PG-Sheets Template (Name of your dataset)

Instructions on PG-Sheets

PG-Sheets is a document composed of questions that can be used as guidelines for documenting Property Graph datasets. The PG-Sheets sections and questions are inspired by the 5W1H (Who, What, Where, When, Why, and How) writing technique, in which each section should answer one of the questions What? Where? When? Why? and How? about the data, and the answer to Who? is embedded in each section. The overarching hypothesis behind PG-Sheets is that if a dataset's documentation successfully addresses all these questions, then it can be considered complete and concise.
Each PG-Sheet is intended to provide all the essential information a reader or potential data user needs to understand the dataset, including its strengths, limitations, and context of use. Each one of the sections has a set of questions which is labeled according to the following priority level:

- MUST - questions that must always be answered and included in a PG-Sheet.
- SHOULD - questions that should be answered, but are not mandatory to have.
- NICE - questions that would be nice to have an answer but can be considered optional.

This template presents each question along with recommendations or expected answer formats. These should be treated as guidelines and adapted to the context of the specific property graph dataset being documented.

Executive Summary (MUST)

This section can be considered an abstract of your documentation; it contains the basic information about your dataset.

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	-
PG-Sheets authors	-

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 the general context of the data.
Information about the instances (MUST) - Schema information about the property graph (entities, labels, properties, and its type). Here the schema of the graph should be exposed in two different ways: (1) graphical way as a figure that contains nodes and edge labels, the attributes of each label and its data type, and the connections every two nodes can have (see example Figure 1), 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 the information in the PG-Sheet.

Label	Attribute/Property	Description
Node or edge label (e.g. Person)	Property key (e.g. name)	Natural language description of what this property refers to

Table 1: Table with description of the properties

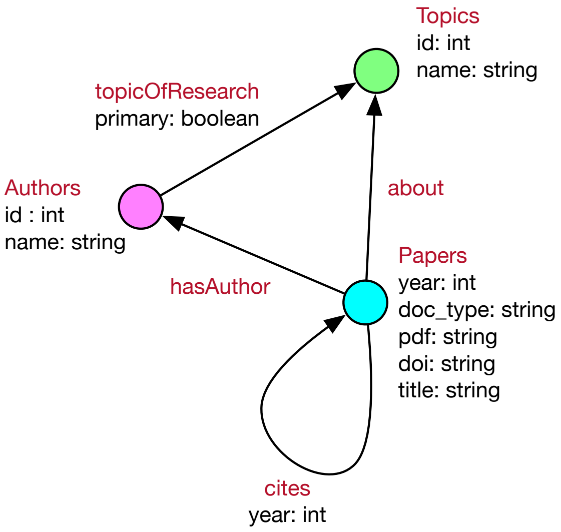

Figure 1: Example of a property graph schema figure

Node Label	Size
Node (e.g. Authors)	Number of nodes labelled Authors

Edge Label	Size
Edge (e.g. about)	Number of edges labelled about

Important additional information (MUST) - This section should contain additional important information about the graph schema. This additional information includes a 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 value distribution, using plots and exposing it as a figure. In case information about constraints is available in a specific language (e.g. PG-Keys), it should be added as a separate file in the repository and referenced in PG-Sheets, however, there must always be a natural language explanation for each constraint referenced in the PG-Sheets document.
Who detected or wrote about each 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 property graph generated according to an external source? Or does the composition of the graph rely on a third-party system? 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 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 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 was this type of data/domain 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 previous data, specify it in the data collection section Similar to the question in the composition section but from the perspective of why this data was sampled or split into another dataset.
If it has been used or is a sample, why was each of the entities chosen? (MUST)

Reasoning behind the modeling of the graph (MUST) - Why the dataset/graph was modeled 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 a website, it should include a link and description of the website.
Time-frame that the data was collected (MUST) Dates in which the collection and processing of the data were 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 to 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. Here we recommend using figures to explain the different steps and methods applied. The figure should be clear enough to understand the many steps used in the property graph dataset construction process.
Is the dataset a sample? (MUST) If yes, 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 processing steps post data collection) 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 a software, link the software on 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 the 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 that 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 have to be anonymized? And if yes then make it clear also in the preprocessing step. If not, then answer why privacy is not an 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 another country?
How to use the dataset? (MUST) (Instructions on loading the dataset/external software that can be used, etc)