

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

Igor Mazurek, Berend Wiewel and Benno Kruit

**N-ary statements** describe relations between multiple values

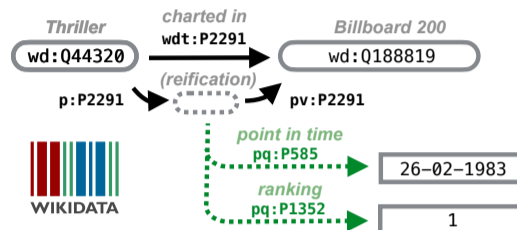


“The album *Thriller* by Michael Jackson reached the top-1 position in the US Billboard 200 chart on February 26th, 1983.”

Thriller (Q44320)

1982 studio album by Michael Jackson

charted in	Billboard 200	
	point in time	26 February 1983
	ranking	1



Thriller (album)

From Wikipedia, the free encyclopedia

Charts

Weekly charts

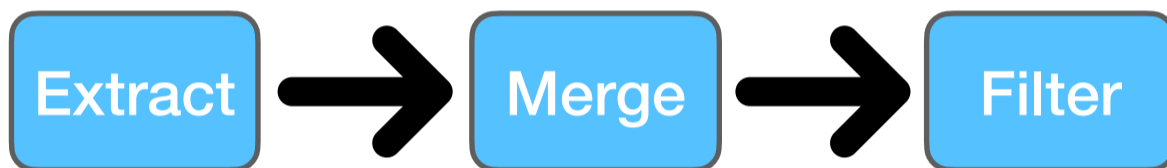
Chart (1983–2010)	Peak position
Australian Albums (Kent Music Report) <sup>(nl)</sup>	1
Canada: Top Albums/CDs (1998– <sup>(nl)</sup>	1
French Albums (SNEP) <sup>(fr)</sup>	50
Swiss Albums (Schweizer Hitparade) <sup>(fr)</sup>	4
UK Albums (OCC) <sup>(en)</sup>	1
US Albums (Billboard) <sup>(en)</sup>	1
US Top 100 (Top 100 Albums & Soundtracks) <sup>(en)</sup>	1



In Wikidata, they are modeled using qualifiers.

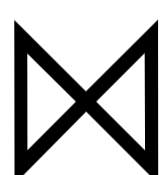
Tables often express n-ary statements.

### This work: Dataset Creation



#### Languages

- Simple English
- Polish
- Dutch



#### Scaling

- Efficient ID lookup
- ID-pair extraction
- Partitioned join



#### Match qualifiers

- Wikidata ID
- Year cell
- Within-cell year

### Evaluation

N-ary	Not N-ary	Incorrect / omitted
117	104	
Prose title	Division_of_Bendigo	tableIndex 0 rowIndex 0
Subject	Q157555	Billy Hughes
Property	P102	member of political party
Object	Q1510970	Nationalist Nationalist Party of Australia

Member	Party	Years
St John Quick	Nationalist	1901–1906
St John Quick	Independence Party	1906–1909
St John Quick	Commonwealth Liberal	1909–1913
John Archer	Labour	1913–1914
Alfred Hampeel	Labour	1915–1917
Billy Hughes	Nationalist	1917–1920
Geoffrey Hurry	Nationalist	1920–1924
Richard Heene	Labour	1924–1931

- 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 dataset	Tables	Pages		T. % of dataset
Simple EN		35.047	18.194	23.809	14.935	68%	10.023	8.155	<b>28.6%</b>	← EN: highest % merged
NL		<b>582.319</b>	243.893	360.778	153.628	62%	87.584	51.867	15%	← NL: most tables
PL		537.366	199.791	366.795	149.858	<b>68.3%</b>	114.694	62.196	21.3%	← 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	<b>7%</b>	58	53	0.6%	301	218	3%	← EN: most tables from ID match
NL		87.584	51.867	4.790	4.094	5.5%	1.055	904	1.2%	14.490	7.508	<b>16%</b>	← NL/PL: most tables from within-year match
PL		114.694	62.196	2.420	2.015	2.1%	1.438	1.329	1.3%	10.532	4.804	<b>9.2%</b>	

EN: most tables from ID match  
 NL/PL: most tables from within-year match

#### Future Work

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

**Download**  
[zenodo.org/record/7025005](https://zenodo.org/record/7025005)

