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

Heiko Paulheim
A Brief History of Knowledge Graphs

- Google’s Announcement
- DBpedia
- YAGO
- NELL
- ResearchCyc
- Freebase
- Wikidata
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
Wikipedia as a Knowledge Graph
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/

Why bother?

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

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

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

See: ISWC 2019 Paper on Uncovering the Semantics of Wikipedia Categories
Cat2Ax: Axiomatizing Wikipedia Categories

\[ \subseteq \text{dbo:genre.}\{\text{dbr:Rock\_Music}\} ? \]
\[ \subseteq \text{dbo:artist.}\{\text{dbr:Rock\_\(\text{(Rapper)}\)}\} ? \]
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 isRockused 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**

<table>
<thead>
<tr>
<th>Approach</th>
<th>Count</th>
<th>Precision [%]</th>
<th>Count</th>
<th>Precision [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Relation axioms</td>
<td></td>
<td>Type axioms</td>
<td></td>
</tr>
<tr>
<td>Cat2Ax</td>
<td>272,707</td>
<td>95.6</td>
<td>430,405</td>
<td>96.8</td>
</tr>
<tr>
<td>C-DF</td>
<td>143,850</td>
<td>83.6</td>
<td>28,247</td>
<td>92.0</td>
</tr>
<tr>
<td>Catriple</td>
<td>306,177</td>
<td>87.2</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Relation assertions</th>
<th></th>
<th>Type assertions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat2Ax</td>
<td>4,424,785 (7,554,980)</td>
<td>87.2 (92.1)</td>
<td>3,342,057 (12,111,194)</td>
<td>90.8 (95.7)</td>
</tr>
<tr>
<td>C-DF</td>
<td>766,921 (2,856,592)</td>
<td>78.4 (93.4)</td>
<td>198,485 (2,352,474)</td>
<td>76.8 (97.1)</td>
</tr>
<tr>
<td>Catriple</td>
<td>6,260,972 (6,836,924)</td>
<td>74.4 (76.5)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>• Tiamat</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>----------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>owl:sameAs</td>
<td>• dbr:Tiamat_(band)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dgo:activeYearsStartYear</td>
<td>• 1987</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dgo:genre</td>
<td>• Symphonic metal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dgo:hometown</td>
<td>• Sweden</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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

This section does not cite any sources. Please help improve this section by adding citations to reliable sources. Unsourced material may be challenged and removed.

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

<table>
<thead>
<tr>
<th>A-K</th>
<th>[ edit ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actress</td>
<td>Acoustic</td>
</tr>
<tr>
<td>Air Liquide</td>
<td>Alarm Will Sound</td>
</tr>
<tr>
<td>Alva Noto</td>
<td>Amon Tobin</td>
</tr>
<tr>
<td>Andy Stott</td>
<td>Aphet Twin</td>
</tr>
<tr>
<td>Apparat</td>
<td>Arovane</td>
</tr>
<tr>
<td>Atjpek</td>
<td>Autotechnique</td>
</tr>
<tr>
<td>B12</td>
<td>Benn Jordan</td>
</tr>
<tr>
<td>Biosphere</td>
<td>Bjork</td>
</tr>
<tr>
<td>Bjork</td>
<td>The Black Dog</td>
</tr>
<tr>
<td>Blanc Mass</td>
<td>Boards of Canada</td>
</tr>
<tr>
<td>Bochum Wetz</td>
<td>Boom Bip</td>
</tr>
<tr>
<td>Bonobo</td>
<td>Brothomstates</td>
</tr>
<tr>
<td>Burial</td>
<td>Budu</td>
</tr>
<tr>
<td>Budu</td>
<td>C418</td>
</tr>
<tr>
<td>Cabaret Voltaire</td>
<td>Casino Versus Japan</td>
</tr>
<tr>
<td>Chris Clark</td>
<td>Chris Eaton</td>
</tr>
<tr>
<td>Ciucul</td>
<td>Czajka</td>
</tr>
<tr>
<td>Daedelus</td>
<td>Deathbed</td>
</tr>
<tr>
<td>Deepchord</td>
<td>Demidevices</td>
</tr>
<tr>
<td>Deru</td>
<td>Richard Devine</td>
</tr>
<tr>
<td>Dopplegangerfield</td>
<td>Dreamers</td>
</tr>
<tr>
<td>Die Wool</td>
<td>Diverse</td>
</tr>
<tr>
<td>Eight Frozen Modules</td>
<td>Emptyscript</td>
</tr>
<tr>
<td>Esben</td>
<td>Felt</td>
</tr>
<tr>
<td>Fennesz</td>
<td>The Field</td>
</tr>
<tr>
<td>Fish</td>
<td>The Flashbulb</td>
</tr>
<tr>
<td>Floating Points</td>
<td>Flying Lotus</td>
</tr>
<tr>
<td>Forest Swords</td>
<td>Funktionsgruppe</td>
</tr>
<tr>
<td>Future Sound Of London</td>
<td>Gas</td>
</tr>
<tr>
<td>Giscom</td>
<td>Global Communication</td>
</tr>
<tr>
<td>Global Groov</td>
<td>Goldie</td>
</tr>
<tr>
<td>Goldie</td>
<td>Zachary Gray</td>
</tr>
<tr>
<td>Gridlock</td>
<td>Himuro Yoshiteru</td>
</tr>
<tr>
<td>Kim Horath</td>
<td>I Am Robot and Proud</td>
</tr>
<tr>
<td>Innovaders</td>
<td>Jan Jelinek</td>
</tr>
<tr>
<td>Jega</td>
<td>Jeiko</td>
</tr>
<tr>
<td>Jin</td>
<td>John Tejada</td>
</tr>
<tr>
<td>Jon Hopkins</td>
<td>Kettler</td>
</tr>
<tr>
<td>Kevin Blechdom</td>
<td>Kidospace</td>
</tr>
<tr>
<td>Kodomo</td>
<td>Kromeiss</td>
</tr>
</tbody>
</table>
Pushing Entity Coverage Further

• Beyond red links (2020)

<table>
<thead>
<tr>
<th>Title</th>
<th>Running time</th>
<th>Year released</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amra Ekla Cinema Banaba (The Innocence)</td>
<td>1205 min (21 hr, 5 min)</td>
<td>2019</td>
<td></td>
</tr>
<tr>
<td>Rosan (The Journey)</td>
<td>875 min (14 hr, 33 min)</td>
<td>1967</td>
<td></td>
</tr>
<tr>
<td>La Fila</td>
<td>803 min (13 hr, 23 min)</td>
<td>2016</td>
<td></td>
</tr>
<tr>
<td>Out 1 (Not me tangere)</td>
<td>775 min (12 hr, 55 min)</td>
<td>1971</td>
<td></td>
</tr>
<tr>
<td>Evolution of a Filipino Family</td>
<td>593 min (9 hr, 53 min)</td>
<td>2004</td>
<td></td>
</tr>
<tr>
<td>Shoah</td>
<td>660 min (9 hr, 28 min)</td>
<td>1885</td>
<td></td>
</tr>
<tr>
<td>Tie Xi Qu: West of the Tracks</td>
<td>551 min (9 hr, 11 min)</td>
<td>2003</td>
<td></td>
</tr>
<tr>
<td>Death in the Land of Encantos</td>
<td>538 min (8 hr, 58 min)</td>
<td>2007</td>
<td></td>
</tr>
<tr>
<td>Dred Souls</td>
<td>490 min (8 hr, 15 min)</td>
<td>2016</td>
<td></td>
</tr>
<tr>
<td>A Lullaby to the Sorrowful Mystery</td>
<td>485 min (8 hr, 5 min)</td>
<td>2016</td>
<td></td>
</tr>
<tr>
<td>O.J. Made in America</td>
<td>463 min (7 hr, 43 min)</td>
<td>2016</td>
<td></td>
</tr>
<tr>
<td>Melancholia</td>
<td>420 min (7 hr, 30 min)</td>
<td>2006</td>
<td></td>
</tr>
<tr>
<td>Såldintangån</td>
<td>419 min (6 hr, 59 min)</td>
<td>1994</td>
<td></td>
</tr>
<tr>
<td>La Roue</td>
<td>419 min (6 hr, 53 min)</td>
<td>1923 (Restoration, 2010)</td>
<td></td>
</tr>
<tr>
<td>The Best of Youth</td>
<td>366 min (6 hr, 6 min)</td>
<td>2003</td>
<td></td>
</tr>
<tr>
<td>Century of Birthning</td>
<td>360 min (6 hr)</td>
<td>2011</td>
<td></td>
</tr>
<tr>
<td>Hear Death</td>
<td>356 min (5 hr, 58 min)</td>
<td>1989</td>
<td></td>
</tr>
<tr>
<td>Karamay</td>
<td>356 min (5 hr, 56 min)</td>
<td>2011</td>
<td></td>
</tr>
<tr>
<td>Little Dorrit</td>
<td>350 min (5 hr, 50 min)</td>
<td>1987</td>
<td></td>
</tr>
<tr>
<td>Cartos</td>
<td>339 min (5 hr, 39 min)</td>
<td>2010</td>
<td></td>
</tr>
<tr>
<td>Mula sa Kung Año ang Noon</td>
<td>330 min (5 hr, 30 min)</td>
<td>2014</td>
<td></td>
</tr>
<tr>
<td>Napoleón</td>
<td>332 min (5 hr, 32 min)</td>
<td>1927 (Restoration, 2016)</td>
<td></td>
</tr>
<tr>
<td>1900</td>
<td>317 min (5 hr, 17 min)</td>
<td>1976</td>
<td></td>
</tr>
<tr>
<td>Happy Hour</td>
<td>317 min (5 hr, 17 min)</td>
<td>2015</td>
<td></td>
</tr>
<tr>
<td>Batang West Side</td>
<td>310 min (5 hr, 15 min)</td>
<td>2001</td>
<td></td>
</tr>
<tr>
<td>The Deluge</td>
<td>315 min (5 hr, 15 min)</td>
<td>1974</td>
<td></td>
</tr>
<tr>
<td>Fanny and Alexander</td>
<td>312 min (5 hr, 12 min)</td>
<td>1982</td>
<td></td>
</tr>
<tr>
<td>Tschech</td>
<td>304 min (5 hr, 4 min)</td>
<td>1994</td>
<td></td>
</tr>
</tbody>
</table>

• Beyond explicit lists (2021)

Members [edit]
- Marcel Zürcher – guitar, keyboards (2005–present)
- Nils Finklestein – guitar (2015–present)
- Paul Kieler – drums (2015–present)

Former members [edit]
- Christoph “Hook” Michielitz – drums, electronic percussion
- Frank Köului – drums
- Eva Gössling – saxophone (1981)
- Christina Schneekneburger – keyboards
- Walter Jäger – ?
- Darren Minter – drums (1993)
- George Lewis – drums (1997)
- Oliver Röhl – drums
- Achim Farber – drums

Discography [edit]

Albums [edit]
- Stahlwerk/Symphone (1981)
- Vole Kraft Vorant (1982)
- Entering the Arena (1985)
- I (1985)
- II – The Final Option (1993)
- The Final Remakes (1994)
- Paradise Now (1997)
- The Machinists of Joy (2013)
- Stahlwerk/Requiem (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
Entity Extraction from List Pages

• Lists form (shallow) hierarchies
Entity Extraction from List Pages

• Idea: align with category graph
• Equivalence:
  – “List of Japanese Writers”
    \[\leftrightarrow\] Category:Japanese Writers
• Subsumption:
  – “List of Japanese Speculative Fiction Writers”
    \[\rightarrow\] 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
- 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, …

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

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Enum</th>
<th></th>
<th>Table</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>Baseline (pick first entity)</td>
<td>74</td>
<td>96</td>
<td>84</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>80</td>
<td>90</td>
<td>84</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>82</td>
<td>78</td>
<td>80</td>
</tr>
<tr>
<td>Random Forest</td>
<td>85</td>
<td>90</td>
<td>87</td>
</tr>
<tr>
<td>XG-Boost</td>
<td><strong>90</strong></td>
<td>83</td>
<td>86</td>
</tr>
<tr>
<td>Neural Network (MLP)</td>
<td>86</td>
<td>84</td>
<td>85</td>
</tr>
<tr>
<td>SVM</td>
<td>86</td>
<td>60</td>
<td>71</td>
</tr>
</tbody>
</table>

Pie chart showing distribution of entity types: Place (29%), Species (25%), Work (9%), Other (9%), Organisation (9%).
Beyond List Pages

- Many pages contain list-like constructs
  - 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?" (1990)
  - "Chinese Democracy" (2008)

with Hollywood Rose [edit]

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 (1988; "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 Outpatients (1996; "Anxious Disease" ft. Slash)
- New Looney Tunes (2016; "Rock the Rock")

Filmography [edit]

<table>
<thead>
<tr>
<th>Title</th>
<th>Year</th>
<th>Role</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Dead Pool</td>
<td>1988</td>
<td>Musician at funeral</td>
<td></td>
</tr>
<tr>
<td>Grand Theft Auto: San Andreas (video game)</td>
<td>2004</td>
<td>DJ Tommy &quot;The Nightmare&quot; Smith in the K-GST radio</td>
<td>Voice</td>
</tr>
<tr>
<td>That Metal Show</td>
<td>2011</td>
<td>Himself</td>
<td></td>
</tr>
<tr>
<td>Jimmy Kimmel Live!</td>
<td>2012</td>
<td>Himself</td>
<td></td>
</tr>
<tr>
<td>New Looney Tunes (TV show)</td>
<td>2018</td>
<td>Himself</td>
<td>Voice</td>
</tr>
<tr>
<td>Scooby-Doo and Guess Who? (TV Show)</td>
<td>2021</td>
<td>Himself</td>
<td>Voice</td>
</tr>
</tbody>
</table>
Beyond List Pages
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

![Graphs showing type confidence and type consistency](image)
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%

<table>
<thead>
<tr>
<th>Assertion Type</th>
<th>Raw</th>
<th>Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Types (DBpedia)</td>
<td>11,459,047</td>
<td>7,721,039</td>
</tr>
<tr>
<td>Types (CaLiGraph)</td>
<td>47,249,624</td>
<td>29,128,677</td>
</tr>
<tr>
<td>Relations</td>
<td>732,820</td>
<td>542,018</td>
</tr>
<tr>
<td>Relations (via CaLiGraph)</td>
<td>1,381,075</td>
<td>796,910</td>
</tr>
</tbody>
</table>
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

**Die Krupps**
- Marcel Zürcher – guitar, keyboards (2005–present)
- Nils Finkjesen – guitar (2016–present)
- Paul Keller – drums (2018–present)

**Eisbrecher**
- Alex Wesselsky – vocals (2003–present)
- Noel Pix – lead guitar, programming, production (2003–present)
- Jürgen Pfangger – rhythm guitar (2007–present)
- Achim Farber – drums (2011–present)
- Rupert Keplinger – bass (2013–present)

### Former members

- Christoph "Noel" Mischeff – drums, electronic percussion
- Frank Kolges – drums
- Tina Schnekenburger – synccussion, bass
- Walter Jager – ?
- Darren Minter – drums (1993)
- George Lewis – drums (1997)
- Oliver Rohl – drums

- Achim Farber – drums

### Former live members


### Touring members

- Sebastien Angrand – drums (2010)
List of twin towns and sister cities in Moldova

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 here are towns, the list also comprises villages, cities, districts, counties, etc. with similar links.

C [edit]

Mannheim, Germany

- Planned capital
- City in Baden-Württemberg
- Twin town or sister city
- Coat of arms with the Palatine Lion
- French exonym for German toponyms
- Twin town or sister city in Lithuania
- University town in Germany
- owl:NamedIndividual
- City or town in Germany
- Most polluted city in the world
CaLiGraph Challenges & Open Issues

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

---

**Members**
- Marcel Zürcher – guitar, keyboards (2005–present)
- Nils Finkeisen – guitar (2015–present)
- Paul Keller – drums (2018–present)

**Former members**
- Christoph "Nook" Micholleit – drums, electronic percussion
- Frank Köllges – drums
- Tina Schneekenburger – syncussion, bass
- Walter Jäger – ?
- Darren Minter – drums (1993)
- George Lewis – drums (1997)
- Oliver Röhl – drums
- **Achim Färber – drums**
Knowledge Graph Creation Beyond Wikipedia
A Bird’s Eye View on DBpedia EF

- DBpedia Extraction Framework
- Input:
  - A Wikipedia Dump
    (+ mappings)
- Output:
  - DBpedia
A Satellite View on DBpedia EF

• DBpedia Extraction Framework
• Input:
  – A Media Wiki Dump
    (+ mappings)
• Output:
  – A Knowledge Graph
What if…?

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

<table>
<thead>
<tr>
<th>WikiApiary Stats</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Sites:</td>
<td>25,327</td>
</tr>
<tr>
<td>Semantic Sites:</td>
<td>1,699</td>
</tr>
<tr>
<td>Farm Sites:</td>
<td>7,499</td>
</tr>
<tr>
<td>Tracked generators:</td>
<td>709</td>
</tr>
<tr>
<td>Tracked extensions:</td>
<td>8,691</td>
</tr>
<tr>
<td>Tracked skins:</td>
<td>3,347</td>
</tr>
<tr>
<td>Registered farms:</td>
<td>218</td>
</tr>
<tr>
<td>Active users:</td>
<td>11,097,461</td>
</tr>
<tr>
<td>Pages:</td>
<td>824,500,467</td>
</tr>
<tr>
<td>Total edits:</td>
<td>5,559,750,218</td>
</tr>
</tbody>
</table>
Why?

- More is better (maybe)

WikiApiary Stats

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Sites</td>
<td>25,327</td>
</tr>
<tr>
<td>Semantic Sites</td>
<td>1,699</td>
</tr>
<tr>
<td>Farm Sites</td>
<td>7,498</td>
</tr>
<tr>
<td>Tracked generators</td>
<td>709</td>
</tr>
<tr>
<td>Tracked extensions</td>
<td>8,691</td>
</tr>
<tr>
<td>Tracked skins</td>
<td>3,347</td>
</tr>
<tr>
<td>Registered farms</td>
<td>218</td>
</tr>
<tr>
<td>Active users</td>
<td>11,097,461</td>
</tr>
<tr>
<td>Pages</td>
<td>824,500,467</td>
</tr>
<tr>
<td>Total edits</td>
<td>5,559,750,218</td>
</tr>
</tbody>
</table>
Why?

- Overcoming Wikipedia’s coverage bias
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

fandom

POWERED BY wikia

385,000+ COMMUNITIES & 50,000,000+ PAGES
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
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

<table>
<thead>
<tr>
<th>WikiApiary Stats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Sites:</td>
</tr>
<tr>
<td>Semantic Sites:</td>
</tr>
<tr>
<td>Farm Sites:</td>
</tr>
<tr>
<td>Tracked generators:</td>
</tr>
<tr>
<td>Tracked extensions:</td>
</tr>
<tr>
<td>Tracked skins:</td>
</tr>
<tr>
<td>Registered farms:</td>
</tr>
<tr>
<td>Active users:</td>
</tr>
<tr>
<td>Pages:</td>
</tr>
<tr>
<td>Total edits:</td>
</tr>
</tbody>
</table>
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: VIVIsctVI–present
- Role: Composer
- Born: May 17, 1965
- Mercer, Pennsylvania, USA

1 Nomination / 1 Win

- 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  
... 
}
Naive Data Fusion and Linking to DBpedia

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

<table>
<thead>
<tr>
<th>F1 score...</th>
<th>Internal Linking</th>
<th>Linking to DBpedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classes</td>
<td>.979</td>
<td>.898</td>
</tr>
<tr>
<td>Properties</td>
<td>.836</td>
<td>.865</td>
</tr>
<tr>
<td>Instances</td>
<td>.879</td>
<td>.657</td>
</tr>
</tbody>
</table>

- Results
  - Classes and properties work OK
  - Instances are trickier
  - Internal linking seems easier
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
The Golden Hammer Bias

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

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

See: ESWC 2020 Paper on OAEI Knowledge Graph Track
Big Picture

Dump Downloader → DBpedia Extraction Framework

MediaWiki Dumps

Extracted RDF

Consolidated Knowledge Graph

DBkWik Linked Data Endpoint

Type
SDType
Light
Materialization

Ontology
Domain/Range
Subclass

Knowledge Graph Fusion

Internal Linking
Instance Matcher
Schema Matcher

Interlinking
Instance Matcher
Schema Matcher

9/30/21  Heiko Paulheim
DBkWik 1.1

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

<table>
<thead>
<tr>
<th></th>
<th>Raw</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instances</td>
<td>14,212,535</td>
<td>11,163,719</td>
</tr>
<tr>
<td>Typed instances</td>
<td>1,880,189</td>
<td>1,372,971</td>
</tr>
<tr>
<td>Triples</td>
<td>107,833,322</td>
<td>91,526,001</td>
</tr>
<tr>
<td>Avg. indegree</td>
<td>0.624</td>
<td>0.703</td>
</tr>
<tr>
<td>Avg. outdegree</td>
<td>7.506</td>
<td>8.169</td>
</tr>
<tr>
<td>Classes</td>
<td>71,580</td>
<td>12,029</td>
</tr>
<tr>
<td>Properties</td>
<td>506,487</td>
<td>128,566</td>
</tr>
</tbody>
</table>
Towards DBkWik 1.2

• Again, we have an entity resolution problem
  – with entities from 300k sources

• Strategies
  • (i) pairwise (O(n²))
  • (ii-iv) transitive pairs
  • (iii-v) incremental merge

• Ordering by
  • smallest/largest first
  • source similarity
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