

PGSheets Template (Name of your dataset)

INSTRUCTIONS ON PGSHEETS

PGSheets is a set of questions to be used as guidelines for documenting Property Graph datasets. PGSheets is inspired by 5W1H and the 5 Why's used in journalistic writing and each section of the document answer one of the questions What? Where? When? Why? and How? about the data and the answer to Who? is embedded in each one of the sections. Your PGSheet should be complete, clear and concise and should contain all needed information about the property graph for a reader/data-user to understand the possible pros and cons of using your data.

Each one of the sections, has a set of question which is labelled with **MUST**, **SHOULD** or **NICE**, which means (1) **MUST** - questions that must always be answered and included in a PGSheet, (2) **SHOULD** - questions that should be answered but are not mandatory to have, (3) **NICE** - questions that would be nice to have an answer in a PGSheet but can be considered optional.

EXECUTIVE SUMMARY **MUST**

Dataset Name	name of the dataset
Number of Node Labels	-
Number of Edge Labels	-
Total number of nodes	-
Total number of edges	-
Version	-
Source	-
Dataset authors	-
How to cite	-
PGSheets authors	Your name

COMPOSITION (WHAT?)

Context/Domain of the data (MUST) - What is the domain of the data and its context. For example, if the dataset is part of the biomedical domain, financial data, temporal information and what's the general context of the data.

Information about the instances (MUST) - Schema information about the property graph (entities, labels, properties and its type). In here the schema of the graph should be exposed in two different ways: (1) graphical way as a figure which contains nodes and edge labels, the attributes of each label and its data type and the connections every two nodes can have and, (2) in a tabular format which contains the same information but in a tabular format for ease of understanding. The answer should also include who made the schema and who wrote down the information in the PGSheets.

Important additional information (MUST) - This section should contain important additional information to the schema of the graph. This additional information includes known distribution/range of values of considered important attributes, constraints and semantic information. These should be described in this document in a natural language format or in the case of values distribution, using plots and exposing it as a figure. In case information about constraints are available in a specific language, it should be added as a separate file in the repository and not included in the PG-Sheets.

Who detected or wrote about each one of the additional information should also be indicated in this section. Why such constraints/information is important and in which context needs to be included in this answer as well.

Does it rely on external sources? (MUST) - Was the dataset generated according to an external source? or does the composition of the graph rely on a third-party system. By who and why this choice was made. Essentially in this question, we want to know if the dataset is generated for example according to an external source or tool that might change the content of the data)

Dataset splits or other samples available in the repository that contains this same data or the other way around (MUST) - Point to other splits or samples of the same data in the repository, declare if the schema is the same or not. Who made the split and why this decision was made. More details about the collection aspects should be declared in the data collection section.

How is it different from other versions? (MUST) If the answer to the previous question is yes, then in this answer should be declared what is different from its split.

MOTIVATION (WHY?)

Why was the dataset created? (MUST) - What is the motivation behind the creation of this dataset. Does this dataset have a specific characteristics that previous datasets for the same task did not cover.

Which task was it intended to solve? (MUST) - What kind of problem was it intended to solve, e.g., frequent subgraph mining, path query evaluation, entity linking, etc. Why this type of data/domain was collected to solve this task.

Has this data been used before? (MUST) (If yes, give examples or links to where it was used - if it is a sample of a previous data, specify it in the data collection section) Similar to the question in composition section but from the perspective of why this data was sampled or split into another dataset.

If it has been used or it is a sample, why each one of the entities were chosen? (MUST)

Reasoning behind the modeling of the graph (MUST) - Why the dataset/graph was modelled in the way that it was exposed in the Composition section of the paper.

DATA COLLECTION AND PRE-PROCESSING (WHEN? HOW?)

Source of data (MUST) - If it is collected from sensors or other appliances, this question should describe the hardware specification. IF the data was collected from an website, it should include an link an description of the website.

Time-frame that the data was collected (MUST) Dates in which the collection and processing of the data was made. **How the dataset was collected? (MUST)** This question includes which method was used to collect the dataset and what kind of pre-processing was applied in this data including who was responsible for each part of the data collection and processing phase and why each method was applied to get to the final dataset.

Is the dataset a sample? (MUST) include information about the full dataset, similar to the previous information.

Who is involved in each one of the collection steps? (MUST) can be answered together with the collection steps questions.

Is the raw data available? (SHOULD) (if there is any) - is the raw data of this dataset (without any post collection processing) available for access, if yes, include the link to the webpage.

Method used for collection available? (SHOULD) (external software for example) - IF the collection method was done by an software, link the software in this page.

DISTRIBUTION AND MAINTENANCE (WHERE?)

License of the dataset (MUST)	-
Date of Distribution (MUST)	-
Dates of Modification (MUST)	Maintenance dates and how did it change in each version
Responsible for hosting the dataset (MUST)	Institution or name of the person responsible
Responsible for maintaining the dataset (MUST)	Contact of who is maintaining the dataset

Papers that use this dataset? (NICE) → Link to any publicly available information, for example papers and references that use the same data.

How can a third-party contribute to this dataset?(NICE) - Instructions on how to contribute to the dataset and who is the person in contact in this case.

Additional Comments (NICE) - Additional comments regarding license or distribution of this dataset (e.g. cannot be distributed on other platforms without the consent of the authors)

USING THE DATASET AND LEGAL/ETHICAL ASPECTS (How?)

Was this dataset approved by an ethical review? (MUST) - If yes, link to the ethical review it was approved from to a page which contains information about the ethical review it was submitted to.

Privacy issues (MUST) (if this was part of the data preprocessing - explain what were the privacy related risks that the raw data contained) - are there any attributes or data that has to be anonymized? and if yes then make it clear also in the preprocessing step. If not, then answer why privacy is not a issue in this dataset.

Ethical issues (MUST) → does it advantage or disadvantage a specific group? If this is unknown, give an overview of the known groups in this data so users know which are the potential groups that they should look into

Does it contain inappropriate information? Or any potentially inappropriate data? (MUST) - If it is not known, declare it in this question.

Does it comply with a data protection regulation law? (MUST) e.g. GDPR, or similar from other country?

How to use the dataset? (MUST) (Instructions on loading the dataset/external software that can be used and etc)