Wikary

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

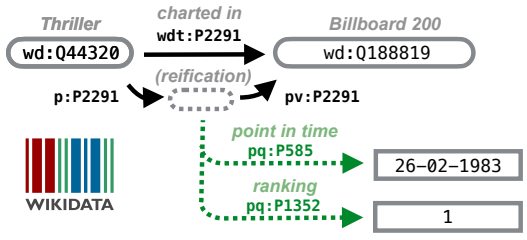
Igor Mazurek, Berend Wiewel and **Benno Kruit**

VU VRIJE UNIVERSITEIT AMSTERDAM

N-ary statements describe relations between multiple values

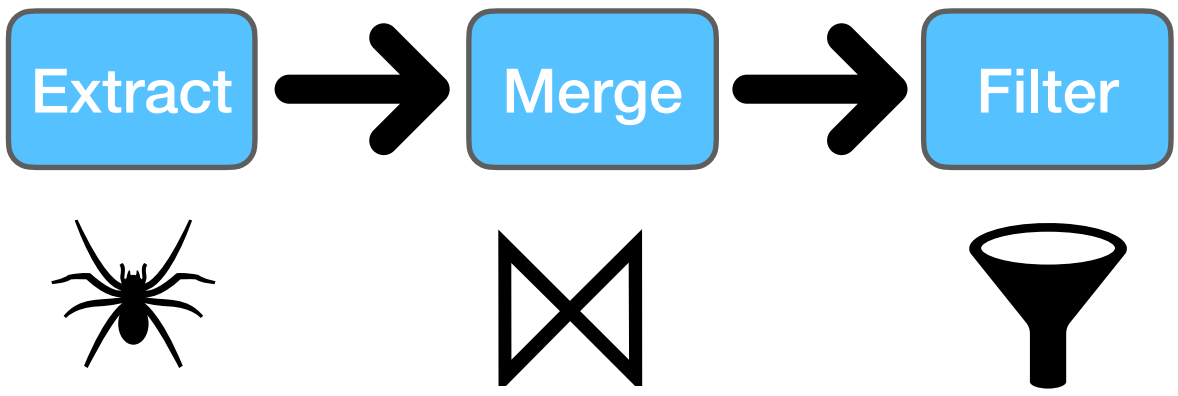
"The album *Thriller* by Michael Jackson reached the top-1 position in the US Billboard 200 chart on February 26th, 1983."

In Wikidata, they are modeled using qualifiers.



Tables often express n-ary statements.

This work: Dataset Creation



Languages

1. Simple English
2. Polish
3. Dutch

Scaling

1. Efficient ID lookup
2. ID-pair extraction
3. Partitioned join

Match qualifiers

1. Wikidata ID
2. Year cell
3. Within-cell year

Evaluation

- Annotation interface
- Filtering step:
  - 98.4% precision
  - 23.3% recall

Data statistics (final dataset contains ~32.000 tables total)

Merge step	Full Dataset		At least one hyperlink matched to Wikidata			Merged with Wikidata statements with qualifiers		
Lang.	Tables	Pages	Tables	Pages	T. % of datase	Tables	Pages	T. % of dataset
Simple EN	35.047	18.194	23.809	14.935	68%	10.023	8.155	**28.6%**
NL	**582.319**	243.893	360.778	153.628	62%	87.584	51.867	15%
PL	537.366	199.791	366.795	149.858	**68.3%**	114.694	62.196	21.3%

← EN: highest % merged

← NL: most tables

← PL: highest % linked

Filter step	Full Dataset		Wikidata ID match			Year cell match			Within-cell year match		
Lang.	Tables	Pages	Tables	Pages	T. % of data	Tables	Pages	T. % of data	Tables	Pages	T. % of data
Simple EN	10.023	8.155	706	605	**7%**	58	53	0.6%	301	218	3%
NL	87.584	51.867	4.790	4.094	5.5%	1.055	904	1.2%	14.490	7.508	**16%**
PL	114.694	62.196	2.420	2.015	2.1%	1.438	1.329	1.3%	10.532	4.804	**9.2%**

EN: most tables from ID match

NL/PL: most tables from within-year match

Download

zenodo.org/record/7025005

Future Work

- More languages (full English)
- Extend to non-matching tables