

Wikary: A Dataset of N-ary Wikipedia Tables Matched to Qualified Wikidata Statements

Igor Mazurek¹, Berend Wiewel¹ and Benno Kruit^{1,*}

¹Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

Abstract

We introduce a dataset of almost 32.000 tables from 3 Wikipedia language versions which have been matched to Wikidata statements with qualifiers at 98.4% precision. The tables express a diverse set of n-ary relations which constitute a new target for semantic table interpretation research.

Keywords

Tabular Data, Knowledge Graph Matching, N-ary Relations, Qualifiers

1. Introduction and Background

Tabular data from databases, documents, or the web contain a wealth of information that could be made more accessible, searchable, and useful by matching it with Knowledge Graphs (KGs). However, integrating tabular data with KGs is still typically a manual process, requiring in-depth knowledge of the KG schema and the domain of interest. Much progress has been made in recent years, and automating this task currently remains an open challenge. In particular, many systems have been developed that match table columns to semantic types (Column-Type Annotation, CTA), table cells to KG entity (Cell-Entity Annotation, CEA), and pairs of table columns to binary KG relations (Column-Property Annotation, CPA). One application of such systems is the extraction of subject-property-object triples from each pair of columns, for extending KGs with new factual statements. The quality of such systems may be evaluated using a variety of benchmark datasets, with the goal of assessing performance on a variety of topical domains and controlled environments [1, 2, 3]. In order to create useful systems that effectively match tabular data to KGs in practice, these benchmarks should therefore reflect the diversity of tabular data as it occurs in real-world usage. Public benchmarks for this problem are becoming increasingly realistic, incentivizing the development and evaluation of usable systems [2].

However, much real-world tabular data does not express binary KG relations but rather represents higher-order, n-ary relations instead [4, 5]. N-ary relations express statements involving multiple (> 2) entities or values, which cannot be decomposed into independent, atomic parts without compromising truthfulness, coherence, or completeness. For example, consider the statement the following statement:


SemTab'22: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, Collocated with the 21st International Semantic Web Conference, 23-27 October 2022, Hybrid, Hangzhou, China

*Corresponding author.

✉ i.w.mazurek@student.vu.nl (I. Mazurek); b.wiewel@student.vu.nl (B. Wiewel); b.b.kruit@vu.nl (B. Kruit)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Example 1.1 (N-ary Statement)

The album *Thriller* by *Michael Jackson* reached the *top-1* position in the *US Billboard 200* chart on *February 26th, 1983*.

This statement cannot be expressed by a single triple, but also cannot be decomposed into independent parts that make sense on their own. Multiple statements like this can be naturally represented in tabular form, and, indeed, real-world tables very often express n-ary relations. As of yet, though, table-to-KG matching benchmarks have not included such n-ary tables as they are not covered by the CPA task framework. To reflect the diversity of real-world data, this class of tables should be considered in table interpretation research.

In the popular Wikidata KG[6], n-ary relations are modeled using qualifiers[7]. Qualifiers extend simple statements with additional context information for the claim and may be represented in RDF in a straightforward way using blank nodes. This way, complex n-ary claims may be represented such as our example 1.1 above:

Example 1.2 (N-ary Statement as RDF in Turtle syntax)

```
wd:Q44320 p:P2291 [           # Thriller (album)
  ps:P2291 wd:Q188819;       # charted in: US Billboard 200
  pq:P585 "1983"^^xsd:gYear; # point in time: 1983
  pq:P1352 "1"              # ranking: 1
] .
```

Qualifiers in Wikidata are most often used to represent temporal scopes of statements and are therefore important from a data modeling perspective.

Because tables in Wikipedia articles express information about well-known entities that are also described by Wikidata, they form a prime candidate for studying n-ary tables in a controlled environment. By only changing the structure of the table, while keeping a tight alignment to a broad-coverage KG, we hope to contribute insights that may generalize to situations in which the entities might not be covered by a KG. Such low-coverage scenarios may occur for tables from other sources such as the web[5], CSVs [8], or relational databases [9].

Contribution Our goal is to encourage the study of the entire variety of web tables encountered in practice while maintaining a grounding in well-studied semantic models. In this paper, we, therefore, introduce a dataset of almost 32.000 tables from 3 Wikipedia language versions which have been matched to Wikidata statements with qualifiers. The large scale allows for the analysis of the diversity of representation of n-ary statements in practice, and by sourcing tables in multiple languages, we aim to diversify the topical coverage of the tables. The dataset is publicly available on Zenodo¹.

2. Dataset creation

The creation process of the dataset can be split into three parts: scraping tables, joining the tables with Wikidata, and filtering high-confidence matches. Finally, we estimate the quality of the dataset using an annotation interface to label a subset of the data manually.

¹<https://doi.org/10.5281/zenodo.7025005>

Tim Allen

Filmography

Year	Title	Role
1988	<i>Tropical Snow</i>	Baggage Handler
	<i>Comedy's Dirtiest Dozen</i>	
1989	<i>Rodney Dangerfield: Opening Night at Rodney's Place</i>	Himself
1990	<i>Tim Allen: Men Are Pigs</i>	
1991	<i>Tim Allen Rewires America</i>	
1994	<i>The Santa Clause</i>	Scott Calvin / Santa Claus
1995	<i>Toy Story</i>	Buzz Lightyear
	<i>Meet Wally Sparks</i>	Himself
1997	<i>Jungle 2 Jungle</i>	Michael Cromwell
	<i>For Richer or Poorer</i>	Brad Sexton

(a)

Toy Story (Q171048)

Statements

voice actor	<div> <div>Q171048</div> <div>Tom Hanks</div> <div>character role</div> <div>Woody</div> </div>
	<div> <div>0 references</div> </div>
voice actor	<div> <div>Q171048</div> <div>Tim Allen</div> <div>character role</div> <div>Buzz Lightyear</div> </div>
	<div> <div>0 references</div> </div>

(b)

Figure 1: (a) Example of a row from a Wikipedia table that expresses an n-ary relation (b) Wikidata statement that expresses the same information

Lang	Database version	Titles	Tables	Pages w/ tables
Simple EN	all nopic 2022-03	274296	35047	18194
NL	all nopic 2021-11	2853121	582319	243893
PL	all nopic 2022-05	2034836	537366	199791

Table 1

Statistical characteristics of each Wikipedia version

Scraping data The Wikimedia Foundation provides two different types of static database dumps of its projects, including Wikipedia. One is the original *wikitext* markup as edited by contributors, and the other is a static HTML export which is suitable for self-hosting². Due to the fact that the wikitext allows for complex nesting of templates, we opted to use the HTML representation to extract the final tables as they appear rendered in articles. This ensures that all table elements that are viewable by readers are also available for extraction.

We decided to compare three languages - simple English, Dutch, and Polish. The choice was made based on our linguistic competencies as later on, we need to annotate tables which requires an understanding of the given language. Key statistical characteristics of each Wikipedia version together with dump versions are shown in Table 1. For each Wikipedia version considered, we scraped all HTML tables of the “wikitable” class, which is used to indicate content tables in articles. During this step, we also extracted additional data for each table: page title, table index, section title, table caption, and list of hyperlinks per row. The page title and table index allow us to uniquely identify the table.

Merge with Wikidata statements Each row in every table is being assessed individually, which is why for the sake of clarity, in our example we will cover only one row in the table. We will focus on the row shown in Figure 1a which is present on Tim Allen’s page in the *Filmography* section. The next step is converting hyperlinks of each row to Wikidata entities,

²Provided using the Kiwix/Openzim toolchain (<http://www.kiwix.org>)

as well as looking up the Wikidata entity associated with the page itself. The page entity provides additional information about the table, necessary to completely understand the table. Matching hyperlinks to Wikidata identifiers is performed using an index based on all redirects and mappings of article titles to Wikidata IDs.

The next step is creating permutations of all pairs of entities included in each row and the page entity. These are merged into a collection of all Wikidata statements that have qualifiers. This merge can be considered as database-style join, and the keys that we use are pair of identifiers from the Wikipedia table and subject and object from Wikidata. In Figure 1b, a Wikidata statement is shown that shows the same information as the row in the example table. The set of Wikidata statements can be seen as a database table that consists of more than 15.5 million tuples, with the following information: subject, property, object, qualifier property, and qualifier value. From the perspective of record matching, this merge operation may be seen as a blocking step that efficiently produces a large number of matching candidates.

Finding matches The last step of the implementation is finding matches to keep only table rows that likely express n-ary relations. We can distinguish three types of matches, however, all of them are based on qualifier value in a corresponding Wikidata statement. The qualifier value can not be equal to the subject and object and has been in a different cell. Note that the presence of a match does not guarantee that the table expresses the same information as the matched Wikidata statement, nor does it guarantee that the table expresses n-ary relations. Our goal is to find a large number of tables that may express n-ary facts with high likelihood, and these matching approaches are designed to result in high precision. We use the following matching functions:

1. **Wikidata identifier match** In this type of match, the qualifier value is equal to Wikidata entities. We look up the Wikidata entities of hyperlinked pages and compare the qualifier value, if the values are the same then we have a match.
2. **Year cell match** This match applies when the qualifier value is provided as a date, however instead of using the entire date we only extract the year from it. The next step is looking up a cell from the row that contains only a given year as a text in the cell and nothing more. Moreover, this cell can not be a subject or object of a matched statement.
3. **Within cell year match** In this more lenient version of the above-mentioned matching, we also used the year extracted from the qualifier value. However in this case we try to find this year occurring in any place in a cell. This means that all of the matches from the previous type are included in this one. This trade-off leads to higher recall as the year in a cell could be combined with some additional information, which would not be detected by the previous type of matching. Though it also means lower precision because the number indicating the year could be part of some large number of unrelated strings, our evaluation has shown that the loss in precision is limited.

Quality Evaluation In order to estimate the quality of the extracted tables, we created the environment to annotate tables. To facilitate this process, annotators need convenient tools to save time on unnecessary actions like opening a table's HTML one at a time or independently

N-ary	Not N-ary	Incorrect formatting	Skip
prev	next		
Page title	Division_of_Bendigo	tableIndex: 0 rowIndex: 6	Visit wikipedia page
Subject	Q152666	Billy Hughes	Billy Hughes Visit wikidata page
Property	P102	member of political party	Visit wikidata page
Object	Q1516976	Nationalist	Nationalist Party of Australia Visit wikidata page

Member	Party	Results
	Protectionist	1901–1906
Sir John Quick	Independent Protectionist	1906–1909
	Commonwealth Liberal	1909–1913
John Arthur	Labor	1913–1914
Alfred Hampson	Labor	1915–1917
Billy Hughes	Nationalist	1917–1922
Geoffry Hurry	Nationalist	1922–1929
Richard Keane	Labor	1929–1931

Figure 2: Widget used for annotations. The interface shows the full HTML table with all original markup and the matching row highlighted. Above it, the matching Wikidata statement is shown with links to check the identity and background information of matching entities, along with the option to inspect the table in its original context.

finding the table on Wikipedia. Our annotation interface makes use of the PigeonXT³ Python library and runs entirely within a Jupyter Notebook, and is therefore integrated with our extraction approach. As shown in Figure 2 it displays the content of the table along with subject, property, and object per random row and relation pair merged with Wikidata statement. This labeling interface will be released along with the dataset, to facilitate the creation and assessment of new versions of this dataset.

Moreover, during the evaluation process, we discarded rows containing more than 100 hyperlinks per row as the manual review showed that were results of incorrect table formatting.

3. Statistics & Analysis

The Table 2 displays key statistics during the process of finding matches. The third column of the table presents the number of tables where at least one hyperlink was matched to the Wikidata page. Looking at percentages of the total data set this number seems to be quite consistent across all languages. Most of the tables that were successfully merged with Wikidata statements are in Simple English Wikipedia, followed by Polish, and Dutch has the lowest number of 15%. Intriguing is the number of tables per page differs remarkably, 1.23 in Simple English compared to 1.85 in the Polish version.

The number of matches is shown in Table ??, while Simple English and Dutch Wikipedia look similar in terms of *identifier matches*, 7%, and 5.5%, the Polish version has only 2.1%. *Year cell*

³<https://github.com/dennisbakhuis/pigeonXT>

Lang	Dataset		At least one hyperlink matched to Wikidata page			Merged with Wikidata statements with qualifiers		
	Tables	Pages	Tables	Pages	T. % of dataset	Tables	Pages	T. % of dataset
Simple EN	35047	18194	23809	14935	68%	10023	8155	28.6%
NL	582319	243893	360778	153628	62%	87584	51867	15%
PL	537366	199791	366795	149858	68.3%	114694	62196	21.3%

Table 2
Number of tables and pages at distinct point of implementation

Lang / Match type		Simple EN	NL	PL
Wikidata identifier match	Pages	605	4094	2015
	Tables	706	4790	2420
	Rows	2212	14230	6370
	Matches	6895	37382	13978
Year cell match	Pages	53	904	1329
	Tables	58	1055	1438
	Rows	176	3208	7930
	Matches	270	4987	25318
Within cell year match	Pages	217	14445	4773
	Tables	292	7485	10459
	Rows	740	41646	28798
	Matches	1109	50551	58504

Table 3
Number of matches found

match is alike in Dutch and Polish versions, while Simple English has only half of the matches percentage-wise, specifically 0.6%. Results of *within cell year match* show different results in all of the Wikipedias, Dutch has the highest percentage value equal to 16%, Polish has 9.1% and Simple English only 2.9%.

Due to the low number of annotations, we decided to calculate evaluation metrics based on merged results for both different methods of detection and Wikipedia's versions, this resulted in 750 annotated matches. The table was classified as n-ary if one of three types of matches occurred, this results in a precision of 98.4%. Starting from all possible matches returned by the blocking step, the matching step filtered n-ary tables with a recall of 23.3% .

4. Maintenance, Availability and Use-cases

The dataset is available on Zenodo⁴. Our aim is to maintain and expand the dataset in more languages, including the full English Wikipedia. Additionally, we aim to incorporate any user feedback with regard to possible noise and formatting errors, to improve usability. We

⁴<https://doi.org/10.5281/zenodo.7025005>

Lang	Simple EN		NL		PL	
Total n-ary tables	964		18674		12355	
N-ary table characteristics	Rows	Columns	Rows	Columns	Rows	Columns
mean	31.59	9.83	19.16	5.59	27.70	18.40
std	50.02	99.34	43.64	3.34	65.48	74.39
25%	6	3	8	4	8	4
50%	16	4	13	4	13	5
75%	37.25	6	22	6	25	8
max	626	2511	4574	49	1009	701

Table 4
Numer of n-ary tables found

published the full dataset creation pipeline as a set of Python scripts⁵, to facilitate the creation of custom datasets based on our approach, including those based on up-to-date Wikipedia dumps. Additionally, we also included the annotation tool used for quality evaluation.

Use-cases The main use-case of our dataset is as a resource for N-ary information extraction. It can be used to train or benchmark N-ary table interpretation systems within the Wikipedia domain, or for analyzing general structures of n-ary tables (such as functional dependencies[5]) for out-of-domain applications.

As an exemplar task, we have already used this dataset in preliminary research on distinguishing binary and nary tables by statistical classification. We extracted surface-level features (such as column types) from known binary and these n-ary tables, and were able to train classifiers that distinguished them with acceptable performance. Such classifiers may also be used as a pre-processing filter when extracting binary facts from tables, so as to reduce false-positive CPA predictions.

5. Conclusion

We introduce a dataset of almost 32.000 tables from 3 Wikipedia language versions which have been matched to Wikidata statements with qualifiers at 98.4% precision. The tables express a diverse set of n-ary relations which constitute a new target for semantic table interpretation research.

References

- [1] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, V. Cutrona, Results of semtab 2020, in: CEUR Workshop Proceedings, volume 2775, 2020, pp. 1–8.

⁵<https://github.com/igormazurek/wikary/>

- [2] V. Cutrona, J. Chen, V. Efthymiou, O. Hassanzadeh, E. Jiménez-Ruiz, J. Sequeda, K. Srinivas, N. Abdelmageed, M. Hulsebos, D. Oliveira, et al., Results of semtab 2021, Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching 3103 (2022) 1–12.
- [3] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, Semtab 2019: Resources to benchmark tabular data to knowledge graph matching systems, in: European Semantic Web Conference, Springer, 2020, pp. 514–530.
- [4] B. Kruit, P. Boncz, J. Urbani, Extracting n-ary facts from wikipedia table clusters, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 655–664.
- [5] O. Lehmborg, C. Bizer, Profiling the semantics of n-ary web table data, in: Proceedings of the International Workshop on Semantic Big Data, 2019, pp. 1–6.
- [6] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledge base, Communications of the ACM 57 (2014) 78–85.
- [7] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, D. Vrandečić, Introducing wikidata to the linked data web, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 8796 (2014) 50–65. doi:10.1007/978-3-319-11964-9.
- [8] M. Hulsebos, Ç. Demiralp, P. Groth, Gittables: A large-scale corpus of relational tables, arXiv preprint arXiv:2106.07258 (2021). URL: <https://arxiv.org/abs/2106.07258>.
- [9] T. Döhmen, M. Hulsebos, C. Beecks, S. Schelter, Gitschemas: A dataset for automating relational data preparation tasks, in: 2022 IEEE 38th International Conference on Data Engineering Workshops (ICDEW), IEEE, 2022, pp. 74–78.