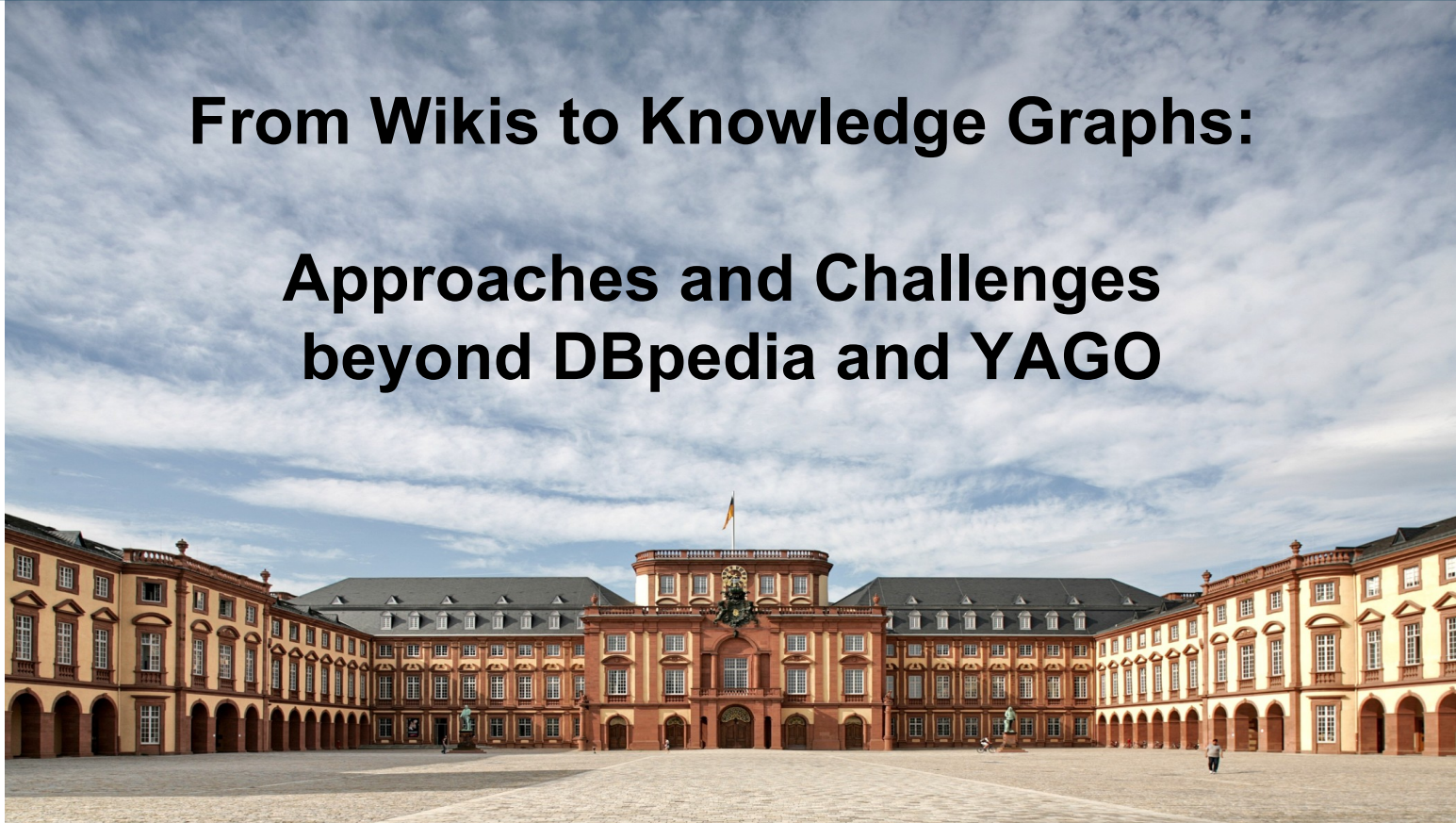


**From Wikis to Knowledge Graphs:  
Approaches and Challenges  
beyond DBpedia and YAGO**



**Heiko Paulheim**

# A Brief History of Knowledge Graphs

● semantic web  
Search term

● knowledge graph  
Search term

+ Add comparison

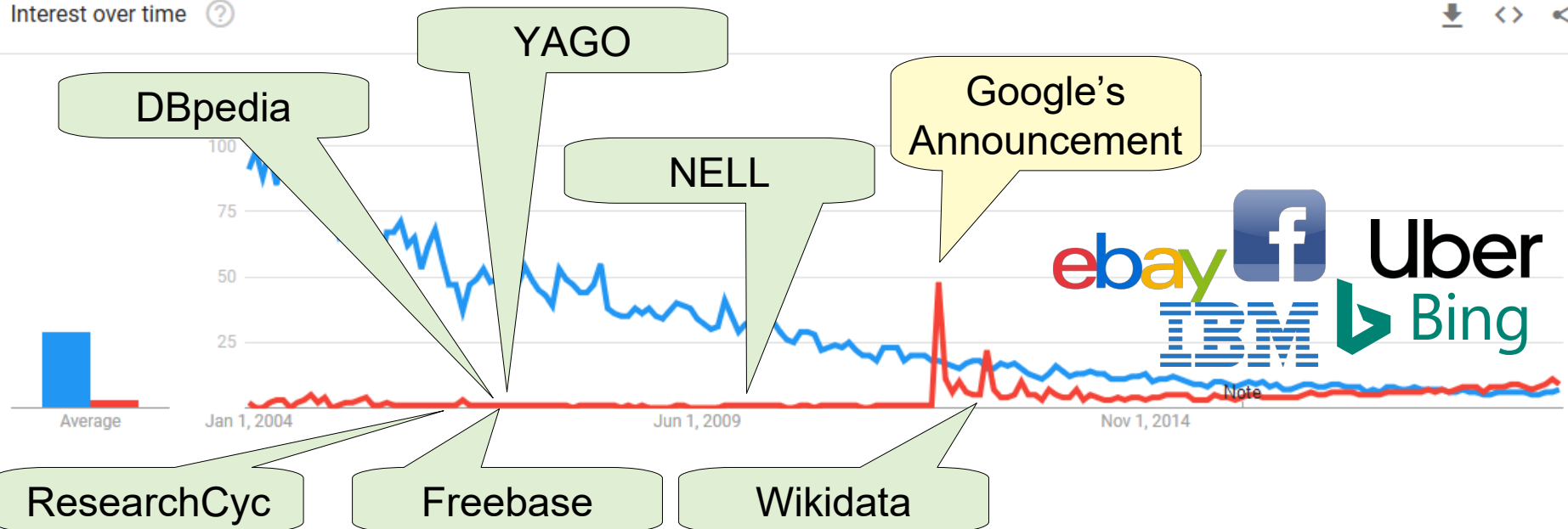
Worldwide ▾

2004 - present ▾

All categories ▾

Web Search ▾

Interest over time ⓘ



# Wikipedia as a Knowledge Graph

- Wikipedia based Knowledge Graphs
  - DBpedia: launched 2007
  - YAGO: launched 2008
  - Extraction from Wikipedia using mappings & heuristics
- Present
  - Two of the most used knowledge graphs
  - ...with Wikidata catching up



# Wikipedia as a Knowledge Graph

The screenshot shows the Wikipedia article for the University of Mannheim. The article text includes a note to not be confused with Mannheim University of Applied Sciences, a paragraph describing the university's founding in 1967 and its origins in the Palatine Academy of Sciences, and another paragraph detailing its undergraduate and graduate programs, campus location, and financial information.

**Contents [hide]**

- 1 History
  - 1.1 20th century
    - 1.1.1 Municipal Commercial College Mannheim (1907–1933)
    - 1.1.2 State College for Economics Mannheim (1946–1967)
    - 1.1.3 University of Mannheim (1967)
  - 1.2 21st century
- 2 Campus
- 3 Organisation and administration
  - 3.1 Schools and Graduate Colleges
  - 3.2 Governance
- 4 Academic profile
  - 4.1 Research institutes and affiliates
  - 4.2 Rankings and reputation
- 5 Student life
  - 5.1 Student organizations
  - 5.2 Sports and athletics
  - 5.3 Traditions
    - 5.3.1 Schlossfest
    - 5.3.2 Schneckenhof Parties
- 6 Notable alumni and faculty members
- 7 See also
- 8 Notes and references
- 9 Further reading
- 10 External links

**History** [ edit ]

The University of Mannheim has no clear foundation date. Its history can be dated back to the establishment of one of its predecessor institutions – the *Kurpfälzische Akademie der Wissenschaften* (Palatine Academy of Sciences) in Mannheim Palace, which was founded by Elector Carl Theodor in 1763. Further predecessors are the *Municipal Commercial College Mannheim* (1907–1933) which was reopened in 1946 as the *State College for Economics Mannheim* and renamed University of Mannheim in 1967.<sup>[4]</sup>

**University of Mannheim**

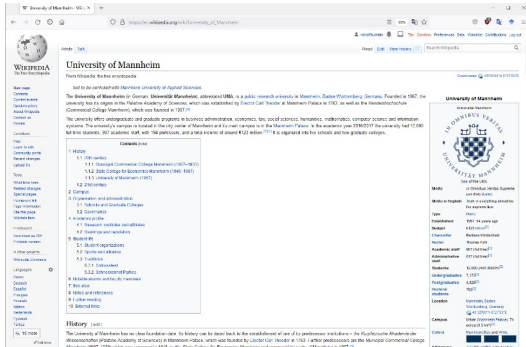
Universität Mannheim

*IN OMNIBUS VERITAS*

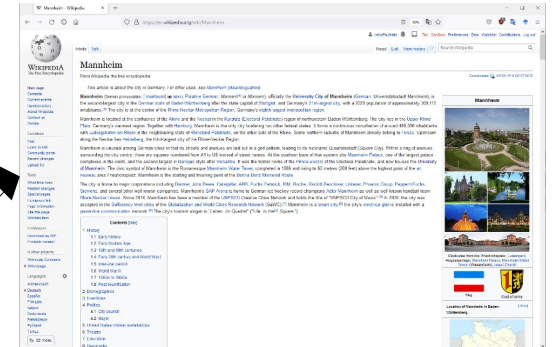
Seal of the UMA

<b>Motto</b>	<i>In Omnibus Veritas Suprema Lex Esto</i> (Latin)
<b>Motto in English</b>	Truth in everything should be the supreme law
<b>Type</b>	Public
<b>Established</b>	1967; 54 years ago
<b>Budget</b>	€123 million <sup>[1]</sup>
<b>Chancellor</b>	Barbara Windscheid
<b>Rector</b>	Thomas Puhl
<b>Academic staff</b>	907 (full time) <sup>[1]</sup>
<b>Administrative staff</b>	617 (full time) <sup>[1]</sup>
<b>Students</b>	12,000 (HWS 2020/21) <sup>[2]</sup>
<b>Undergraduates</b>	7,173 <sup>[2]</sup>
<b>Postgraduates</b>	4,828 <sup>[2]</sup>
<b>Doctoral students</b>	793 <sup>[2]</sup>
<b>Location</b>	Mannheim, Baden-Württemberg, Germany <ul style="list-style-type: none"><li><span><span><span><span><span>49°29′00″N</span> <span>8°27′53″E</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>49.2900°N 8.2753°E</span><span><span>﻿</span> / <span>49.2900; 8.2753</span></span></span></span></span></li></ul>
<b>Campus</b>	Urban (Mannheim Palace), 74 acres (0.3 km²) <sup>[3]</sup>
<b>Colors</b>	Mannheim Blue and White
<b>Affiliations</b>	AACSB; AMBA; CFA Institute;

# Wikipedia as a Knowledge Graph



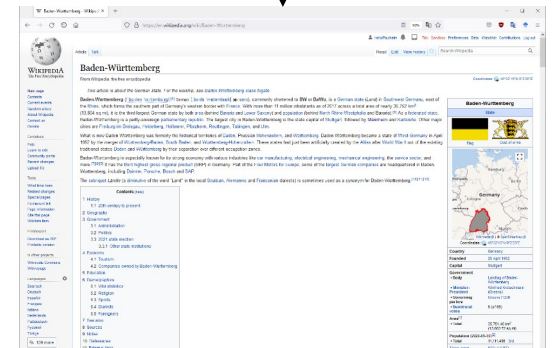
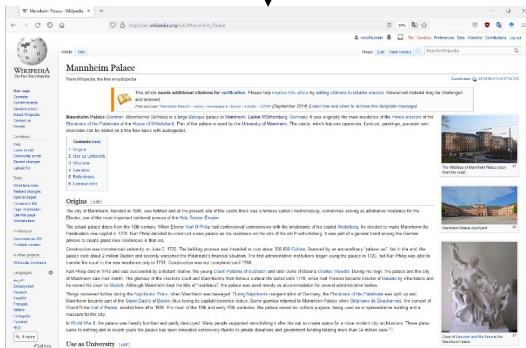
city



campus

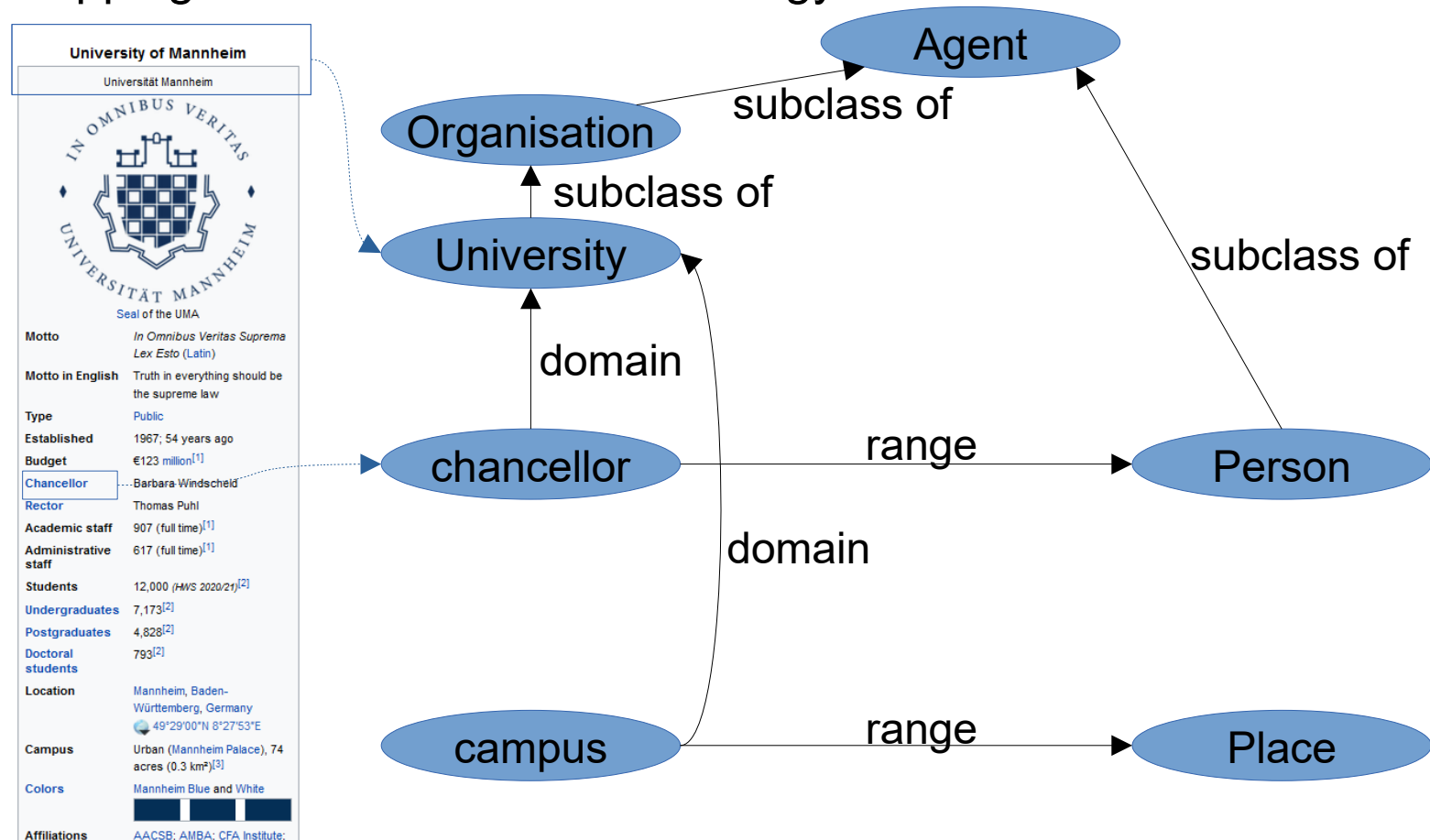
city

state



# Wikipedia as a Knowledge Graph

- Mapping to a central schema/ontology



# Wikipedia as a Knowledge Graph

- General characteristics of DBpedia and YAGO:
  - Central/schema ontology
    - DBpedia: crowdsourcing
    - YAGO: WordNet + categories
  - Mapping of infobox keys
    - DBpedia: crowdsourcing
    - YAGO: engineering
  - One page per entity
    - i.e.: set of entities = set of Wikipedia pages



# Getting the Most out of Wikipedia

- Study for KG-based Recommender Systems\*
  - DBpedia has a coverage of
    - 85% for movies
    - 63% for music artists
    - 31% for books

## Delicious Bookmarks

105,000 bookmarks from 1867 users.

- [README.txt](#)
- [hetrec2011-delicious-2k.zip](#)

## Last.FM

92,800 artist listening records from 1892 users.

- [README.txt](#)
- [hetrec2011-lastfm-2k.zip](#)

## MovieLens + IMDb/Rotten Tomatoes

86,000 ratings from 2113 users.

- [README.txt](#)
- [hetrec2011-movielens-2k.zip](#)

<https://grouplens.org/datasets/>

\*) Di Noia, et al.: *SPRank: Semantic Path-based Ranking for Top-n Recommendations using Linked Open Data*. In: ACM TIST, 2016



# Why bother?

- Experiment w/ recommender systems (LDK 2021)
  - Trained on five versions of DBpedia
  - Biases become evident
    - examined: genre, production country

Genre/KG	de	fr	it	ru	en	$c_e$
Drama	<b>0.198</b>	0.170	0.187	0.172	0.190	0.162
Comedy	0.191	0.192	<b>0.207</b>	0.198	0.166	0.168
Action	0.089	0.010	0.074	<b>0.129</b>	0.112	0.123
Thriller	0.072	0.086	<b>0.097</b>	0.088	0.084	0.095
Romance	0.073	0.055	<b>0.081</b>	0.080	0.052	0.071
Horror	0.043	0.050	0.044	0.043	<b>0.053</b>	0.043
Science Fiction	0.055	0.045	0.044	<b>0.056</b>	0.053	0.073
Adventure	0.053	0.045	0.053	<b>0.070</b>	0.049	0.063
Children's	0.041	<b>0.053</b>	0.052	0.026	0.046	0.031
Crime	0.029	0.039	0.025	0.044	<b>0.045</b>	0.038

Country/KG	de	fr	it	ru	en	$c_e$
USA	0.728	0.750	0.762	0.761	<b>0.782</b>	0.744
UK	0.136	<b>0.143</b>	0.098	0.091	<b>0.108</b>	0.110
France	0.028	<b>0.030</b>	0.036	<b>0.037</b>	0.026	0.033
Germany	<b>0.012</b>	0.018	0.012	0.030	<b>0.034</b>	0.025
Italy	<b>0.016</b>	0.009	0.013	0.009	0.009	0.013
Canada	0.020	<b>0.009</b>	<b>0.021</b>	0.005	<b>0.006</b>	0.015
Australia	0.017	0.010	0.013	0.008	<b>0.020</b>	0.016
Japan	0.006	0.005	<b>0.012</b>	0.004	0.006	0.007
Spain	0.006	0.004	<b>0.006</b>	0.002	0.005	0.005
Mexico	0.004	0.001	0.005	<b>0.006</b>	0.002	0.008

Voit & Paulheim (2021): Bias in Knowledge Graphs.

# Why bother?

- One key take away of that paper:
- Rethink parameter tuning and ablation studies!
  - We see ablation studies on methods, parameters, etc.
  - But rarely on knowledge graphs
  - However, there are considerable differences
    - observed in this work: factor of 2-3
- Especially:
  - **entity coverage** and **level of detail**

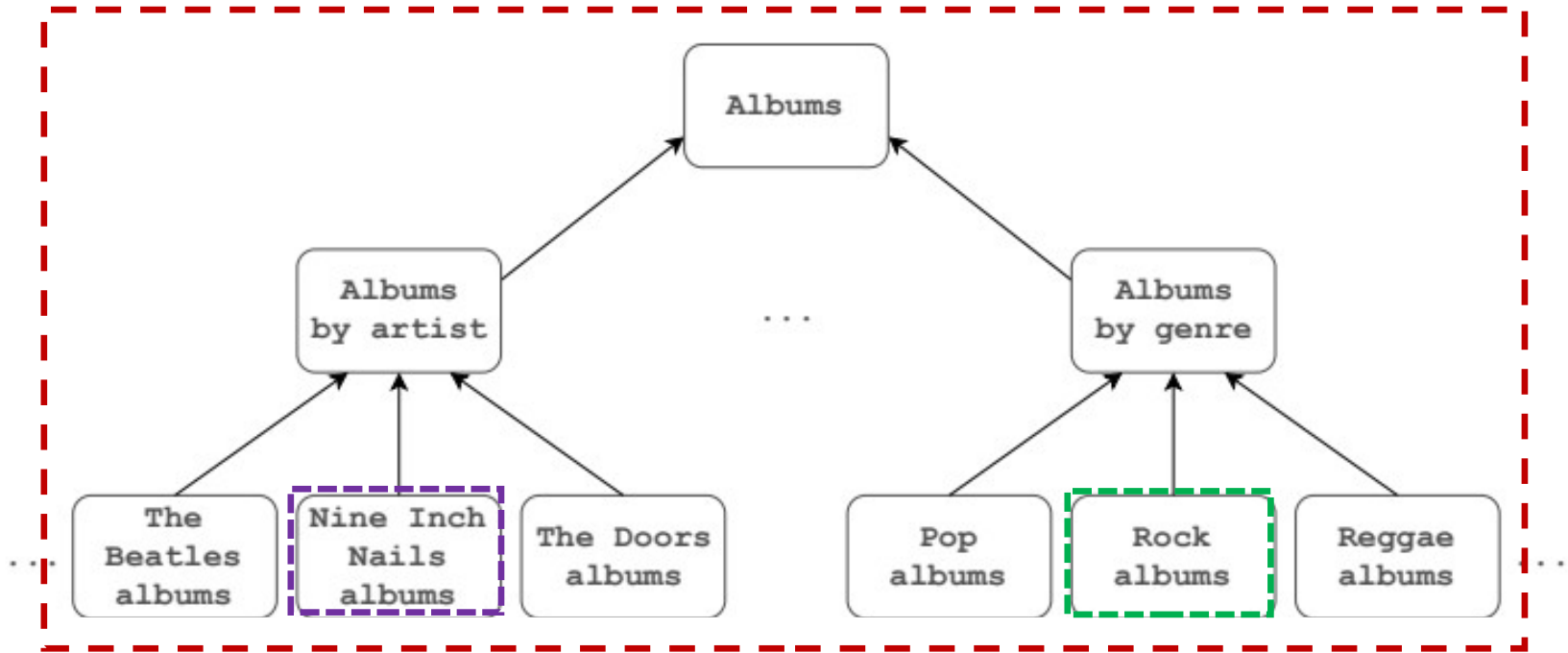


Voit & Paulheim (2021): Bias in Knowledge Graphs.

# Increasing Level of Detail

- YAGO uses categories for types
  - e.g., Category:American Industrial Groups
  - but does not analyze them further
- `:NineInchNails a :AmericanIndustrialGroup`
  - “Things, not Strings”?
- `:NineInchNails a :MusicalGroup ;  
hometown :United_States ;  
genre :Industrial .`

# Cat2Ax: Axiomatizing Wikipedia Categories



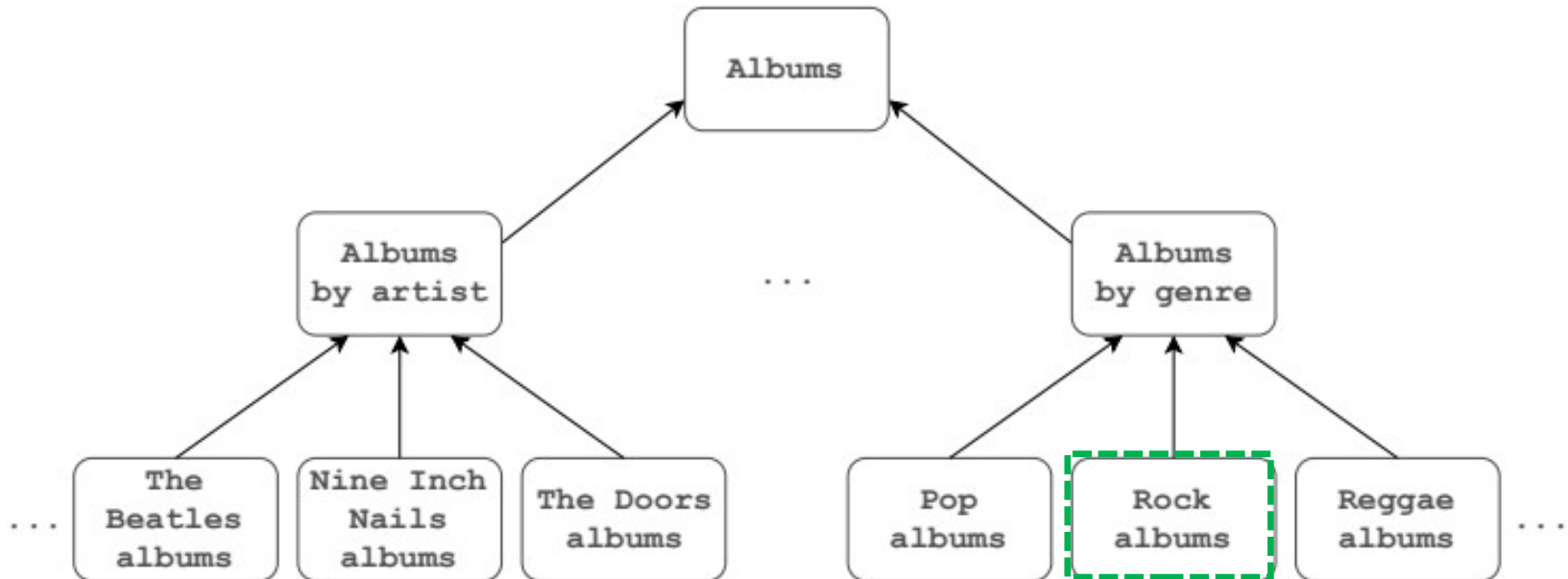
$\subseteq$  `dbo:Album`

$\subseteq$  `dbo:artist.{dbr:Nine_Inch_Nails}`

$\subseteq$  `dbo:genre.{dbr:Rock_Music}`

See: ISWC 2019 Paper on Uncovering the Semantics of Wikipedia Categories

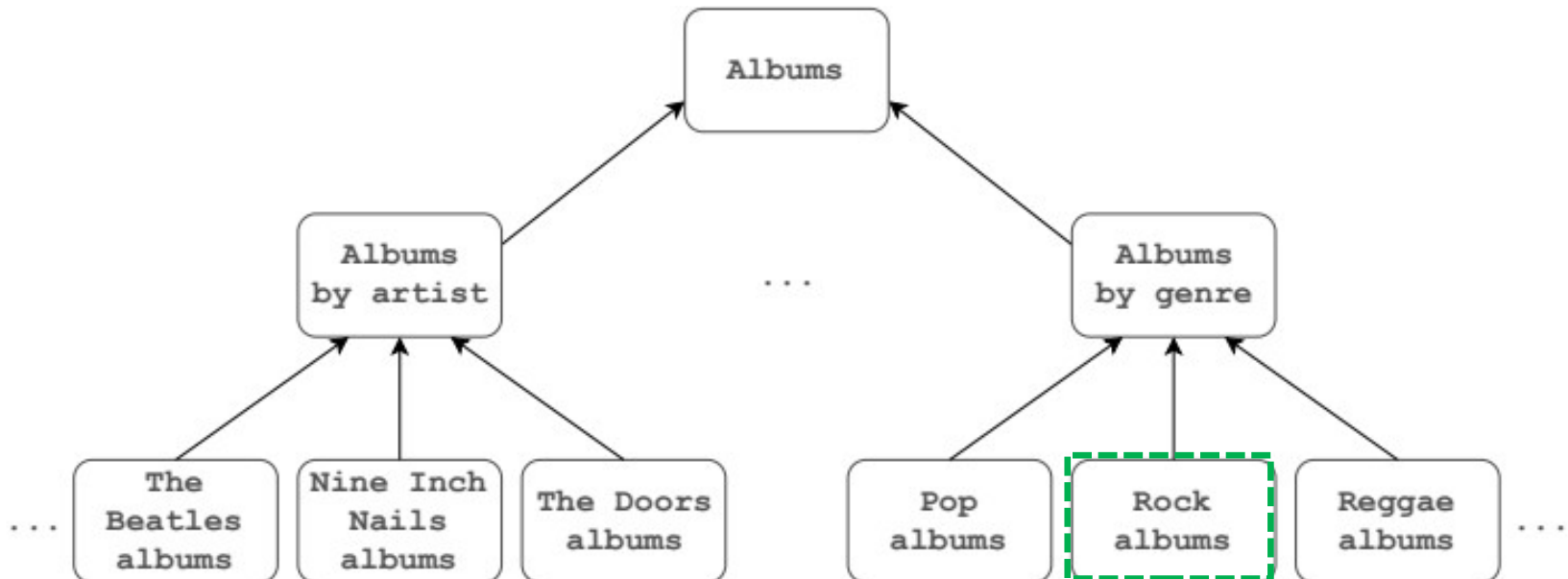
# Cat2Ax: Axiomatizing Wikipedia Categories



$\subseteq$  `dbo:genre.{dbr:Rock_Music}` ?

$\subseteq$  `dbo:artist.{dbr:Rock_(Rapper)}` ?

# Cat2Ax: Axiomatizing Wikipedia Categories



- Frequency: how often does the pattern occur in a category?
  - i.e.: share of instances that have `dbo:genre.{dbr.Rock_Music}`?
- Lexical score: likelihood of term as a surface form of object
  - i.e.: how often is *Rock* used to refer to `dbr:Rock_Music`?
- Sibling score: how likely are sibling categories sharing similar patterns?
  - i.e., are there sibling categories with a high score for `dbo:genre`?

# Cat2Ax: Axiomatizing Wikipedia Categories

- Results

Approach	Count	Precision [%]	Count	Precision [%]
	Relation axioms		Type axioms	
Cat2Ax	272,707	95.6	430,405	96.8
C-DF	143,850	83.6	28,247	92.0
Catriple	306,177	87.2	–	–
	Relation assertions		Type assertions	
Cat2Ax	4,424,785 (7,554,980)	87.2 (92.1)	3,342,057 (12,111,194)	90.8 (95.7)
C-DF	766,921 (2,856,592)	78.4 (93.4)	198,485 (2,352,474)	76.8 (97.1)
Catriple	6,260,972 (6,836,924)	74.4 (76.5)	–	–

# CaLiGraph Example

rdfs:label • Tiamat

owl:sameAs • [dbr:Tiamat\\_\(band\)](#)

elgo:activeYearsStartYear • 1987

elgo:genre • [Symphonic metal](#)

elgo:hometown • [Sweden](#)

Category: Musical Groups established in 1987

List of symphonic metal bands

Category: Swedish death metal bands  
List of Swedes in Music



# Improving Entity Coverage: Lists in Wikipedia

- Only existing pages have categories
  - Lists may also link to non-existing pages

## List of intelligent dance music artists

From Wikipedia, the free encyclopedia



This section **does not cite any sources**. Please help improve this section by adding citations to reliable sources. Unsourced material may be removed and the article may become a list. *Find sources:* "List of intelligent dance music artists" – news · newspapers · books · scholar · JSTOR (June 2015) *(Learn how and when to remove this message)*

This is a list of notable music artists who play **intelligent dance music** (IDM) genre.

### Contents [hide]

- #
- A-K
- L-Z
- References

### # [edit]

- 808 State
- µ-ziq

### A-K [edit]

- |                    |                        |                       |                        |                               |                              |
|--------------------|------------------------|-----------------------|------------------------|-------------------------------|------------------------------|
| • Actress          | • Benn Jordan          | • Casino Versus Japan | • Dopplereffekt        | • Funkstörung                 | • Jan Jelinek                |
| • Acoustic         | • Biosphere            | • Ceephax Acid Crew   | • Chris Douglas        | • The Future Sound Of London  | • Jega                       |
| • Air Liquide      | • Björk <sup>[1]</sup> | • Cex                 | • Drexciya             | • Gas                         | • Jello                      |
| • Alarm Will Sound | • The Black Dog        | • Christ              | • Eight Frozen Modules | • Gescom                      | • Jlin                       |
| • Alva Noto        | • Blanck Mass          | • Chris Clark         | • Emptyset             | • Global Communication        | • John Tejada                |
| • Amon Tobin       | • Boards of Canada     | • Ciocolan            | • Esem                 | • Global Goon                 | • Jon Hopkins <sup>[3]</sup> |
| • Andy Stott       | • Bochum Welt          | • Cylob               | • FaltyDL              | • Goldie                      | • Kettel                     |
| • Aphex Twin       | • Boom Bip             | • Daedelus            | • Fennesz              | • Zachary Gray <sup>[2]</sup> | • Kevin Blechdom             |
| • Apparat          | • Brothomstates        | • Deadbeat            | • The Field            | • Gridlock                    | • Kid606                     |
| • Arovane          | • Burial               | • Deepchord           | • The Flashbulb        | • Himuro Yoshiteru            | • Kodomo                     |
| • Atypic           | • Bvdub                | • Demdike Stare       | • Floating Points      | • Kim Hiorthøy                | • Koreless <sup>[4]</sup>    |
| • Autechre         | • C418                 | • Deru                | • Flying Lotus         | • I am Robot and Proud        |                              |
| • B12              | • Cabaret Voltaire     | • Richard Devine      | • Forest Swords        | • Innovaders                  |                              |

## Delicious Bookmarks

105,000 bookmarks from 1867 users.

- [README.txt](#)
- [hetrec2011-delicious-2k.zip](#)

## Last.FM

92,800 artist listening records from 1892 users.

- [README.txt](#)
- [hetrec2011-lastfm-2k.zip](#)

## MovieLens + IMDb/Rotten Tomatoes

86,000 ratings from 2113 users.

- [README.txt](#)
- [hetrec2011-movielens-2k.zip](#)

# Pushing Entity Coverage Further

- Beyond red links (2020)

Cinematic films			
Title	Running time	Year released	Notes
<i>Amra Ekta Cinema Banabo (The Innocence)</i>	1265 min (21 hr, 5 min)	2019	[31][32]
<i>Resan (The Journey)</i>	873 min (14 hr, 33 min)	1987	[33]
<i>La Flor</i>	803 min (13 hr, 23 min)	2018	[34]
<i>Out 1 (Noli me tangere)</i>	775 min (12 hr, 55 min)	1971	[35]
<i>Evolution of a Filipino Family</i>	593 min (9 hr, 53 min)	2004	[36]
<i>Shoah</i>	566 min (9 hr, 26 min)	1985	[37]
<i>Tie Xi Qu: West of the Tracks</i>	551 min (9 hr, 11 min)	2003	[38]
<i>Death in the Land of Encantos</i>	538 min (8 hr, 58 min)	2007	[39]
<i>Dead Souls</i>	495 min (8 hr, 15 min)	2018	[40]
<i>A Lullaby to the Sorrowful Mystery</i>	485 min (8 hr, 5 min)	2016	[41]
<i>O.J.: Made in America</i>	463 min (7 hr, 43 min)	2016	[42]
<i>Melancholia</i>	450 min (7 hr, 30 min)	2008	[43]
<i>Sátántangó</i>	419 min (6 hr, 59 min)	1994	[44]
<i>La Roue</i>	413 min (6 hr, 53 min)	1923 (Restoration, 2019)	[45]
<i>The Best of Youth</i>	366 min (6 hr, 6 min)	2003	[46]
<i>Century of Birthing</i>	360 min (6 hr)	2011	[47]
<i>Near Death</i>	358 min (5 hr, 58 min)	1989	[48]
<i>Karamay</i>	356 min (5 hr, 56 min)	2011	[49]
<i>Little Dorrit</i>	350 min (5 hr, 50 min)	1987	[50]
<i>Carlos</i>	339 min (5 hr, 39 min)	2010	[51]
<i>Mula sa Kung Ano ang Noon</i>	338 min (5 hr, 38 min)	2014	[52]
<i>Napoléon</i>	332 min (5 hr, 32 min)	1927 (Restoration, 2016)	[53]
<i>1900</i>	317 min (5 hr, 17 min)	1976	[54]
<i>Happy Hour</i>	317 min (5 hr, 17 min)	2015	[55]
<i>Batang West Side</i>	315 min (5 hr, 15 min)	2001	[56]
<i>The Deluge</i>	315 min (5 hr, 15 min)	1974	[57]
<i>Fanny and Alexander</i>	312 min (5 hr, 12 min)	1982	[58]
<i>Tsahal</i>	304 min (5 hr, 4 min)	1994	[59]

- Beyond explicit lists (2021)

## Members [\[ edit \]](#)

- Jürgen Engler – vocals, guitar, keyboards, synthesizers and programming, metallic percussion (1980–1985, 1989–1997, 2005–present)
- Ralf Dörper – keyboards, synthesizers and programming (1980–1982, 1985, 1989–1997, 2005–present)
- Marcel Zürcher – guitar, keyboards (2005–present)
- Nils Finkeisen - guitar (2015–present)
- Paul Keller - drums (2018–present)

## Former members [\[ edit \]](#)

- Bradley Bills - live drums (2013–2014)
- Rüdiger Esch - bass guitar (1989–1997, 2005–2011)
- Christoph "Nook" Michelfeit - drums, electronic percussion
- Bernhard Malaka - bass guitar (1980–1982)
- Hendrik Thiesbrummel - live drums (2016–2018)
- Frank Köllges - drums
- Eva Gossling - saxophone (1981)
- Christina Schnekenburger - keyboards
- Walter Jäger - ?
- Christopher Lietz - programming, samples (1995–1997)
- Lee Altus - guitar (1992–1997)
- Darren Minter - drums (1993)
- George Lewis - drums (1997)
- Oliver Röhl – drums
- Achim Farber – drums
- Volker Borchert – drums (1992, 2015–2016)

## Discography [\[ edit \]](#)

### Albums [\[ edit \]](#)

- *Stahlwerksynfonie* (1981)
- *Volle Kraft Voraus!* (1982)
- *Entering the Arena* (1985)
- *I* (1992)
- *II - The Final Option* (1993)
- *The Final Remixes* (1994)
- *III - Odyssey of the Mind* (1995)
- *Paradise Now* (1997)
- *The Machinists of Joy* (2013)
- *V - Metal Machine Music* (2015)
- *Stahlwerkrequiem* (2016)
- *Live Im Schatten Der Ringe* (2016)

# Entity Extraction from List Pages

- Red and grey links
  - Red links point to entities that do not exist
  - “Grey links”
    - are actually not links
    - i.e., entities to be discovered

Cinematic films

Title	Running time	Year released	Notes
<a href="#">Amra Ekta Cinema Banabo (The Innocence)</a>	1265 min (21 hr, 5 min)	2019	[31][32]
<a href="#">Resan (The Journey)</a>	873 min (14 hr, 33 min)	1987	[33]
<a href="#">La Flor</a>	803 min (13 hr, 23 min)	2018	[34]
<a href="#">Out 1 (Noli me tangere)</a>	775 min (12 hr, 55 min)	1971	[35]
<a href="#">Evolution of a Filipino Family</a>	593 min (9 hr, 53 min)	2004	[36]
<a href="#">Shoah</a>	566 min (9 hr, 26 min)	1985	[37]
<a href="#">Tie Xi Qu: West of the Tracks</a>	551 min (9 hr, 11 min)	2003	[38]
<a href="#">Death in the Land of Encantos</a>	538 min (8 hr, 58 min)	2007	[39]
<a href="#">Dead Souls</a>	495 min (8 hr, 15 min)	2018	[40]
<a href="#">A Lullaby to the Sorrowful Mystery</a>	485 min (8 hr, 5 min)	2016	[41]
<a href="#">O.J.: Made in America</a>	463 min (7 hr, 43 min)	2016	[42]
<a href="#">Melancholia</a>	450 min (7 hr, 30 min)	2008	[43]
<a href="#">Sátántangó</a>	419 min (6 hr, 59 min)	1994	[44]
<a href="#">La Roue</a>	413 min (6 hr, 53 min)	1923 (Restoration, 2019)	[45]
<a href="#">The Best of Youth</a>	366 min (6 hr, 6 min)	2003	[46]
<a href="#">Century of Birthing</a>	360 min (6 hr)	2011	[47]
<a href="#">Near Death</a>	358 min (5 hr, 58 min)	1989	[48]
<a href="#">Karamay</a>	356 min (5 hr, 56 min)	2011	[49]
<a href="#">Little Dorrit</a>	350 min (5 hr, 50 min)	1987	[50]
<a href="#">Carlos</a>	339 min (5 hr, 39 min)	2010	[51]
<a href="#">Mula sa Kung Ano ang Noon</a>	338 min (5 hr, 38 min)	2014	[52]
<a href="#">Napoléon</a>	332 min (5 hr, 32 min)	1927 (Restoration, 2016)	[53]
<a href="#">1900</a>	317 min (5 hr, 17 min)	1976	[54]
<a href="#">Happy Hour</a>	317 min (5 hr, 17 min)	2015	[55]
<a href="#">Batang West Side</a>	315 min (5 hr, 15 min)	2001	[56]
<a href="#">The Deluge</a>	315 min (5 hr, 15 min)	1974	[57]
<a href="#">Fanny and Alexander</a>	312 min (5 hr, 12 min)	1982	[58]
<a href="#">Tsayal</a>	304 min (5 hr, 4 min)	1994	[59]

# Entity Extraction from List Pages

- Lists form (shallow) hierarchies

## Lists of writers

From Wikipedia, the free encyclopedia

The following are lists of writers:

- Bestsellers
- Biographers
- Buddhism
- Business theorists
- Catholicism
- Children's literature
- Christian fiction
- Cricket
- Crime
- Detective fiction
- Drama
- Essays
- Fantasy
- Fiction
- Historical novels
- Horror fiction
- Horsemanship
- Illustrations
- Manga
- Music theory
- Mysteries
- Non-fiction
- Novels
- Occult
- Plays
- Poetry
- Politics
- Role-playing games
- Romantic novels
- Science fiction
- Self-help
- Short stories
- Software
- Technical writers
- Thrillers
- Translations
- Western fiction
- Young adult

Top of page

### Contents [hide]

- 1 Lists by language (non-English)
- 2 Lists by ethnicity or nationality
- 3 Lists of women writers and works
- 4 Lists by publisher
- 5 See also
- 6 External links

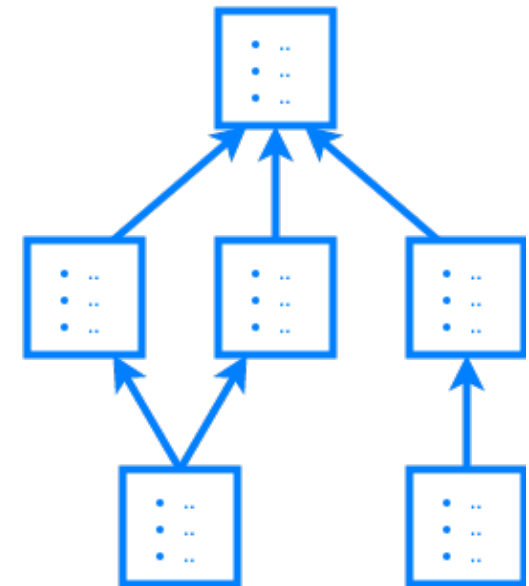
## Lists by language (non-English) [edit]

- Ancient Greek
- Arabic
- Bengali
- Catalan
- Chichewa
- Dutch
- French
- German
- Gujarati
- Greek
- Hebrew
- Hindi
- Leonese
- Lithuanian
- Malayalam
- Marathi
- Nepali
- Odia
- Pukhto or Pashto
- Persian
- Polish
- Portuguese
- Russian
- Spanish
- Swedish
- Tamil
- Turkish
- Urdu
- Welsh

Top of page

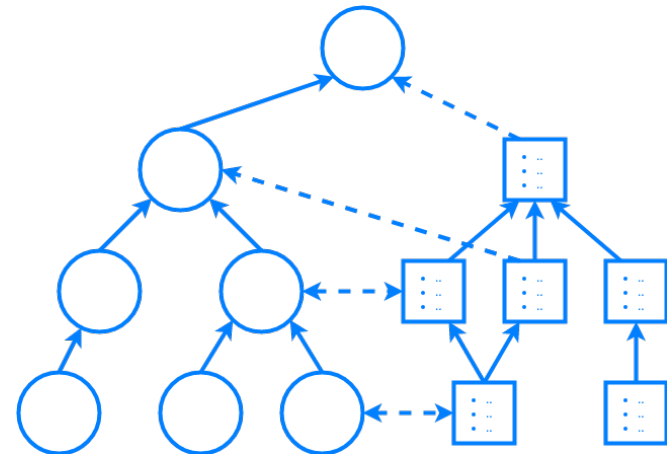
## Lists by ethnicity or nationality [edit]

- African writers
- African-American writers
- Albanian writers
- Algerian writers
- American writers
- Arab American writers
- Armenian authors
- Barbadian writers
- Beninese writers
- Black British writers
- Bosniak writers
- Brazilian writers
- Canadian writers
- Chinese writers
- English writers
- Egyptian writers
- Georgian writers
- German writers
- Ghanaian writers
- Greek writers
- Guyanese writers
- Irish writers
- Italian writers
- Jewish American writers
- Kenyan writers
- Korean American writers
- Macedonian writers
- Mexican writers
- Nepali writers
- New Zealand writers
- Nigerian writers
- Norwegian writers
- American novelists
- Pakistani writers
- Peruvian writers
- Romanian writers
- Russian authors
- Scottish writers
- Serbian writers
- Slovak authors
- Slovenian writers
- Somali writers



# Entity Extraction from List Pages

- Idea: align with category graph
- Equivalence:
  - “List of Japanese Writers”  
↔ Category: Japanese Writers
- Subsumption:
  - “List of Japanese Speculative Fiction Writers”  
→ Category: Japanese Writers



# Classifying Red Links

- Not all entities on a list page belong to the same category
  - Idea:
    - Learn classifier to tell subject entities from non-subject entities
  - Distant learning approach
    - Positive examples:
      - Entities that are in the corresponding category
    - Negative examples
      - Entities that are in a category which is disjoint
      - e.g., Book <> Writer
- [Patricia Aakhus](#) (1952–2012), *The Voyage of Mael Duin's Curragh*
  - [Atia Abawi](#)
  - [Edward Abbey](#) (1927–1989), *The Monkey Wrench Gang*
  - [Lynn Abbey](#) (born 1948), *Daughter of the Bright Moon*
  - [Belle Kendrick Abbott](#) (1842–1893), *Leah Mordecai*
  - [Eleanor Hallowell Abbott](#) (1872–1958), poet, novelist and short story writer
  - [Hailey Abbott](#), *Summer Boys*
  - [Megan Abbott](#) (born 1971), *Die A Little*
  - [Shana Abé](#), *A Rose in Winter*
  - [Louise Abeita](#) (1926–2014), Native American Isleta Pueblo writer, *I am a Pueblo Indian Girl*
  - [Robert H. Abel](#) (1941–2017)
  - [Aberjhani](#)
  - [Walter Abish](#) (born 1931), *How German Is It*
  - [Abiola Abrams](#) (born 1976), TV host, art filmmaker and author, *Dare*
  - [Diana Abu-Jaber](#) (born 1960), *Arabian Jazz*
  - [Susan Abulhawa](#), *Mornings in Jenin*
  - [Kathy Acker](#) (1947–1997), *Blood and Guts in High School*
  - [Cherry Adair](#), *Black Magic*
  - [Alice Adams](#) (1926–1999), *Beautiful Girl*
  - [Victoria Aveyard](#) (born 1990), *Red Queen* series

# Classifying Red Links

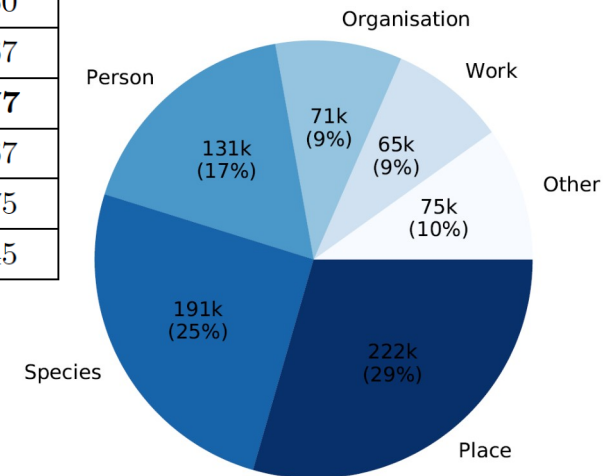
- Using a mix of features
  - Page layout, position of entities, statistical, linguistics, ...

	Feature Type	Features
Shared	Page	# sections
	Positional	Position of section in LP
	Linguistic	Section title, POS/NE tag of entity and its direct context
Enum	Page	# entries, Avg. entry indentation level, Avg. entities/words/characters per entry, Avg. position of first entity
	Positional	Position of entry in enumeration, Indentation level of entry, # of sub-entries of entry, Position of entity in entry
	Custom	# entities in current entry, # mentions of entity in same/other enumeration of LP
Table	Page	# tables, # rows, # columns, Avg. rows/columns per table, Avg. entities/words/characters per row/column, Avg. first column with entity
	Positional	Position of table in LP, Position of row/column in table, Position of entity in row
	Linguistic	Column header is synonym/hyponym of word in LP title
	Custom	# entities in current row, # mentions of current entity in same/other table of LP

# Classifying Red Links

- Enumerations work slightly better than tables
- Unevenly balanced
  - >70% place, species, and person

Algorithm	Enum			Table		
	P	R	F1	P	R	F1
Baseline (pick first entity)	74	96	84	64	53	58
Naive Bayes	80	90	84	34	<b>91</b>	50
Decision Tree	82	78	80	67	66	67
Random Forest	85	<b>90</b>	<b>87</b>	85	71	<b>77</b>
XG-Boost	<b>90</b>	83	86	<b>90</b>	53	67
Neural Network (MLP)	86	84	85	78	72	75
SVM	86	60	71	73	33	45





# Beyond List Pages

- Many pages contain list-like constructs
- Usually
  - small
  - same type
  - same relation to page entity
  - more grey links

## Axl Rose

From Wikipedia, the free encyclopedia

...

### Discography [ edit ]

#### with Guns N' Roses [ edit ]

- *Appetite for Destruction* (1987)
- *G N' R Lies* (1988)
- *Use Your Illusion I* (1991)
- *Use Your Illusion II* (1991)
- *"The Spaghetti Incident?"* (1993)
- *Chinese Democracy* (2008)

#### with Hollywood Rose [ edit ]

- *The Roots of Guns N' Roses* (2004)

#### with Rapidfire [ edit ]

- *Ready to Rumble EP* (2014)

#### Guest appearances [ edit ]

- *The Decline of Western Civilization Part II: The Metal Years – Original Motion Picture Soundtrack* by various artists (1988; "Under My Wheels" ft. Alice Cooper, Slash and Izzy Stradlin)
- *The End of the Innocence* by Don Henley (1989; "I Will Not Go Quietly")
- *Fire and Gasoline* by Steve Jones (1989; "I Did U No Wrong")
- *Pawnshop Guitars* by Gilby Clarke (1994; "Dead Flowers")
- *Anxious Disease* by The Outpatience (1996; "Anxious Disease" ft. Slash)
- *Angel Down* by Sebastian Bach (2007; "Back in the Saddle," "(Love Is) a Bitchslap," "Stuck Inside")
- *New Looney Tunes* (2018, "Rock the Rock")<sup>[122]</sup>

### Filmography [ edit ]

Title	Year	Role	Notes
<i>The Dead Pool</i>	1988	Musician at funeral	Uncredited
<i>Grand Theft Auto: San Andreas</i> (video game)	2004	DJ Tommy "The Nightmare" Smith in the K-DST radio	Voice
<i>That Metal Show</i>	2011	Himself	
<i>Jimmy Kimmel Live!</i>	2012	Himself	
<i>New Looney Tunes</i> (TV show) <sup>[123]</sup>	2018	Himself	Voice
<i>Scooby-Doo and Guess Who?</i> (TV Show)	2021	Himself	Voice

# Beyond List Pages

(artist,  
Axl\_Rose)  
 $\exists topSection.\{„Discography“\}$   
 $\sqsubseteq \exists artist.\{>PageEntity<\}$

(type,  
MusicalWork)  
 $\exists topSection.\{„Discography“\}$   
 $\sqsubseteq MusicalWork$

(musicalBand,  
Guns\_N'\_Roses)  
 $\exists topSection.\{„Discography“\}$   
 $\cap sectionEntityType.\{Band\}$   
 $\sqsubseteq$   
 $\exists musicalBand.\{>SectionEntity<\}$

Axl Rose

Page Entity

---

From Wikipedia, the free encyclopedia

Discography

Top Section

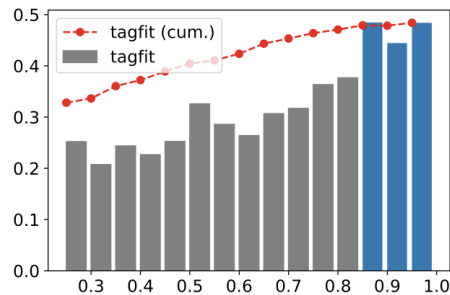
with Guns N' Roses

# Beyond List Pages

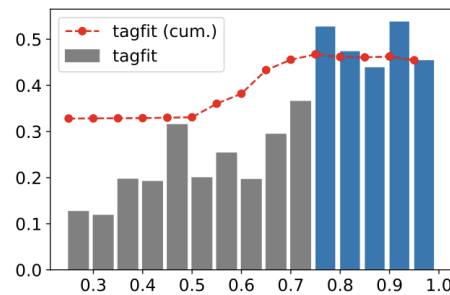
- Learning descriptive rules for listings, e.g.
  - `topSection("Discography") → artist.{>PageEntity<}`
  - Learning across pages to mitigate small data problems
- Metrics:
  - Support: no. of listings covered by rule antecedent
  - Confidence: frequency of rule consequent over all covered listings
  - Consistency: mean absolute deviation of overall confidence and listing confidence
    - i.e., does the rule work equally well across all covered listings

# Beyond List Pages

- Entity detection:
  - Specialize SpaCy tagger with entities on Wikipedia list pages
  - Use SpaCy tags for filtering (e.g., PER for Person etc.)
    - Based on majority vote per class
  - tag fit (i.e., proportion of “fitting” tags for class axioms) used for thresholding



(a) Type confidence

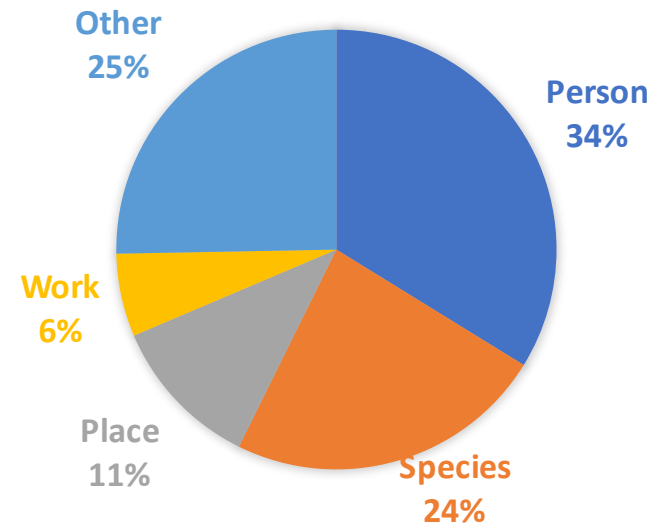


(b) Type consistency

# Beyond List Pages

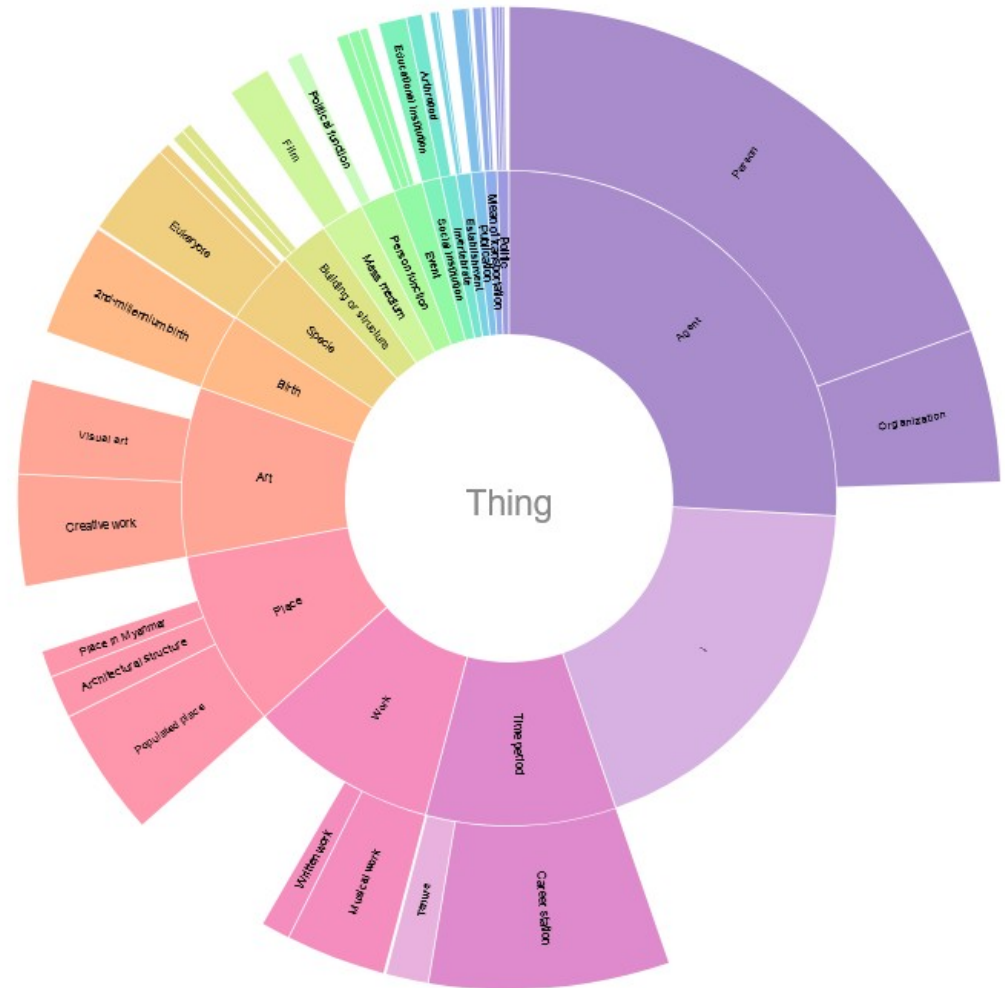
- We can learn
  - ~5M rules for types
  - ~3k rules for relations
- Identify ~2M new entities
  - incl. type and relations within KG
- Post hoc inspection of axioms:
  - Accuracy >90%

Assertion Type	Raw	Filtered
Types (DBpedia)	11,459,047	7,721,039
Types (CaLiGraph)	47,249,624	29,128,677
Relations	732,820	542,018
Relations (via CaLiGraph)	1,381,075	796,910



# CaLiGraph at a Glance

- Latest version 2.1
  - 15M entities
    - incl. 8M from listings
  - Caveat:
    - disambiguation!



# Entity Disambiguation

- Examples: Wikipedia pages of *Die Krupps* and *Eisbrecher*

## Members [\[ edit \]](#)

- Jürgen Engler – vocals, guitar, keyboards, synthesizers and programming, metallic percussion (1980–1985, 1989–1997, 2005–present)
- Ralf Dörper – keyboards, synthesizers and programming (1980–1982, 1985, 1989–1997, 2005–present)
- Marcel Zürcher – guitar, keyboards (2005–present)
- Nils Finkaisen – guitar (2015–present)
- Paul Keller – drums (2018–present)

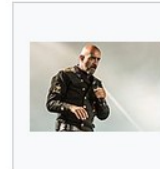
## Former members [\[ edit \]](#)

- Bradley Bills – live drums (2013–2014)
- Rüdiger Esch – bass guitar (1989–1997, 2005–2011)
- Christoph "Nook" Michelfeit – drums, electronic percussion
- Bernward Malaka – bass guitar (1980–1982)
- Hendrik Thiesbrummel – live drums (2016–2018)
- Frank Köllges – drums
- Eva Gossling – saxophone (1981)
- Tina Schnekenburger – syncussion, bass
- Walter Jäger – ?
- Christopher Lietz – programming, samples (1995–1997)
- Lee Altus – guitar (1992–1997)
- Darren Minter – drums (1993)
- George Lewis – drums (1997)
- Oliver Röhl – drums
- Achim Färber – drums
- Volker Borchert – drums (1992, 2015–2016)

## Members [\[ edit \]](#)

- Alexx Wesselsky – vocals (2003–present)
- Noel Pix – lead guitar, programming, production (2003–present)
- Jürgen Plangger – rhythm guitar (2007–present)
- Maximilian Schauer – keyboards, programming (live and session: 2003–2007, session only: 2008–present)
- Achim Färber – drums (2011–present)
- Rupert Keplinger – bass (2013–present)

Eisbrecher, line-up at Rockharz Open Air 2018



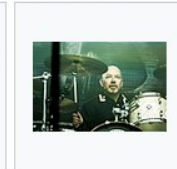
Alexander "Alexx"  
Wesselsky



Jochen "Noel Pix"  
Seibert



Jürgen Plangger



Achim Färber



Rupert Keplinger

## Former live members [\[ edit \]](#)

- Felix Primc – rhythm guitar (2003–2007)
- Micheal Behnke – bass (2003–2007)
- Martin Motnik – bass (2007–2008)
- Olli Pohl – bass (2008–2010, 2015)
- Dominik Palmer – bass (2010–2013)
- Rene Greil – drums (2003–2011)

## Touring members [\[ edit \]](#)

- Sebastien Angrand – drums (2010)

# CaLiGraph Glitches



Formats ▾

[Sparql Endpoint](#)

## About: [clgr:Mannheim](#)

Property	Value
<a href="#">rdfs:label</a>	<ul style="list-style-type: none"><li>Mannheim</li></ul>
<a href="#">clgo:country</a>	<ul style="list-style-type: none"><li>Moldova</li><li>Germany</li></ul>
<a href="#">rdf:type</a>	<ul style="list-style-type: none"><li>Planned capital</li><li>City in Baden-Württemberg</li><li>Twin town or sister city in Germany</li><li>Coat of arms with the Palatine Lion</li><li>French exonym for German toponyms</li><li>Twin town or sister city in Lithuania</li><li>University town in Germany</li><li>owl:NamedIndividual</li><li>City or town in Germany</li><li>Most polluted city in the world</li></ul>

### List of twin towns and sister cities in Moldova

From Wikipedia, the free encyclopedia

This is a list of places in **Moldova** having standing links to local communities in other countries. In most cases, the association, especially when formalised by local government, is known as "town twinning" (though other terms, such as "partner towns" or "sister cities" are sometimes used instead), and while most of the places are towns, the list also comprises villages, cities, districts, counties, etc. with similar links.

Index: [A](#) · [B](#) · [C](#) · [D](#) · [E](#) · [F](#) · [G](#) · [H](#) · [I](#) · [J](#) · [K](#) · [L](#) · [M](#) · [N](#) · [O](#) · [P](#) · [Q](#) · [R](#) · [S](#) · [T](#) · [U](#) · [V](#) · [W](#) · [X](#) · [Y](#) · [Z](#) · [References](#)

**C** [\[ edit \]](#)

#### Chişinău<sup>[3][4]</sup>

- Alba Iulia, Romania
- Ankara, Turkey
- Bucharest, Romania
- Chernivtsi, Ukraine
- Grenoble, France
- Iaşi, Romania

#### Comrat

- Bălţi, Moldova<sup>[1]</sup>

#### Criuleni

- Jurbarkas, Lithuania<sup>[5]</sup>

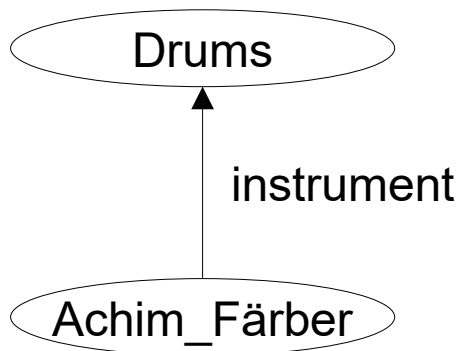
- Kiev, Ukraine
- Mannheim, Germany
- Minsk, Belarus
- Odessa, Ukraine
- Reggio Emilia, Italy
- Sacramento, United States

- Tbilisi, Georgia
- Tel Aviv, Israel
- Vilnius, Lithuania
- Yerevan, Armenia



# CaLiGraph Challenges & Open Issues

- Entity disambiguation
- Usage of formal ontologies (e.g., *country* is functional)
- Extracting information directly from the context



## Members [\[ edit \]](#)

- Jürgen Engler – vocals, guitar, keyboards, [synthesizers](#) and [programming](#), metallic [percussion](#) (1980–1985, 1989–1997, 2005–present)
- Ralf Dörper – keyboards, synthesizers and programming (1980–1982, 1985, 1989–1997, 2005–present)
- Marcel Zürcher – guitar, keyboards (2005–present)
- Nils Finkeisen – guitar (2015–present)
- Paul Keller – drums (2018–present)

## Former members [\[ edit \]](#)

- Bradley Bills – live drums (2013–2014)
- [Rüdiger Esch](#) – bass guitar (1989–1997, 2005–2011)
- Christoph "Nook" Michelfeit – drums, electronic percussion
- [Bernward Malaka](#) – bass guitar (1980–1982)
- Hendrik Thiesbrummel – live drums (2016–2018)
- Frank Köllges – drums
- Eva Gosling – saxophone (1981)
- Tina Schnekenburger – syncussion, bass
- Walter Jäger – ?
- Christopher Lietz – programming, samples (1995–1997)
- [Lee Altus](#) – guitar (1992–1997)
- Darren Minter – drums (1993)
- George Lewis – drums (1997)
- Oliver Röhl – drums
- [Achim Färber](#) – drums
- Volker Borchert – drums (1992, 2015–2016)

# Knowledge Graph Creation Beyond Wikipedia



# A Bird's Eye View on DBpedia EF

- DBpedia Extraction Framework
- Input:
  - A Wikipedia Dump
  - (+ mappings)
- Output:
  - DBpedia



**DBpedia  
Extraction  
Framework**



# A Satellite View on DBpedia EF

- DBpedia Extraction Framework
- Input:
  - A Media Wiki Dump  
(+ mappings)
- Output:
  - A Knowledge Graph



**DBpedia  
Extraction  
Framework**



# What if...?

- What if we went from Wikipedia every MediaWiki?
- According to WikiApiary, there's thousands...

WikiApiary Stats	
Active Sites:	25,327
Semantic Sites:	1,699
Farm Sites:	7,498
Tracked generators:	709
Tracked extensions:	8,691
Tracked skins:	3,347
Registered farms:	218
Active users:	11,097,461
Pages:	824,500,467
Total edits:	5,559,750,218



# Why?

- More is better (maybe)



Pages to date  
**349,05M**

April ↑ 0.80 % month over month



**333,65M** ↑ 11.28 % year over year

12 month average (Apr 2018 - Apr 2019)

## WikiApiary Stats

Active Sites:	25,327
Semantic Sites:	1,699
Farm Sites:	7,498
Tracked generators:	709
Tracked extensions:	8,691
Tracked skins:	3,347
Registered farms:	218
Active users:	11,097,461
Pages:	824,500,467
Total edits:	5,559,750,218

# Why?

- Overcoming Wikipedia's coverage bias

### Notability

### Subject-specific guidelines

- Academics ·
- Astronomical objects
- Books · Events
- Films · Geographic features
- Music · Numbers
- Organizations and companies
- People · Sports and athletes
- Web content

### See also

- Wikipedia essays
- Guide to deletion
- Common deletion outcomes
- Why was my article deleted?

V · T · E



- Main page
- Recent changes
- Statistics
- Random page
- FAQ
- deutsch
- français
- Nederlands
- svenska
- Tools
- What links here
- Related changes
- Special pages

POPULAR PAGES · COMMUNITY · EXPLORE


in: [HasTwitter](#), [HasMySpace](#), [HasFacebook](#), and 2 more

## Speedy Deletion Wikia Main Page

[EDIT](#) [COMMENTS](#) [SHARE](#)

The purpose of this wiki is the same that of Wikipedia, it is to create an encyclopedia which is a comprehensive summary of information from all branches of knowledge.

The only difference between the Wikipedia and this wiki is that we do not have the same criteria for deletion. We want to cover all the artists, actors, athletes and companies that Wikipedia does not want to document. So together with Wikipedia we will have comprehensive knowledge, because Wikipedia deletes so much.



<http://speedydeletion.wikia.com>

Recent Wiki Activity

- [Amar'e Stoudemire](#) Frank Ntilikina · 1 minute ago
- [Carmelo Anthony](#) Frank Ntilikina · 7 minutes ago

### Main Page

**Deletionpedia** is a **radical inclusionist** wiki for rescuing articles from Wikipedia's **deletionism**. It was started by [Guaka](#) on December 24th 2013 and so far we've rescued **53,934 articles**.

As of July 21st 2015 there are also versions in other languages: [French](#), [Dutch](#), [German](#) and [Swedish](#).

Some bot code is available at [GitHub](#).

You should be able to actually sign up and edit articles.

### How does this work?

Articles that are under discussion on Wikipedia are automatically copied here by [Robyt](#). If the article is retained on Wikipedia the article is emptied on Deletionpedia. If the article is removed on Wikipedia we don't have to do anything here. So if an article is not deleted we won't delete the article here, [Robyt](#) will just put a template linking back to Wikipedia. But articles are often relisted for deletion again soon.

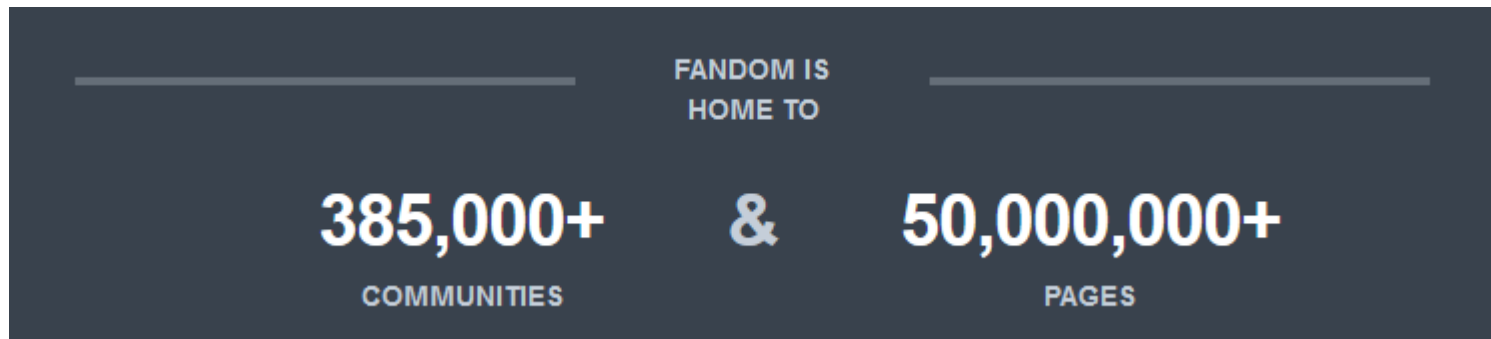
You are welcome to [sign up](#) and help with the project.

It's okay (encouraged even) to edit articles here once they have definitely been deleted on Wikipedia (unlike DPv1), but it's advisable to wait until articles have been definitely deleted on Wikipedia. Like that you can still go into the history of an article and find your edits.

Click on [random page](#) to get an idea about what kind of stuff gets deleted on Wikipedia. A lot is great quality articles written by people who care and spent a lot of time on them, including research and adding references.

# A Brief History of DBkWik

- Started as a student project in 2017
- Task: run DBpedia EF on a large Wiki Farm
  - ...and see what happens





# DBkWik vs. DBpedia

- Challenges
  - Getting dumps: only a fraction of Fandom Wikis has dumps
  - Downloadable from Fandom: 12,840 dumps
  - Tried: auto-requesting dumps

## Database dumps

Database dumps can be used as a personal backup (FANDOM produces separate backups of all wikis automatically) or for maintenance bots

Current pages

(This version is usually best for bot use)

2017-04-05 04:09:48

Current pages and history

(Warning: this file may be very large)

2017-04-05 04:09:48

Request an update

(Dumps are usually generated weekly)



Please [see](#) for more info

„Hallo JPwiki123,

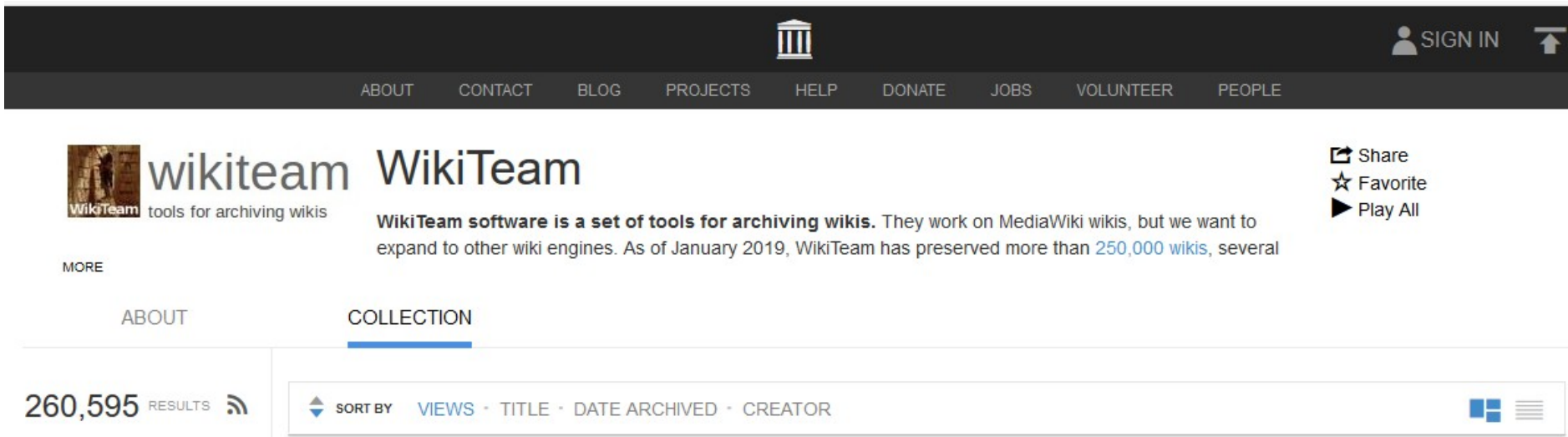
*nur Admins des entsprechenden Wikis können nun neue Datenbank-Dumps anfordern, da ein Missbrauch dieser Funktion zu einer Serverüberlastung auf unserer Seite führen kann. Wenn du auf dem Wiki, für das du einen Dump brauchst, selbst kein Admin bist, kannst du entweder jemanden aus dem Admin-Team darum bitten oder du kannst uns wissen lassen, um welches Wiki es sich handelt. Dann fordern wir gerne einen für dich an!*

Viele Grüße

E [REDACTED]  
Community Support Manager“

# Obtaining Dumps

- We had to change our strategy: WikiTeam software
  - Produces dumps by crawling Wikis
  - Fandom has not blocked us so far :-)
  - Current collection: >300k Wikis
    - will go into DBkWik 1.2 release



The screenshot shows the WikiTeam website interface. At the top, there is a dark navigation bar with a logo and links for ABOUT, CONTACT, BLOG, PROJECTS, HELP, DONATE, JOBS, VOLUNTEER, and PEOPLE. A 'SIGN IN' button is also visible. Below the navigation bar, the main content area features the WikiTeam logo and a description: 'WikiTeam software is a set of tools for archiving wikis. They work on MediaWiki wikis, but we want to expand to other wiki engines. As of January 2019, WikiTeam has preserved more than 250,000 wikis, several'. To the right of the description are social sharing options: Share, Favorite, and Play All. Below the description, there are tabs for 'ABOUT' and 'COLLECTION'. The 'COLLECTION' tab is active, showing a search results page with '260,595 RESULTS' and a sorting menu with options: SORT BY VIEWS · TITLE · DATE ARCHIVED · CREATOR. A blue square icon and a hamburger menu icon are also present in the bottom right of the search results area.

# DBkWik vs. DBpedia

- Mappings do not exist
  - no central ontology
  - i.e., only raw extraction possible
- Duplicates exist
  - origin: pages about the same entity in different Wikis
  - unlike Wikipedia: often not explicitly linked
- Different configurations of MediaWiki



WikiApiary Stats	
Active Sites:	25,327
Semantic Sites:	1,699
Farm Sites:	7,498
Tracked generators:	709
Tracked extensions:	8,691
Tracked skins:	3,347
Registered farms:	218
Active users:	11,097,461
Pages:	824,500,467
Total edits:	5,559,750,218

# Absence of Mappings and Ontology

- Every infobox becomes a class:

```
{infobox actor  
→ mywiki:actor a owl:Class
```

- Every infobox key becomes a property

```
|role = Harry's mother  
→ mywiki:role a rdf:Property
```

- The resulting ontology is very shallow
  - No class hierarchy
  - No distinction of object and data properties
  - No domains and ranges



# Duplicates

- Collecting Data from a Multitude of Wikis

**Trent Reznor**



**Instruments:** Vocals, Guitar, Keyboards, Bass, Marimba, Saxophone, Small Percussion

**Years:** 1988–present

**Tours:** VIVIsectVI–present

**Trent Reznor**




**1 Nomination / 1 Win**

**Role** Composer

**Born** May 17, 1965  
Mercer, Pennsylvania, USA

**Trent Reznor**



**Born**  
May 17, 1965  
New Castle, Pennsylvania, United States

**Other David Lynch Projects**  
*Lost Highway* (Soundtrack - "Videodrones; Questions," "Driver Down")  
"Came Back Haunted" (Music video)

# Representational Variety

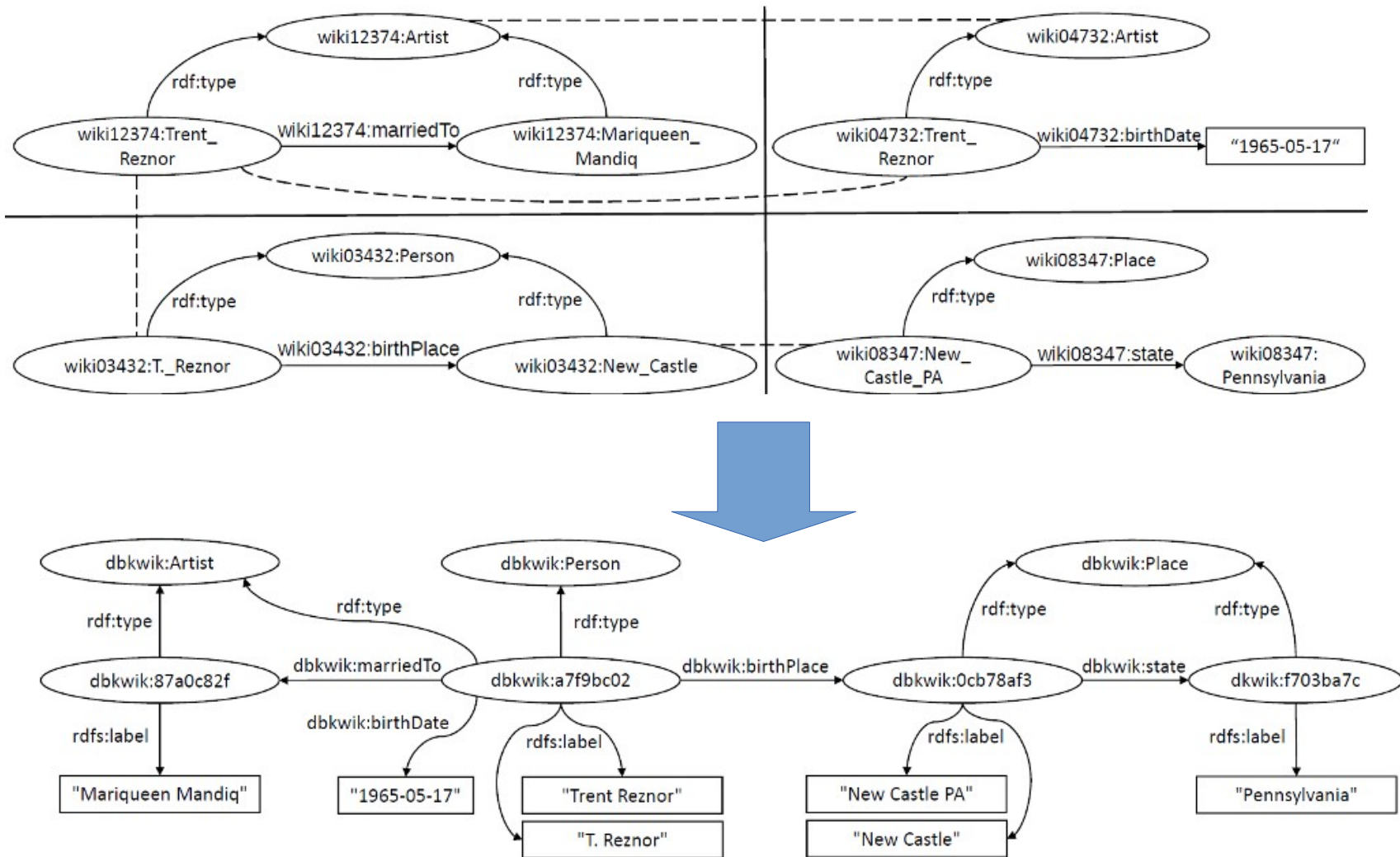
- No conventions across Wikis (besides using MediaWiki syntax)

```
{Person
|name = Trent Reznor
|image = TrentReznor.jpg
|caption - Reznor at the [[83rd Academy Awards]]
|nominations = 1
|wins = 1
|role = Composer
|birthdate = May 17, 1965
|birthloc = Mercer, Pennsylvania
```

```
{{Infobox musician
| Name = Trent Reznor
| Birth_name = Michael Trent Reznor
| Born = May 17, [[1965]] (age 53)
| Origin = [[Mercer]],
[[Pennsylvania]], [[United States]]
```

```
{{Infobox cast
|Name=Trent Reznor
|Image=
|ImageCaption=
|character=
|crew=
|Born={{d|May|17|1965}} New Castle,
Pennsylvania, United States
...
}
```

# Data Fusion



# Naive Data Fusion and Linking to DBpedia

- String similarity for schema matching (classes/properties)
- doc2vec similarity on original pages for instance matching

F1 score...	Internal Linking	Linking to DBpedia
Classes	.979	.898
Properties	.836	.865
Instances	.879	.657

- Results
  - Classes and properties work OK
  - Instances are trickier
  - Internal linking seems easier

maybe...



# Improving Linking and Fusion

- Started a new track at OAEI in 2018
  - annual benchmark for matching tools
- From 2019, some tools starting beating the baseline
  - albeit by a small margin only

System	Time	#testcases	class				property				instance				overall			
			Size	Prec.	F-m.	Rec.	Size	Prec.	F-m.	Rec.	Size	Prec.	F-m.	Rec.	Size	Prec.	F-m.	Rec.
AGM	10:47:38	5	14.6	0.23 (0.23)	0.09 (0.09)	0.06 (0.06)	49.4	0.66 (0.66)	0.32 (0.32)	0.21 (0.21)	5169.0	0.48 (0.48)	0.25 (0.25)	0.17 (0.17)	5233.2	0.48 (0.48)	0.25 (0.25)	0.17 (0.17)
AML	0:45:46	4	27.5	0.78 (0.98)	0.69 (0.86)	0.61 (0.77)	58.2	0.72 (0.91)	0.59 (0.73)	0.49 (0.62)	7529.8	0.72 (0.90)	0.71 (0.88)	0.69 (0.86)	7615.5	0.72 (0.90)	0.70 (0.88)	0.69 (0.86)
baselineAltLabel	0:11:48	5	16.4	1.00 (1.00)	0.74 (0.74)	0.59 (0.59)	47.8	0.99 (0.99)	0.79 (0.79)	0.66 (0.66)	4674.2	0.89 (0.89)	0.84 (0.84)	0.80 (0.80)	4739.0	0.89 (0.89)	0.84 (0.84)	0.80 (0.80)
baselineLabel	0:12:30	5	16.4	1.00 (1.00)	0.74 (0.74)	0.59 (0.59)	47.8	0.99 (0.99)	0.79 (0.79)	0.66 (0.66)	3641.2	0.95 (0.95)	0.81 (0.81)	0.71 (0.71)	3706.0	0.95 (0.95)	0.81 (0.81)	0.71 (0.71)
DOME	1:05:26	4	22.5	0.74 (0.92)	0.62 (0.77)	0.53 (0.66)	75.5	0.79 (0.99)	0.77 (0.96)	0.75 (0.93)	4895.2	0.74 (0.92)	0.70 (0.88)	0.67 (0.84)	4994.8	0.74 (0.92)	0.70 (0.88)	0.67 (0.84)
FCAMap-KG	1:14:49	5	18.6	1.00 (1.00)	0.82 (0.82)	0.70 (0.70)	69.0	1.00 (1.00)	0.98 (0.98)	0.96 (0.96)	4530.6	0.90 (0.90)	0.84 (0.84)	0.79 (0.79)	4792.6	0.91 (0.91)	0.85 (0.85)	0.79 (0.79)
LogMap	0:15:43	5	26.0	0.95 (0.95)	0.84 (0.84)	0.76 (0.76)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	26.0	0.95 (0.95)	0.01 (0.01)	0.00 (0.00)
LogMapBio	2:31:01	5	26.0	0.95 (0.95)	0.84 (0.84)	0.76 (0.76)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	26.0	0.95 (0.95)	0.01 (0.01)	0.00 (0.00)
LogMapKG	2:26:14	5	26.0	0.95 (0.95)	0.84 (0.84)	0.76 (0.76)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	29190.4	0.40 (0.40)	0.54 (0.54)	0.86 (0.86)	29216.4	0.40 (0.40)	0.54 (0.54)	0.84 (0.84)
LogMapLt	0:07:28	4	23.0	0.80 (1.00)	0.56 (0.70)	0.43 (0.54)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	6653.8	0.73 (0.91)	0.67 (0.84)	0.62 (0.78)	6676.8	0.73 (0.91)	0.66 (0.83)	0.61 (0.76)
POMAP++	0:14:39	5	2.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	19.4	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Wiktionary	0:20:14	5	21.4	1.00 (1.00)	0.80 (0.80)	0.67 (0.67)	75.8	0.97 (0.97)	0.98 (0.98)	0.98 (0.98)	3483.6	0.91 (0.91)	0.79 (0.79)	0.70 (0.70)	3581.8	0.91 (0.91)	0.80 (0.80)	0.71 (0.71)

# The Golden Hammer Bias

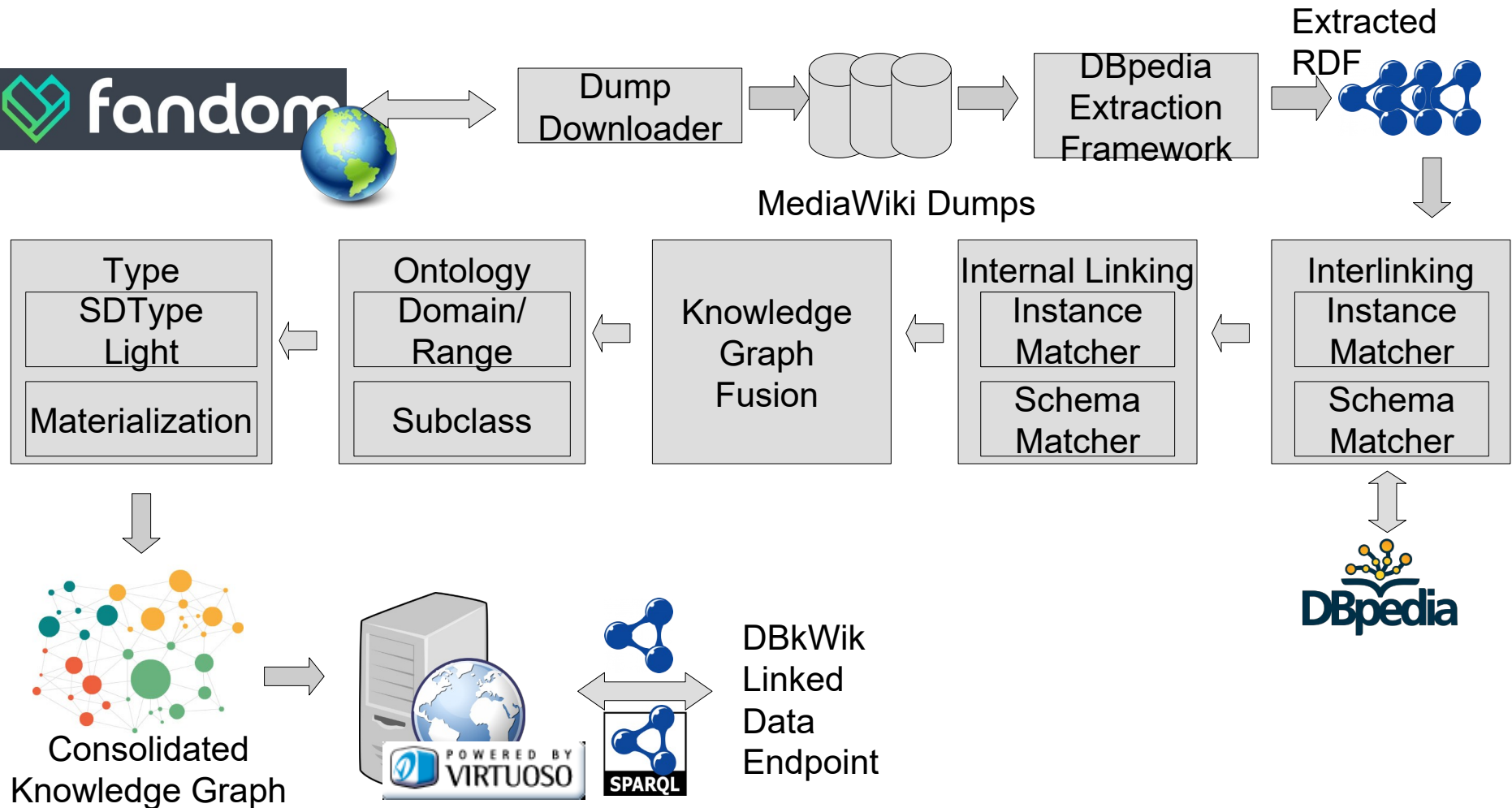
- Challenge:
  - OAEI tools expect two **related** KGs
  - but: 300k KGs can only be matched without manual pre-inspection



Matcher	mcu lyrics		memoryalpha lyrics		starwars lyrics	
	matches	precision	matches	precision	matches	precision
AML	2,642	0.12	7,691	0.00	3,417	0.00
baselineAltLabel	588	0.44	1,332	0.02	1,582	0.04
baselineLabel	513	0.54	1,006	0.06	1,141	0.06
FCAMap-KG	755	0.40	2,039	0.14	2,520	0.02
LogMapKG	29,238	0.02	-	-	-	-
LogMapLt	2,407	0.08	7,199	0.00	2,728	0.04
Wiktionary	971	0.12	3,457	0.02	4,026	0.00

See: ESWC 2020 Paper on OAEI Knowledge Graph Track

# Big Picture



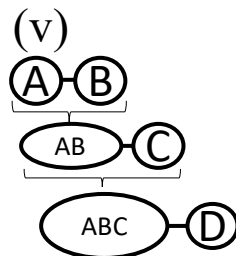
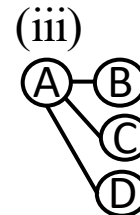
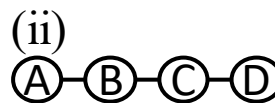
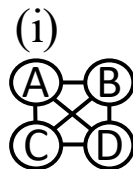
# DBkWik 1.1

- Source: ~15k Wiki dumps from Fandom
  - 52.4GB of data (roughly the size of the English Wikipedia)

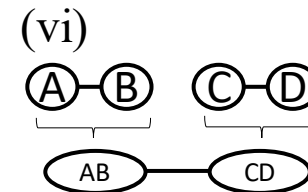
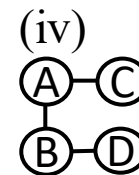
	Raw	Final
Instances	14,212,535	11,163,719
Typed instances	1,880,189	1,372,971
Triples	107,833,322	91,526,001
Avg. indegree	0.624	0.703
Avg. outdegree	7.506	8.169
Classes	71,580	12,029
Properties	506,487	128,566

# Towards DBkWik 1.2

- Again, we have an entity resolution problem
  - with entities from 300k sources
- Strategies
  - (i) pairwise ( $O(n^2)$ )
  - (ii-iv) transitive pairs
  - (iii-v) incremental merge
- Ordering by
  - smallest/largest first
  - source similarity



Order based



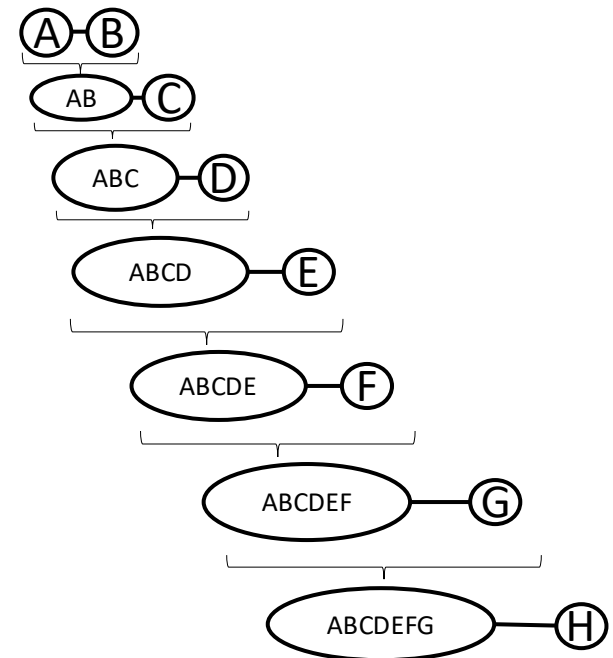
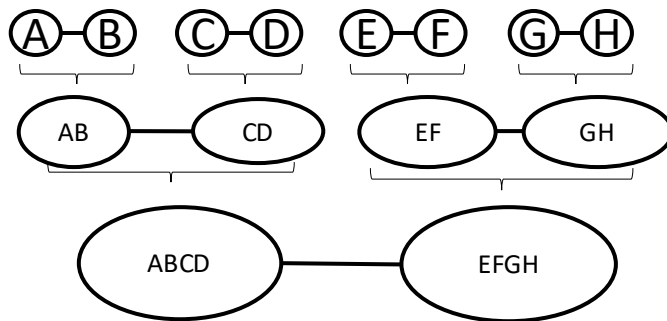
Similarity based

Transitive Pairs

Incremental Merge

# Towards DBkWik 1.2

- Preliminary results
  - Incremental merge works best (quality close to pairwise)
- Main challenge: runtime
  - One match&merge step takes ~20 minutes
  - i.e., 300k steps are more than 11 years!
- Idea: parallelization
  - best case: tree height of 18
  - best runtime (fully parallel): six hours

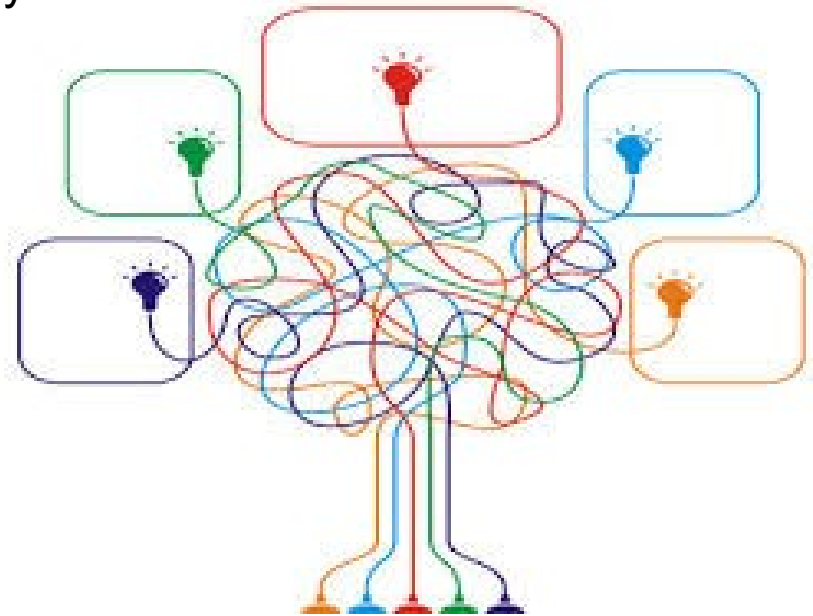


# Summary

- DBpedia and YAGO
  - one source (Wikipedia, multiple languages)
  - one entity per page paradigm
- CaLiGraph
  - one source (Wikipedia, multiple languages would be possible)
  - extraction from list-like constructs
    - further possible extension: list-like constructs *outside* of Wikipedia
  - current open challenge: **entity resolution**
- DBkWik
  - extraction from thousands of Wikis
  - current open challenge: **entity resolution**
    - in particular: scalability!

# Further Open Challenges

- More detailed profiling of knowledge graphs
  - e.g., do we reduce or increase bias?
  - and: is that good or bad?
- Task-based downstream evaluations
  - Does it improve, e.g., recommender systems?
- Fusion policies
  - schema level,  
e.g., many shallow ontologies  
→ one deep ontology?
  - instance level,  
e.g., identify outdated information





# Contributors

- Contributors (past&present)



Sven Hertling



Alexandra  
Hofmann



Samresh  
Perchani

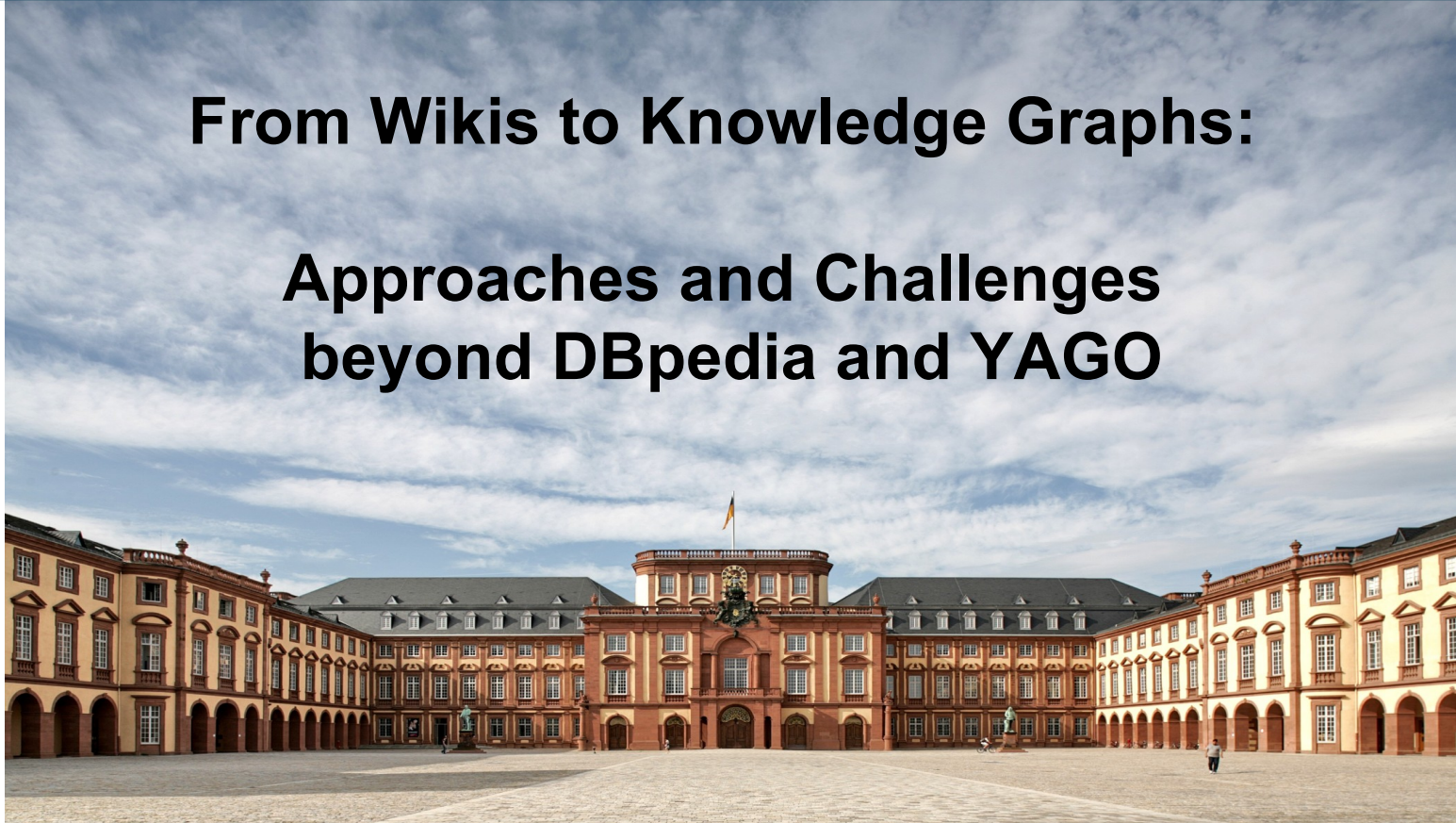


Jan Portisch



Nicolas  
Heist

**From Wikis to Knowledge Graphs:  
Approaches and Challenges  
beyond DBpedia and YAGO**



**Heiko Paulheim**