

---

# WikiDBs: Dataset Documentation

---

Liane Vogel<sup>1</sup>, Jan-Micha Bodensohn<sup>2,1</sup>, Carsten Binnig<sup>1,2</sup>

<sup>1</sup>Technical University of Darmstadt, Germany

<sup>2</sup>German Research Center for Artificial Intelligence (DFKI), Darmstadt, Germany

1 We use the *datasheet for datasets* dataset documentation framework proposed by [1]. The questions  
2 are taken from version v8 of the paper.

## 3 1 Motivation

4 **For what purpose was the dataset created? Was there a specific task in mind? Was there a**  
5 **specific gap that needed to be filled? Please provide a description.** WikiDBs is created to foster  
6 the development of foundation models for relational databases. Currently, there is a lack in large-scale  
7 collections of relational databases, most existing datasets contain only individual tables that are not  
8 connected by foreign key relationships.

9 **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g.,**  
10 **company, institution, organization)?** The dataset is created by researchers from the Systems Group  
11 from the Computer Science Department at Technical University of Darmstadt.

12 **Who funded the creation of the dataset?** This work has been supported by the BMBF and the state  
13 of Hesse as part of the NHR Program, as well as the HMWK cluster project 3AI. We want to thank  
14 hessian.AI, and DFKI Darmstadt for their support.

15 **Any other comments?** N/A

## 16 2 Composition

17 **What do the instances that comprise the dataset represent (e.g., documents, photos, people,**  
18 **countries)?** Each instance of the dataset represents one relational database, consisting of the tables  
19 of the database in CSV format, a schema description in JSON format and a visualization in form of  
20 an ERD-Diagram as a PDF and .DOT file.

21 **How many instances are there in total (of each type, if appropriate)?** In total, WikiDBs 20k has  
22 20,000 instances. The full WikiDBs dataset will contain 100,000 instances.

23 **Does the dataset contain all possible instances or is it a sample (not necessarily random) of**  
24 **instances from a larger set?** It would be possible to create even more than 100,000 databases  
25 from Wikidata using our method. As our method is based on property relationships in Wikidata, we  
26 uniformly traversed all suitable (i.e. resulting in tables with more than 20 rows) relationships until  
27 20,000 instances were reached. We make our code openly available, so that databases tailored to  
28 specific use-cases can be created if necessary.

29 **What data does each instance consist of?** Each instance consists of:

- 30
- tables: a folder with CSV files, one CSV file per table in the database

- 31 • tables\_with\_item\_ids: a folder with CSV files, one CSV file per table in the database, each  
32 cell value is a tuple containing the Wikidata Q-ID in addition to the cell value (cell value,  
33 Q-ID)
- 34 • schema.json: a JSON file containing information on every table's column names and  
35 datatypes, as well as foreign key connections to other tables in the database
- 36 • schema\_diagram.pdf and schema\_diagram.dot: an ERD diagram visualizing the database  
37 tables and foreign key connections and the underlying dot file
- 38 • renaming.json: information on the inputs and outputs of our paraphrasing step (please refer  
39 to our paper on how we paraphrased table and column names using GPT-4)

40 **Is there a label or target associated with each instance?** No, there are no designated labels or  
41 targets for each database.

42 **Is any information missing from individual instances?** For reviewing purposes due to the high  
43 costs of paraphrasing, we paraphrased the table names and column names only of 17,000 of the  
44 20,000 databases in order to save costs before incorporating reviewer feedback. The 3,000 databases  
45 that have not been paraphrased yet do not include the "renaming.json" file. We plan to paraphrase the  
46 whole corpus of 100,000 databases for the final version of this paper and dataset.

47 **Are relationships between individual instances made explicit (e.g., users' movie ratings, social  
48 network links)?** There are no relationships between different databases in the dataset. Relationships  
49 between individual tables per database are made explicit in each schema.json file describing the  
50 schema of each database.

51 **Are there recommended data splits (e.g., training, development/validation, testing)?** Yes, we  
52 provide the training/validation/testing split that we used for our experiments along with the dataset  
53 for download. The splits were created by splitting the 20,000 files into 14,000 for training, 2,000 for  
54 validation, and 4,000 for testing. We kept the 3,000 databases that were not paraphrased yet in the  
55 training set and otherwise splitted the rest randomly for the three splits.

56 **Are there any errors, sources of noise, or redundancies in the dataset?** As our corpus is grounded  
57 in Wikidata, the underlying data may potentially be noisy, untruthful, and hard to attribute to  
58 individual authors. Since our automatic paraphrasing procedure is based on large language models, it  
59 is vulnerable to their well-known weaknesses, including hallucinations and social biases.

60 **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g.,  
61 websites, tweets, other datasets)?** WikiDBs is self-contained. We make the Q-IDs of items from  
62 Wikidata available, to open up the opportunity to adapt our corpus for a variety of table-based end  
63 tasks such as schema matching, entity matching, and deduplication.

64 **Does the dataset contain data that might be considered confidential (e.g., data that is pro-  
65 tected by legal privilege or by doctor-patient confidentiality, data that includes the content of  
66 individuals' non-public communications)?** All included data is already openly available in the  
67 Wikidata knowledge base. Wikidata is actively monitored and moderated, adhering to strict guide-  
68 lines and policies regarding personal information ([https://www.wikidata.org/wiki/Wikidata:  
69 Living\\_people](https://www.wikidata.org/wiki/Wikidata:Living_people)). Our used paraphrasing method is unlikely to add additional information beyond  
70 the provided inputs, therefore the publication of our dataset does not expose any new personally  
71 identifiable information.

72 **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening,  
73 or might otherwise cause anxiety?** Our dataset is grounded in Wikidata, which is actively monitored  
74 and moderated. However there might still be content that does not fulfill the Wikidata guidelines and  
75 might be harmful.

76 **Does the dataset relate to people? If not, you may skip the remaining questions in this section.**  
77 The dataset contains data on individuals that is already openly available in Wikidata. Wikidata  
78 adheres to strict guidelines and policies regarding personal information ([https://www.wikidata.](https://www.wikidata.org/wiki/Wikidata:Living_people)

79 `org/wiki/Wikidata:Living_people`). We do not add any additional data related to people to the  
80 dataset apart from the data from Wikidata.

### 81 **3 Collection Process**

82 **How was the data associated with each instance acquired?** All data from the dataset is openly  
83 available in Wikidata. We use the inherent triple structure of Wikidata (subject, predicate, object) to  
84 build relational tables and foreign key connections between them.

85 **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or  
86 sensors, manual human curation, software programs, software APIs)?** The dataset is based  
87 on the Wikidata dump 'latest-all.json.gz' of May 15, 2024, downloaded from <https://dumps.wikimedia.org/wikidatawiki/entities>. To process the data, it was loaded into a MongoDB  
88 database, the database creation is written in Python. For rephrasing, the OpenAI API was used to  
89 prompt GPT-4o (*gpt-4o-2024-05-13*).  
90

91 **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic,  
92 probabilistic with specific sampling probabilities)?** It would be possible to create even more than  
93 100,000 databases from Wikidata using our method. As our method is based on property relationships  
94 in Wikidata, we uniformly traversed all suitable (i.e. resulting in tables with more than 20 rows)  
95 relationships until 20,000 instances were reached. We make our code openly available, so that  
96 databases tailored to specific use-cases can be created if necessary.

97 **Who was involved in the data collection process (e.g., students, crowdworkers, contractors)  
98 and how were they compensated (e.g., how much were crowdworkers paid)?** Only the authors  
99 were involved in the data collection process.

100 **Over what timeframe was the data collected?** The dataset is based on the Wikidata dump 'latest-  
101 all.json.gz' of May 15, 2024.

102 **Were any ethical review processes conducted (e.g., by an institutional review board)?** No.

103 **Does the dataset relate to people? If not, you may skip the remainder of the questions in  
104 this section.** The dataset contains data on individuals that is already openly available in Wikidata.  
105 Wikidata adheres to strict guidelines and policies regarding personal information ([https://www.wikidata.org/wiki/Wikidata:Living\\_people](https://www.wikidata.org/wiki/Wikidata:Living_people)). As Wikidata is made available under Creative  
106 Commons Public Domain License, no individuals needed to be notified about the data collection.  
107

### 108 **4 Preprocessing/cleaning/labeling**

109 **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing,  
110 tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing  
111 of missing values)?** Yes, the data from the Wikidata dump 'latest-all.json.gz' of May 15, 2024 was  
112 reformatted into a format suitable to process for the approach of creating relational databases and  
113 loaded into a MongoDB instance.

114 **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support  
115 unanticipated future uses)?** Yes, we plan to provide the pre-processed data as a MongoDB export  
116 along with our code to do the pre-processing.

117 **Is the software that was used to preprocess/clean/label the data available?** Yes, we openly  
118 release our code to preprocess the data.

119 **Any other comments?** N/A

## 120 5 Uses

121 **Has the dataset been used for any tasks already?** Initial experiments have been conducted for the  
122 tasks of predicting missing table names, column names, and cell values using the dataset within the  
123 paper of this submission. The dataset has not yet been used for further tasks.

124 **Is there a repository that links to any or all papers or systems that use the dataset?** The dataset  
125 has not been used yet apart from this submission. It is planned to link to papers or systems that use  
126 the dataset on a Website of the dataset.

127 **What (other) tasks could the dataset be used for?** The dataset can be used to train a foundation  
128 model on relational databases. The included Q-IDs which link every cell value to the corresponding  
129 Wikidata item open up the opportunity to adapt our corpus for a variety of table-based end tasks such  
130 as schema matching, entity matching, and deduplication.

131 **Is there anything about the composition of the dataset or the way it was collected and prepro-  
132 cessed/cleaned/labeled that might impact future uses?** As our corpus is grounded in Wikidata,  
133 the underlying data may potentially be noisy, untruthful, and hard to attribute to individual authors.  
134 Since our automatic paraphrasing procedure is based on large language models, it is vulnerable to  
135 their well-known weaknesses, including hallucinations and social biases.

136 **Are there tasks for which the dataset should not be used?** The dataset should not be used for any  
137 tasks that result in unfair treatment of individuals or groups.

138 **Any other comments?**

## 139 6 Distribution

140 **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution,  
141 organization) on behalf of which the dataset was created?** The dataset will be openly available  
142 under CC-BY 4.0 license.

143 **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** The dataset  
144 will be uploaded to Zenodo where it receives a DOI. Additionally we plan to upload it to the  
145 HuggingfaceDataset Hub.

146 **When will the dataset be distributed?** The dataset will be published when the submitted paper gets  
147 published to incorporate reviewer feedback into the final version.

148 **Will the dataset be distributed under a copyright or other intellectual property (IP) license,  
149 and/or under applicable terms of use (ToU)?** The dataset will be openly available under CC-BY  
150 4.0 license.

151 **Have any third parties imposed IP-based or other restrictions on the data associated with the  
152 instances?** No.

153 **Do any export controls or other regulatory restrictions apply to the dataset or to individual  
154 instances?** No.

155 **Any other comments?** N/A

## 156 7 Maintenance

157 **Who will be supporting/hosting/maintaining the dataset?** The dataset will be hosted on Zenodo  
158 and the HuggingfaceDatasets Hub.

159 **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** Com-  
160 ments and Questions can be sent to Liane Vogel (liane.vogel@cs.tu-darmstadt.de).

161 **Is there an erratum?** We plan to use Zenodo, and Zenodo allows to upload further versions if error  
162 corrections are necessary along with the documentation of what was changed.

163 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**  
164 If necessary, the authors will fix errors, update the dataset, and upload updated versions to Zenodo.

165 **If the dataset relates to people, are there applicable limits on the retention of the data associated**  
166 **with the instances (e.g., were the individuals in question told that their data would be retained**  
167 **for a fixed period of time and then deleted)?** N/A

168 **Will older versions of the dataset continue to be supported/hosted/maintained?** Zenodo hosts all  
169 versions of the dataset that are ever uploaded.

170 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for**  
171 **them to do so?** Yes, we make our source code openly available, to make extentions/augmentations  
172 etc. and customized versions of the dataset possible.

173 **Any other comments?** N/A

## 174 **References**

175 [1 ] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021).  
176 Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92.