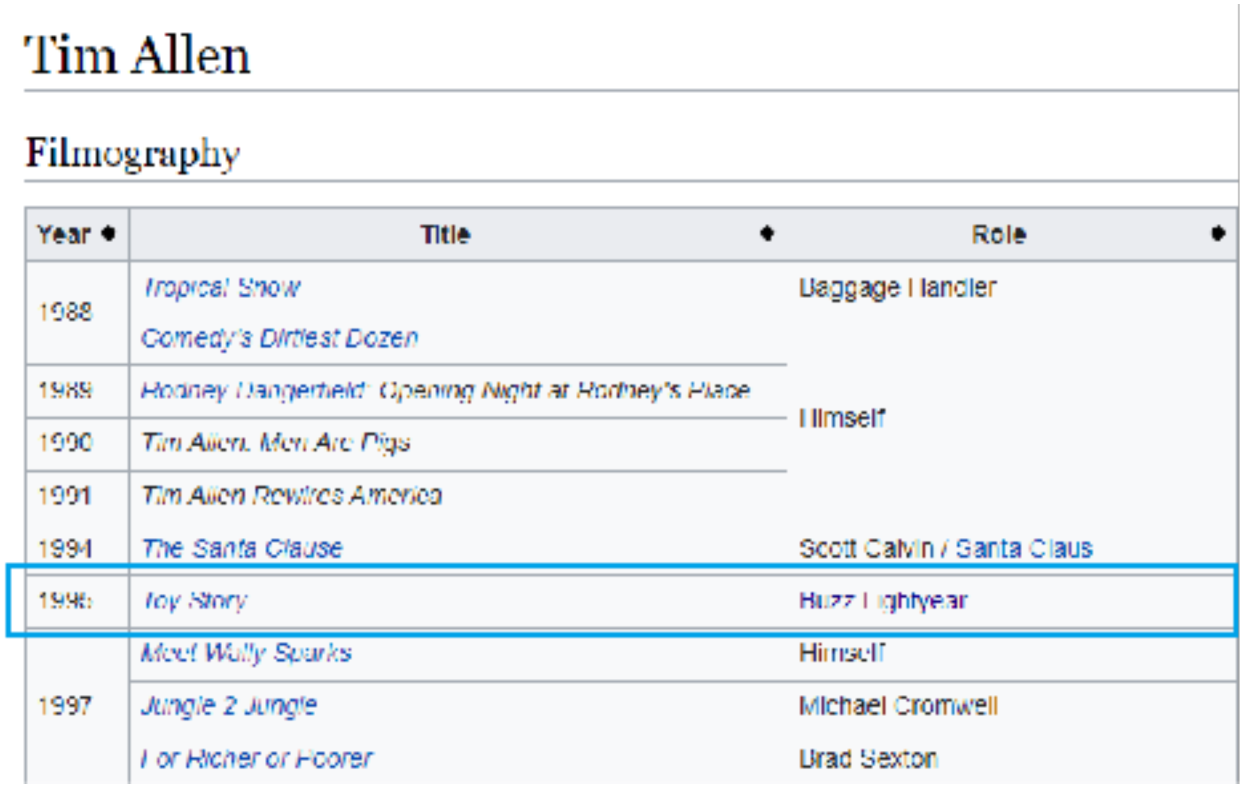
Wikary

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

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

Wikary

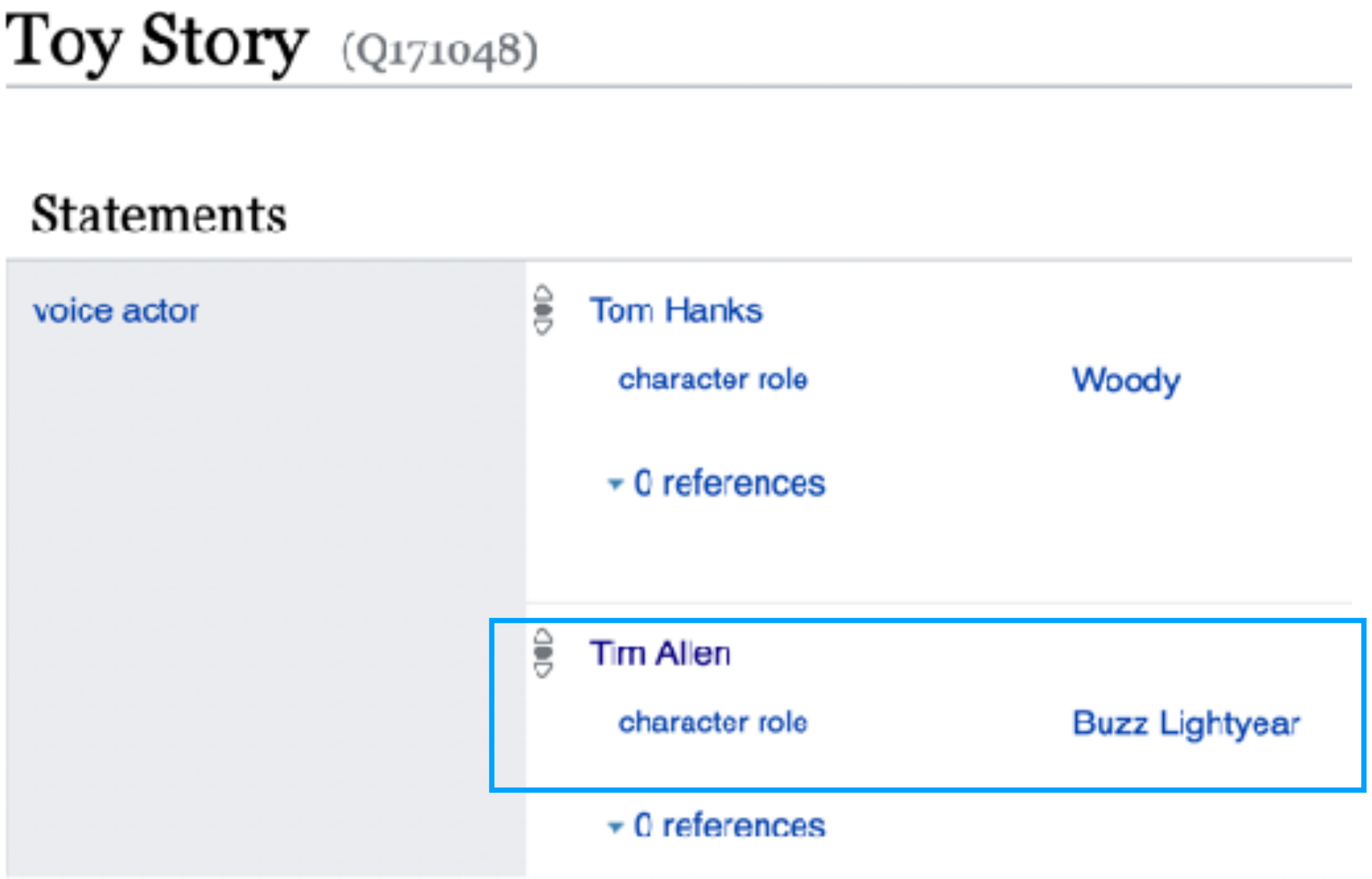
N-ary **Wikipedia** tables matched to **Wikidata**



Tim Allen

Filmography

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



Toy Story (Q171048)

Statements

voice actor

Tom Hanks
character role Woody
▾ 0 references

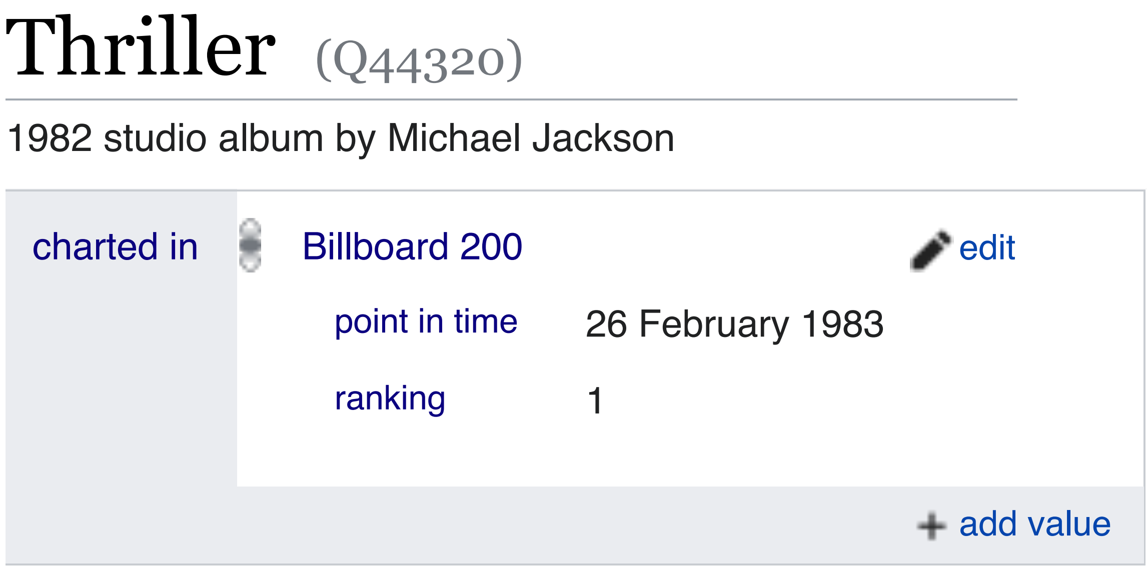
Tim Allen
character role Buzz Lightyear
▾ 0 references

N-ary statements

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

N-ary statements

with Wikidata qualifiers



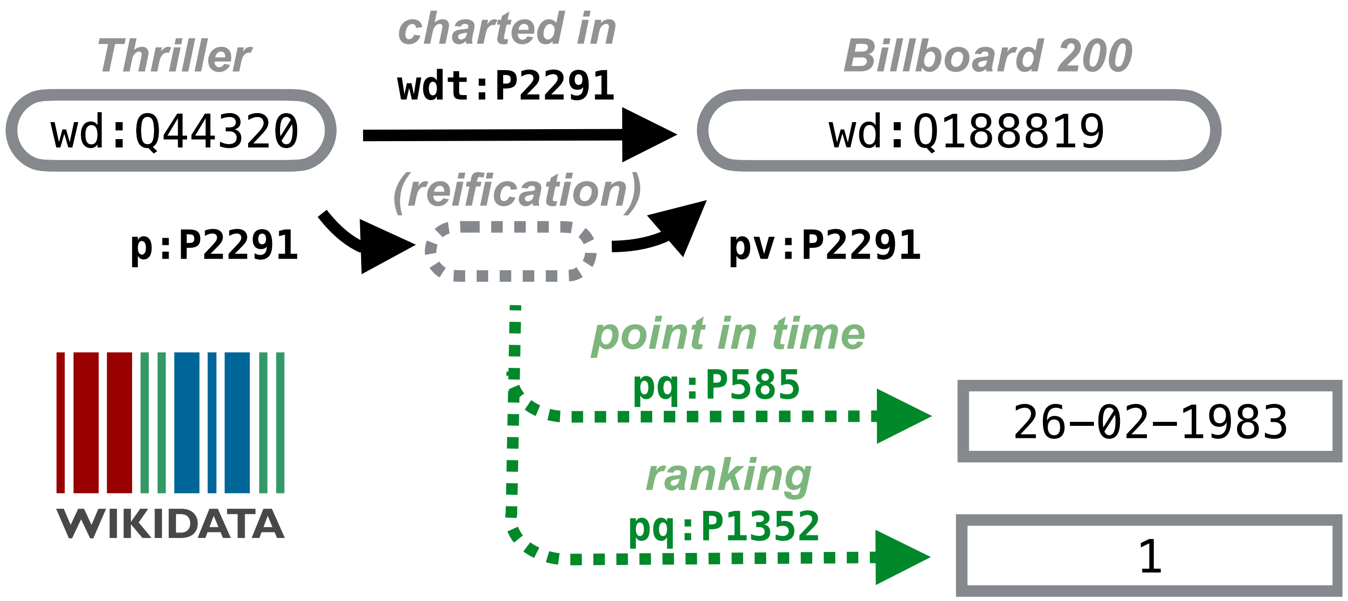
Thriller (Q44320)

1982 studio album by Michael Jackson

charted in — Billboard 200 — edit

point in time — 26 February 1983

ranking — 1



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

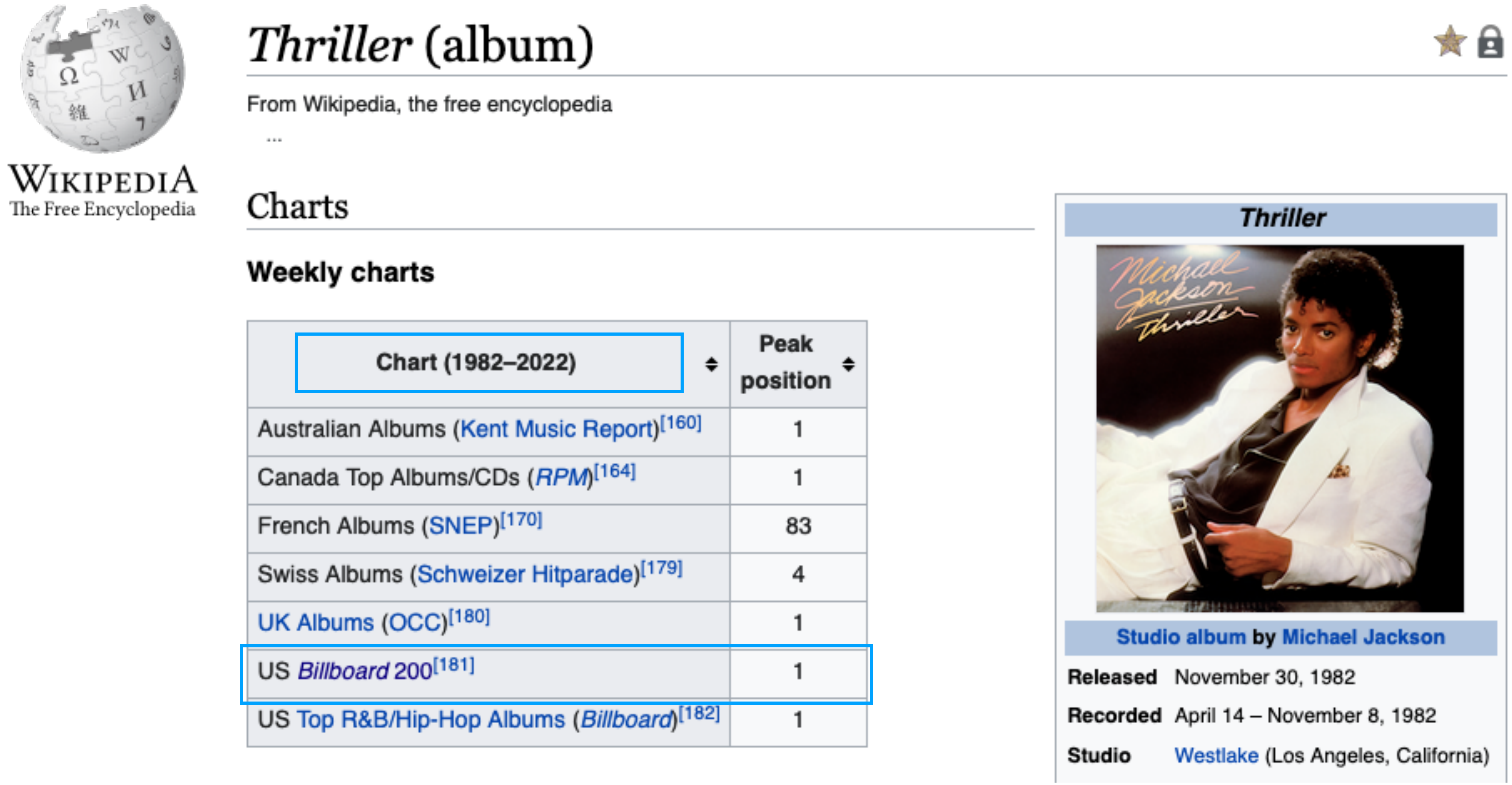
N-ary statements

reified as RDF blank nodes

```
# Thriller (album)
wd:Q44320 p:P2291 [

  # charted in: US Billboard 200
  ps:P2291 wd:Q188819;

  # point in time: 26th February 1983
  pq:P585 ”1983-02-26”^^xsd:date;

  # ranking: 1
  pq:P1352 ”1”
].
```

N-ary statements

in Wikipedia tables



WIKIPEDIA
The Free Encyclopedia

Thriller (album)

From Wikipedia, the free encyclopedia

...

Charts

Weekly charts

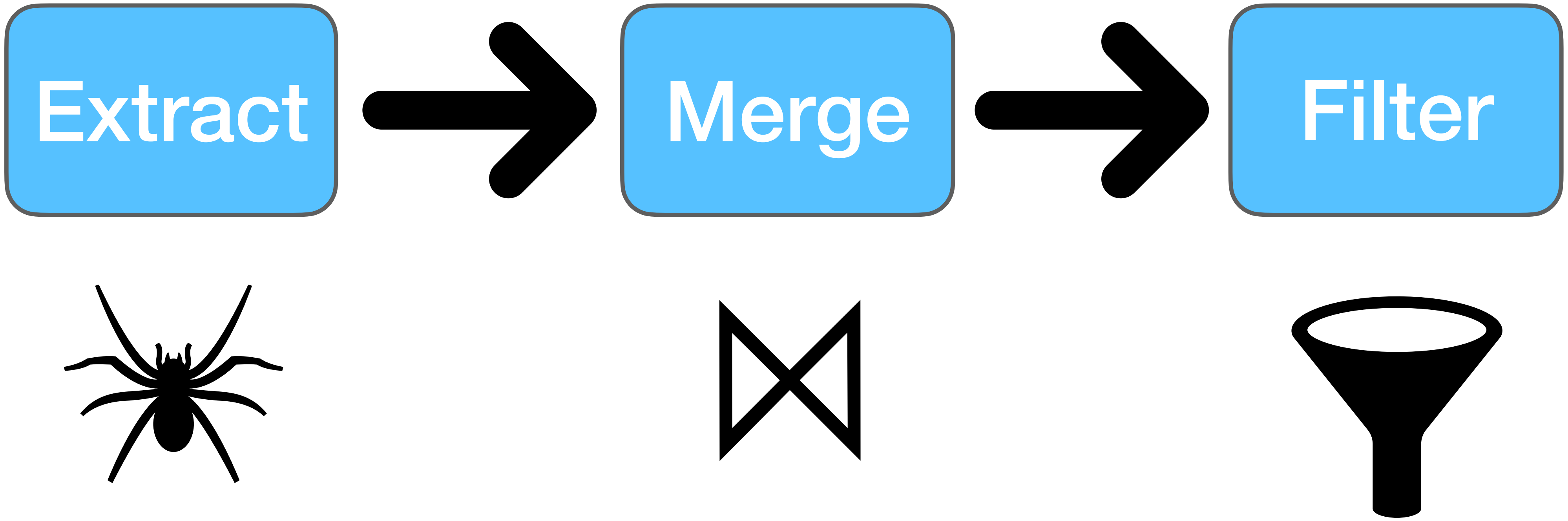
Chart (1982–2022)	Peak position
Australian Albums (Kent Music Report)[160]	1
Canada Top Albums/CDs (*RPM*)[164]	1
French Albums (SNEP)[170]	83
Swiss Albums (Schweizer Hitparade)[179]	4
UK Albums (OCC)[180]	1
US *Billboard* 200[181]	1
US Top R&B/Hip-Hop Albums (*Billboard*)[182]	1

Thriller

Studio album by Michael Jackson

Released	November 30, 1982
Recorded	April 14 – November 8, 1982
Studio	Westlake (Los Angeles, California)

Wikary dataset creation



Extract → Merge → Filter

Languages

1. Simple English
2. Polish
3. Dutch

Scaling

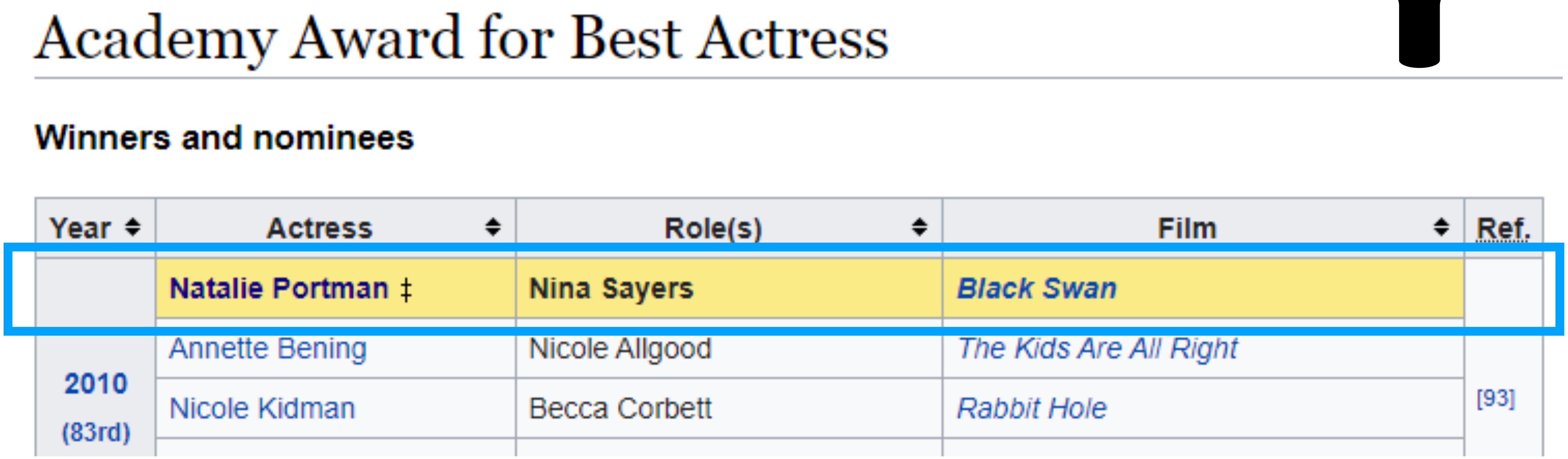
1. Hyperlink lookup
2. ID-pair extraction
3. Partitioned join

Qualifier match

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

Wikidata ID match

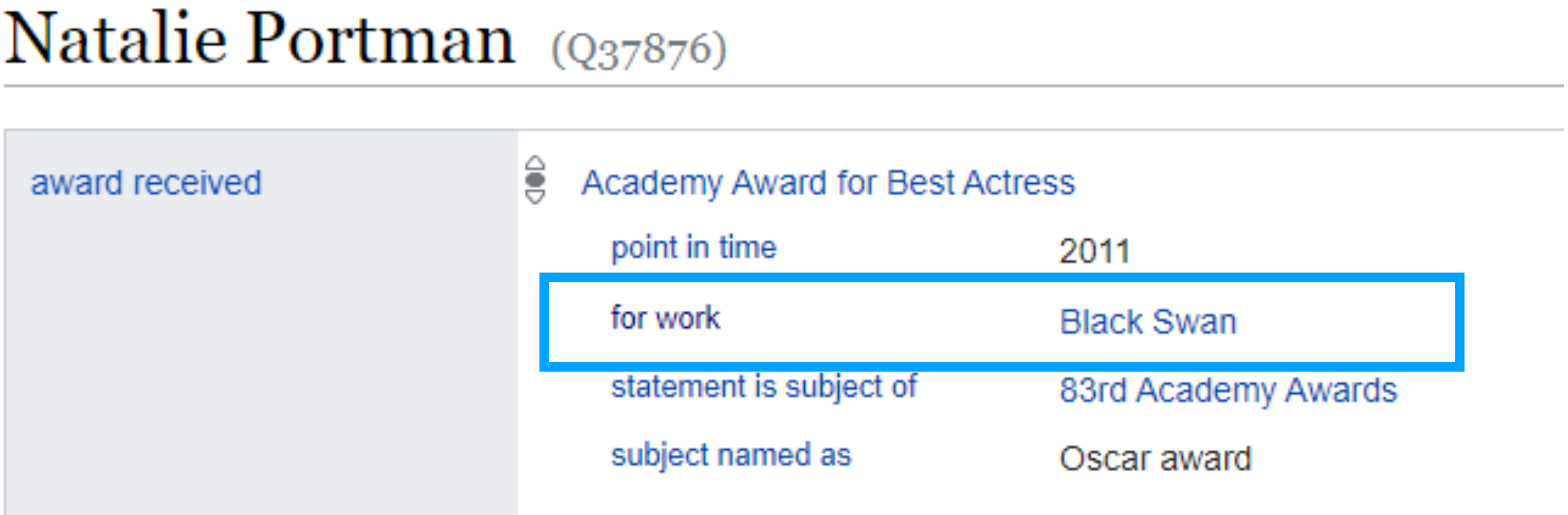
Qualifier filter



Academy Award for Best Actress

Winners and nominees

Year	Actress	Role(s)	Film	Ref.
2010 (83rd)	**Natalie Portman ‡**	**Nina Sayers**	***Black Swan***	[93]
	Annette Bening	Nicole Allgood	*The Kids Are All Right*	
	Nicole Kidman	Becca Corbett	*Rabbit Hole*	

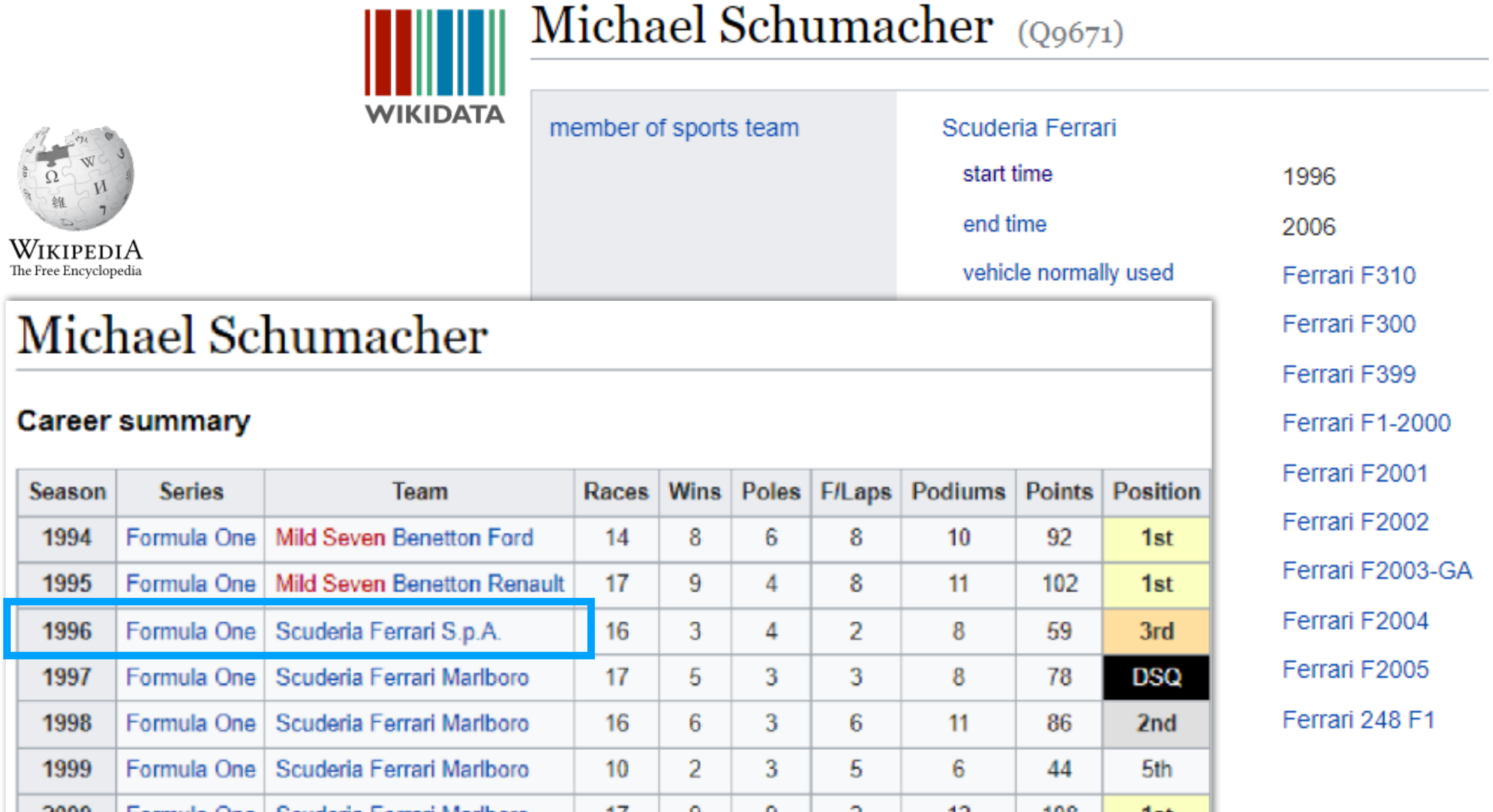


Natalie Portman (Q37876)

award received	Academy Award for Best Actress	
	point in time	2011
	for work	Black Swan
	statement is subject of	83rd Academy Awards
	subject named as	Oscar award

Year cell match

Qualifier filter



Michael Schumacher (Q9671)

member of sports team

Scuderia Ferrari

start time	1996
end time	2006
vehicle normally used	Ferrari F310 Ferrari F300 Ferrari F399 Ferrari F1-2000 Ferrari F2001 Ferrari F2002 Ferrari F2003-GA Ferrari F2004 Ferrari F2005 Ferrari 248 F1

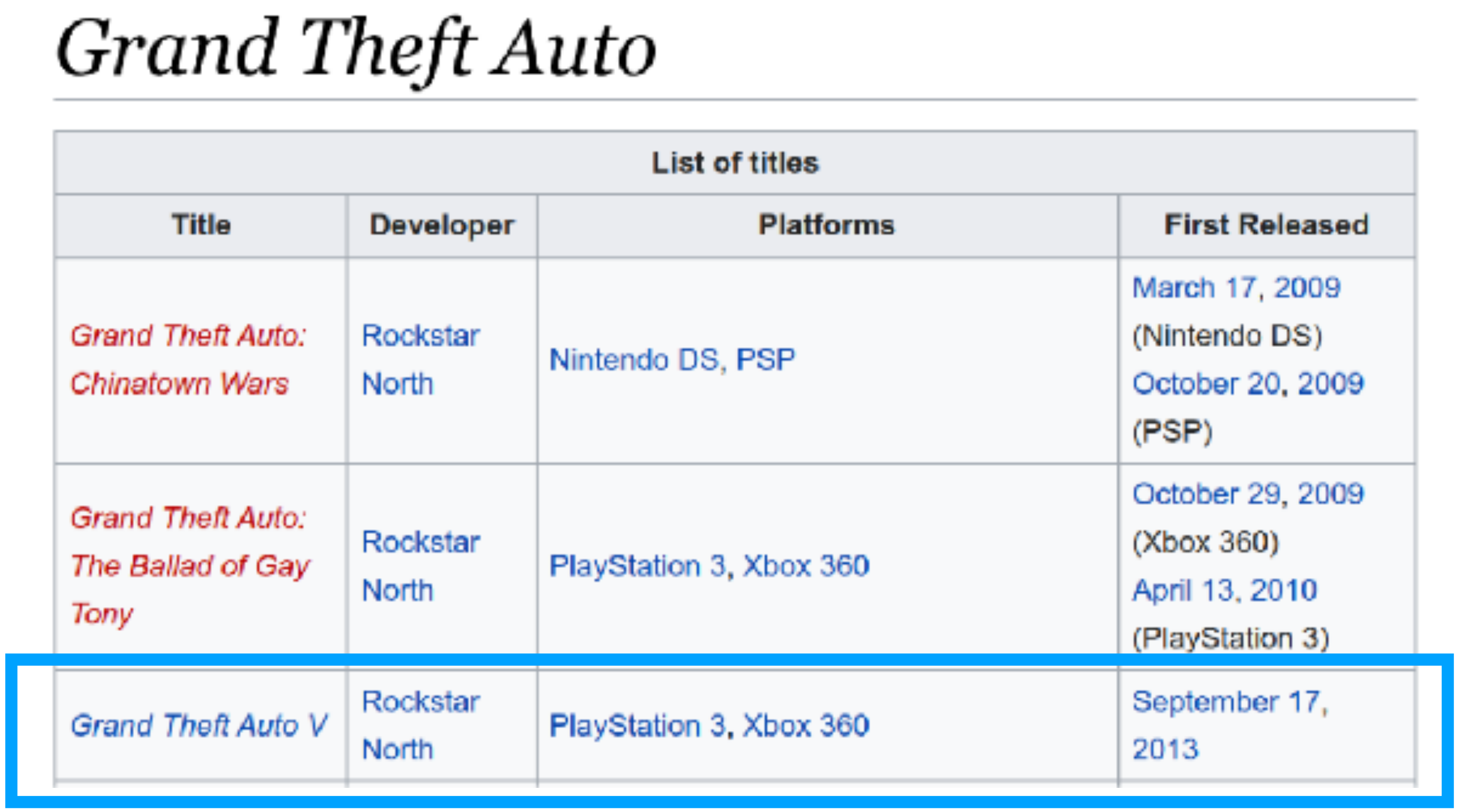
Michael Schumacher

Career summary

Season	Series	Team	Races	Wins	Poles	F/Laps	Podiums	Points	Position
1994	Formula One	Mild Seven Benetton Ford	14	8	6	8	10	92	1st
1995	Formula One	Mild Seven Benetton Renault	17	9	4	8	11	102	1st
1996	Formula One	Scuderia Ferrari S.p.A.	16	3	4	2	8	59	3rd
1997	Formula One	Scuderia Ferrari Marlboro	17	5	3	3	8	78	DSQ
1998	Formula One	Scuderia Ferrari Marlboro	16	6	3	6	11	86	2nd
1999	Formula One	Scuderia Ferrari Marlboro	10	2	3	5	6	44	5th
2000	Formula One	Scuderia Ferrari Marlboro	17	9	9	2	12	108	1st

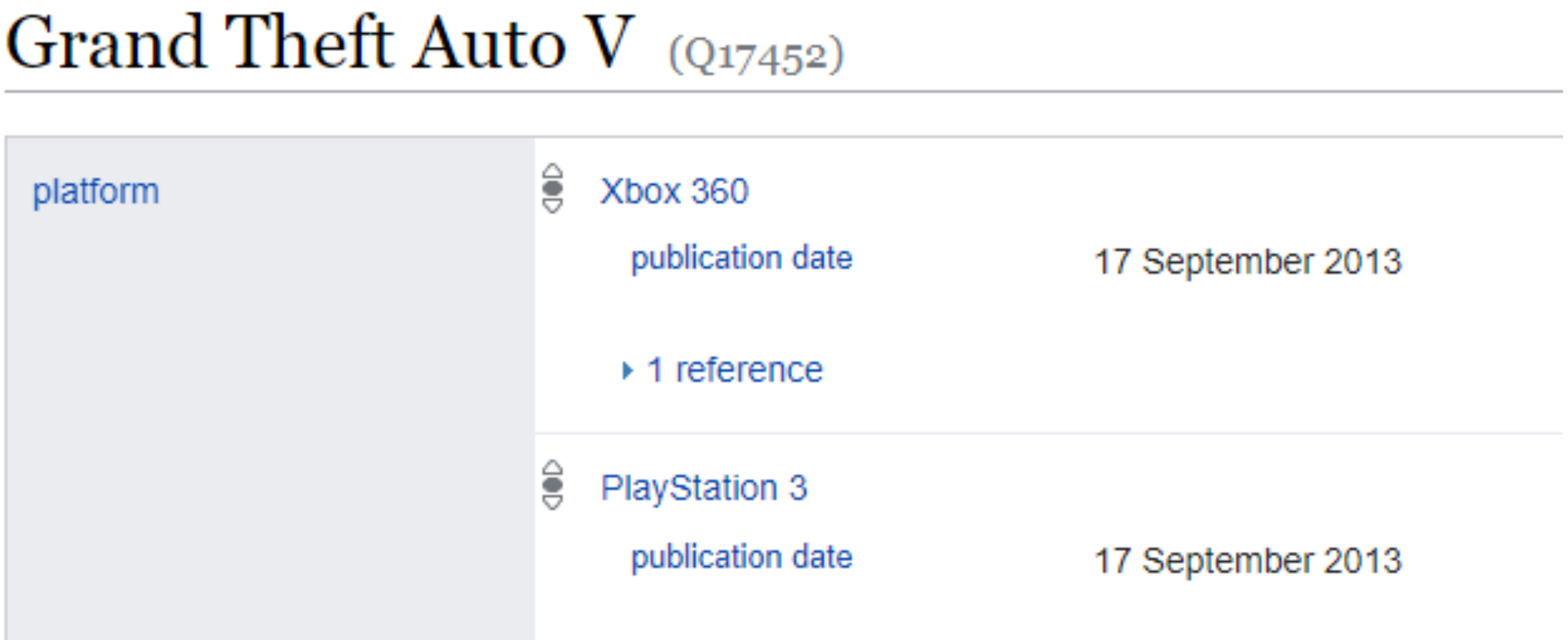
Within-cell year match

Qualifier filter



Grand Theft Auto

List of titles			
Title	Developer	Platforms	First Released
Grand Theft Auto: Chinatown Wars	Rockstar North	Nintendo DS, PSP	March 17, 2009 (Nintendo DS) October 20, 2009 (PSP)
Grand Theft Auto: The Ballad of Gay Tony	Rockstar North	PlayStation 3, Xbox 360	October 29, 2009 (Xbox 360) April 13, 2010 (PlayStation 3)
Grand Theft Auto V	Rockstar North	PlayStation 3, Xbox 360	September 17, 2013



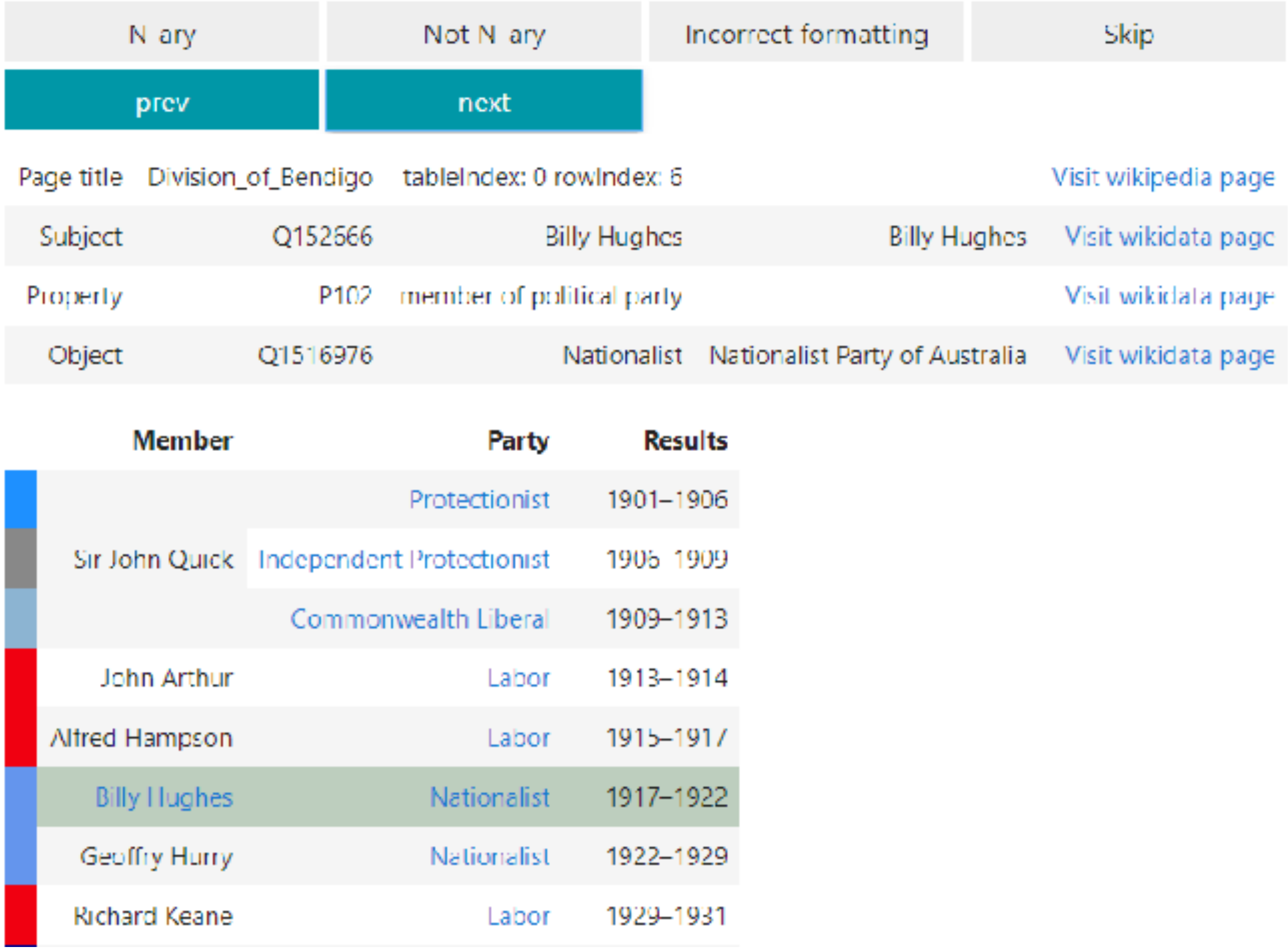
Grand Theft Auto V (Q17452)

platform	Xbox 360	
	publication date	17 September 2013
	▸ 1 reference	
	PlayStation 3	
	publication date	17 September 2013

Quality Evaluation

with our annotation interface

- Runs in Jupyter notebook
- Shows original table with highlighted match
- Used to annotate 750 matches total
- Filtering:
  - 98.4% precision
  - 23.3% recall



Page title	Division_of_Bendigo	tableIndex: 0 rowIndex: 6		Visit wikipedia page
Subject	Q152566	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
Sir John Quick	Protectionist	1901–1906
	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

Wikary: data statistics

Extracted Wikipedia tables

	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	**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%

Wikary: data statistics

Filtering of merged Wikipedia tables

	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%**

Preliminary use-case

Automatically distinguish binary vs. N-ary table

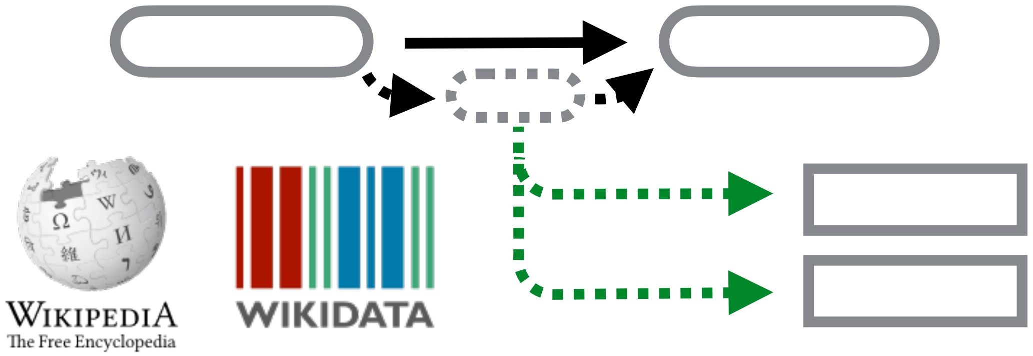
- Trained various linear classifiers
  - N-ary tables **vs.** binary tables from Wikipedia list pages
- Good cross-validation performance
  - 80-90% accuracy
  - Pre-processing for CPA
- **Caveat**: probable data bias
  - samples about different topics

Features
Column names
% of Entity-type columns
% of Numeric-type columns
% of String-type columns
% of Time/date-type columns
Min column-uniqueness
Max column-uniqueness
Mean column-uniqueness

Conclusion

Wikary: N-ary Wikipedia tables matched to Wikidata

- Almost 32.000 tables
- From 3 Wikipedia editions
- Precision-focused qualifier match: 98.4%



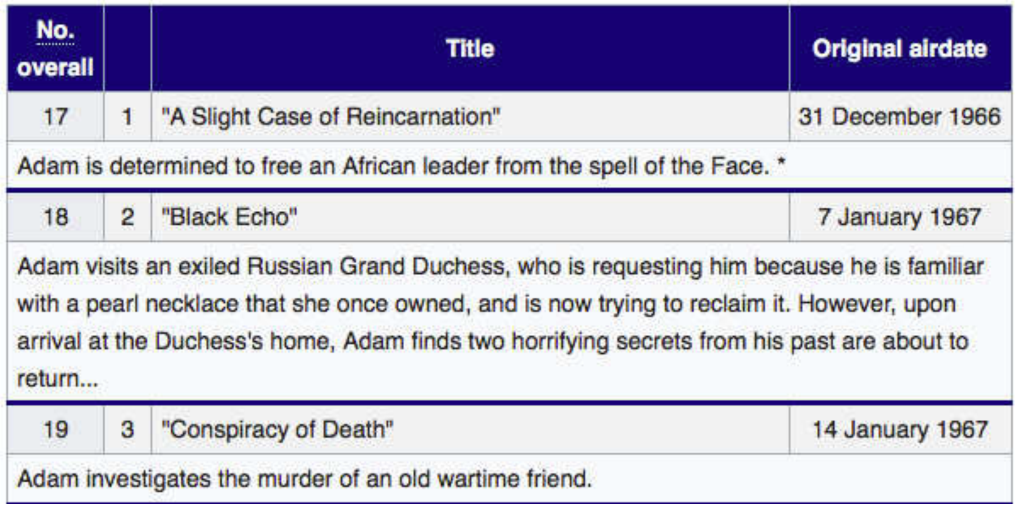
Thank you!

Future Work

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

zenodo.org/record/7025005

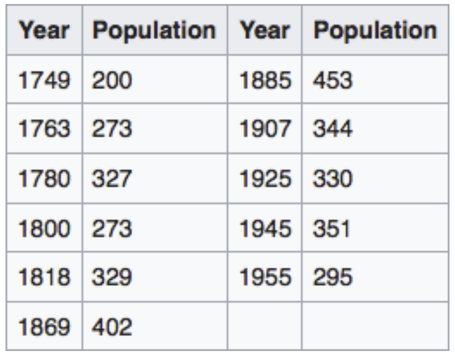
Igor Mazurek
Berend Wiewel
Benno Kruit (b.b.kruit@vu.nl)



No. overall		Title	Original airdate
17	1	"A Slight Case of Reincarnation"	31 December 1966
Adam is determined to free an African leader from the spell of the Face. *			
18	2	"Black Echo"	7 January 1967
Adam visits an exiled Russian Grand Duchess, who is requesting him because he is familiar with a pearl necklace that she once owned, and is now trying to reclaim it. However, upon arrival at the Duchess's home, Adam finds two horrifying secrets from his past are about to return...			
19	3	"Conspiracy of Death"	14 January 1967
Adam investigates the murder of an old wartime friend.			

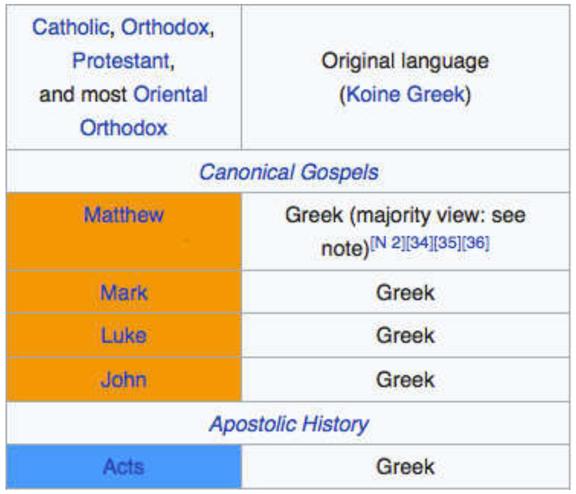
(a) Cells containing long text, wrapped to next row

Extra: Harder tables



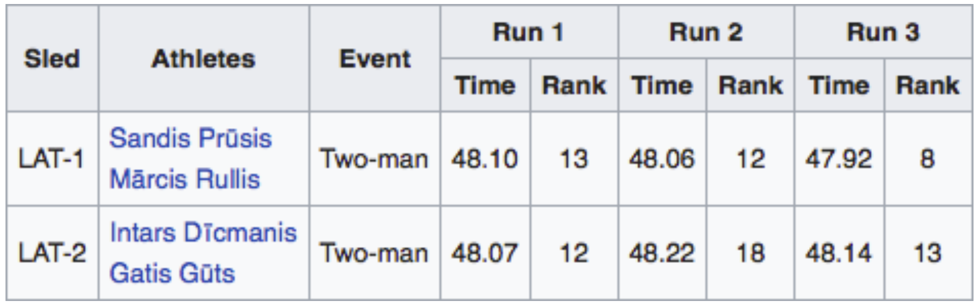
Year	Population	Year	Population
1749	200	1885	453
1763	273	1907	344
1780	327	1925	330
1800	273	1945	351
1818	329	1955	295
1869	402		

(b) Repeated header



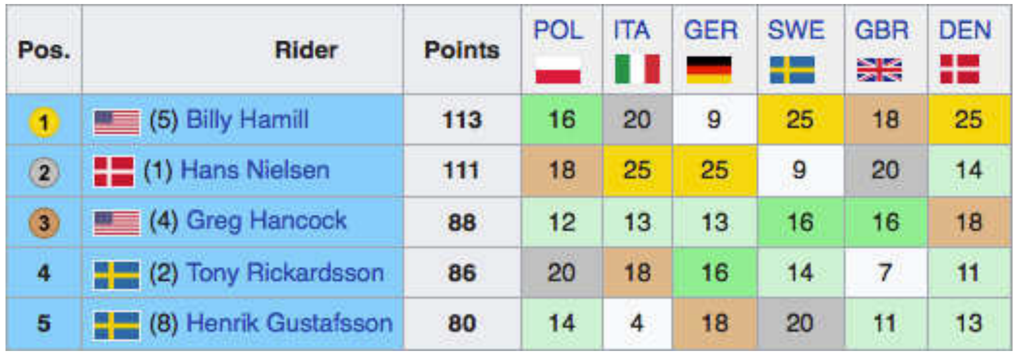
Catholic, Orthodox, Protestant, and most Oriental Orthodox	Original language (Koine Greek)
Canonical Gospels	
Matthew	Greek (majority view: see note)[N 2][34][35][36]
Mark	Greek
Luke	Greek
John	Greek
Apostolic History	
Acts	Greek

(c) Sub-headers



Sled	Athletes	Event	Run 1		Run 2		Run 3	
			Time	Rank	Time	Rank	Time	Rank
LAT-1	Sandis Prūsis Mārcis Rullis	Two-man	48.10	13	48.06	12	47.92	8
LAT-2	Intars Dīcmanis Gatis Gūts	Two-man	48.07	12	48.22	18	48.14	13

(d) Hierarchical header structure



Pos.	Rider	Points	POL 🇵🇱	ITA 🇮🇹	GER 🇩🇪	SWE 🇸🇪	GBR 🇬🇧	DEN 🇩🇰
1	🇺🇸 (5) Billy Hamill	113	16	20	9	25	18	25
2	🇩🇰 (1) Hans Nielsen	111	18	25	25	9	20	14
3	🇺🇸 (4) Greg Hancock	88	12	13	13	16	16	18
4	🇸🇪 (2) Tony Rickardsson	86	20	18	16	14	7	11
5	🇸🇪 (8) Henrik Gustafsson	80	14	4	18	20	11	13

(e) Hybrid relational-matrix table

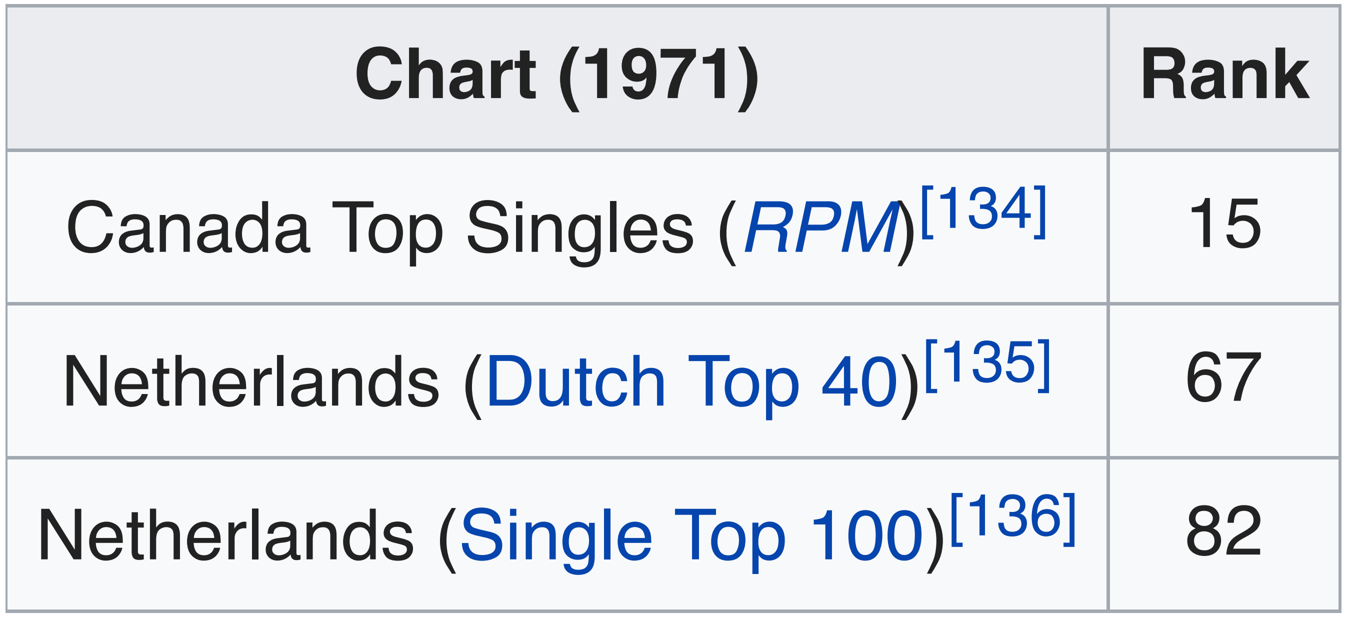


Chart (1971)	Rank
Canada Top Singles (*RPM*)[134]	15
Netherlands (Dutch Top 40)[135]	67
Netherlands (Single Top 100)[136]	82