sensors in both cases. There are processes to keep the losses low and the data quality at a high level. Sensors will be installed to monitor the data collection and as well humans will be check the data collection monthly.

Specify the length of time for which the data will remain re-usable

The plan for data preservation is about 10 years to the past for both datasets. This means, the compability and the naming conventions remain the same to ensure a high data quality level.

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

Costs for the data collection and making the data FAIR are either full costs of sensors equipment and the data cleaning afterwards. These costs will be covered by the research and strategy division of the company.

Clearly identify responsibilities for data management in your project

The main responsibility for data management in this project will have the data manager who is responsible for the data flow in the company. The data collector and the data analyzer will report to him and deliver useful data. For archiving and sharing reasons there will be a role, which is responsible for the preservation and legal publication of the data.

Describe costs and potential value of long term preservation

The data manager has to decide how long what data will be kept. Speaking of the public available datasets the time frame is set to 10 years.

4. Data security

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

The data will be saved in a datacenter hosted in Austria. This ensures legal compliance and these datacenters will also ensure a secure storage due to a redundant storage system.

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

Speaking of ethical aspects both data sources are not handling sensitive or personal data. For both there are no concers of different privacy problems.

6. Other

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

-