Knowledge Graphs
Public Knowledge Graphs
Heiko Paulheim
Previously on “Knowledge Graphs”

• Principles:
  – RDF, RDF-S, SPARQL & co
  – Linked Open Data

• Today
  – A closer look on actually existing knowledge graphs
  – Some useful, large-scale resources
Introduction

- Knowledge Graphs out there (not guaranteed to be complete)

<table>
<thead>
<tr>
<th>Name</th>
<th>Instances</th>
<th>Facts</th>
<th>Types</th>
<th>Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBpedia (English)</td>
<td>4,806,150</td>
<td>176,043,129</td>
<td>735</td>
<td>2,813</td>
</tr>
<tr>
<td>YAGO</td>
<td>4,595,906</td>
<td>25,946,870</td>
<td>488,469</td>
<td>77</td>
</tr>
<tr>
<td>Freebase</td>
<td>49,947,845</td>
<td>3,041,722,635</td>
<td>26,507</td>
<td>37,781</td>
</tr>
<tr>
<td>Wikidata</td>
<td>15,602,060</td>
<td>65,993,797</td>
<td>23,157</td>
<td>1,673</td>
</tr>
<tr>
<td>NELL</td>
<td>2,006,896</td>
<td>432,845</td>
<td>285</td>
<td>425</td>
</tr>
<tr>
<td>OpenCyc</td>
<td>118,499</td>
<td>2,413,894</td>
<td>45,153</td>
<td>18,526</td>
</tr>
<tr>
<td>Google’s Knowledge Graph</td>
<td>570,000,000</td>
<td>18,000,000,000</td>
<td>1,500</td>
<td>35,000</td>
</tr>
<tr>
<td>Google’s Knowledge Vault</td>
<td>45,000,000</td>
<td>271,000,000</td>
<td>1,100</td>
<td>4,469</td>
</tr>
<tr>
<td>Yahoo! Knowledge Graph</td>
<td>3,443,743</td>
<td>1,391,054,990</td>
<td>250</td>
<td>800</td>
</tr>
</tbody>
</table>

Knowledge Graph Creation: CyC

• The beginning
  – Encyclopedic collection of knowledge
  – Started by Douglas Lenat in 1984
  – Estimation: 350 person years and 250,000 rules should do the job
    of collecting the essence of the world’s knowledge

• The present (as of June 2017)
  – ~1,000 person years, $120M total development cost
  – 21M axioms and rules
  – Used to exist until 2017
Knowledge Graph Creation: CyC
Knowledge Graph Creation

• Lesson learned no. 1:
  – Trading efforts against accuracy
Knowledge Graph Creation: Freebase

- The 2000s
  - Freebase: collaborative editing
  - Schema not fixed

- Present
  - Acquired by Google in 2010
  - Powered first version of Google’s Knowledge Graph
  - Shut down in 2016
  - Partly lives on in Wikidata (see in a minute)

coming up soon: was it a good deal or not?
Knowledge Graph Creation: Freebase

- Community based
- Like Wikipedia, but more structured

Arnold Schwarzenegger

- **Types:** Person (People), US Politician (Government), Film actor (Film), Film producer (Film), Pro Athlete (Sports), Sports Award Winner (Sports)
- **Also known as:** Arnold Alois Schwarzenegger, The Governor
- **Gender:** Male
- **Date of Birth:** Jul 30, 1947
- **Place of Birth:** Thai, Austria
- **Country Of Nationality:** United States
- **Profession:** Politician, Bodybuilder, Entrepreneur, Actor
- **Religion:** Roman Catholicism
- **Parents:** Aurelia Jadrny Schwarzenegger, Gustav Schwarzenegger
- **Children:** Christopher Schwarzenegger, Patrick Schwarzenegger, Christina Schwarzenegger, Katherine Schwarzenegger
- **Siblings:** Meinhard Schwarzenegger
- **Spouse (or domestic partner):** Maria Shriver • Apr 26, 1986
- **Height:** 1.88 m
- **IMDB Entry:** http://www.imdb.com/name/nm0000216/
- **Career Start:** 1968
- **Career End:** 1980
Knowledge Graph Creation

• Lesson learned no. 2:
  – Trading formality against number of users
Knowledge Graph Creation: Wikidata

• The 2010s
  – Wikidata: launched 2012
  – Goal: centralize data from Wikipedia languages
  – Collaborative
  – Imports other datasets

• Present
  – One of the largest public knowledge graphs (see later)
  – Includes rich provenance
Knowledge Graph Creation: Wikidata

• Collaborative editing
Knowledge Graph Creation: Wikidata

- Provenance
Wikidata
Knowledge Graph Creation

- Lesson learned no. 3:
  - There is not one truth (but allowing for plurality adds complexity)

Max. simplicity

Max. support for plurality
Knowledge Graph Creation: DBpedia & YAGO

• The 2010s
  – DBpedia: launched 2007
  – YAGO: launched 2008
  – Extraction from Wikipedia using mappings & heuristics

• Present
  – Two of the most used knowledge graphs
  – Next to Wikidata
Wikipedia as a Knowledge Graph
Wikipedia as a Knowledge Graph

city

campus

state
Wikipedia as a Knowledge Graph

- Mapping to a central schema/ontology
DBpedia

Lehmann et al.: DBpedia – A Large-scale, Multilingual Knowledge Base
Extracted from Wikipedia. 2014
Mapping en:Infobox film

This is the mapping for the Wikipedia template Infobox film. Find usages of this Wikipedia template here. Test this mapping or in namespace File or Creator with some example Wikipedia pages. Check which projects can use this template.

Read more about mapping Wikipedia templates.

OntologyClass: Film

This is the definition of an ontology class.

Show all properties available for this class.
Show class in class hierarchy.
Read more about editing the ontology schema.
You can see the result of your edit on DBpedia Live (this is beta).

OntologyProperty: director

This is the definition of an ontology property.
Read more about editing the ontology schema.
You can see the result of your edit on DBpedia Live (this is beta).

Property Mapping

<table>
<thead>
<tr>
<th>Template Mapping</th>
<th>help</th>
</tr>
</thead>
<tbody>
<tr>
<td>map to class</td>
<td>Film</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Property Mapping</th>
<th>help</th>
</tr>
</thead>
<tbody>
<tr>
<td>template property</td>
<td>director</td>
</tr>
<tr>
<td>ontology property</td>
<td>director</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Property Mapping</th>
<th>help</th>
</tr>
</thead>
<tbody>
<tr>
<td>template property</td>
<td>producer</td>
</tr>
<tr>
<td>ontology property</td>
<td>producer</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ontology class</th>
<th>help</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdfs:label (en)</td>
<td>film</td>
</tr>
<tr>
<td>rdfs:label (en)</td>
<td>movie</td>
</tr>
<tr>
<td>rdfs:label (nl)</td>
<td>film</td>
</tr>
<tr>
<td>rdfs:label (da)</td>
<td>film</td>
</tr>
<tr>
<td>rdfs:label (de)</td>
<td>Film</td>
</tr>
<tr>
<td>rdfs:label (el)</td>
<td>ταινία</td>
</tr>
<tr>
<td>rdfs:label (fr)</td>
<td>film</td>
</tr>
<tr>
<td>rdfs:label (ko)</td>
<td>영화</td>
</tr>
<tr>
<td>rdfs:label (zh)</td>
<td>电影</td>
</tr>
<tr>
<td>rdfs:label (ar)</td>
<td>فيلم</td>
</tr>
<tr>
<td>rdfs:label (pl)</td>
<td>film</td>
</tr>
<tr>
<td>rdfs:label (ga)</td>
<td>scannán</td>
</tr>
<tr>
<td>rdfs:label (es)</td>
<td>película</td>
</tr>
<tr>
<td>rdfs:subClassOf</td>
<td>Work</td>
</tr>
<tr>
<td>rdfs:subPropertyOf</td>
<td>dul:coparticipatesWith</td>
</tr>
<tr>
<td>owl:equivalentProperty</td>
<td>schema:director, wikidata:P57</td>
</tr>
<tr>
<td>owl:disjointWith</td>
<td>schema:Movie, wikidata:P1424</td>
</tr>
</tbody>
</table>

10/17/23
Heiko Paulheim
DBpedia
YAGO

- Wikipedia categories for types
  - Plus WordNet as upper structure

- Manual mappings for properties

https://www.cs.princeton.edu/courses/archive/spring07/cos226/assignments/wordnet.html
Knowledge Graph Creation

• Lesson learned no. 4:
  – Heuristics help increasing coverage (at the cost of accuracy)
Knowledge Graph Creation: NELL

• The 2010s
  – NELL: Never ending language learner
  – Input: ontology, seed examples, text corpus
  – Output: facts, text patterns
  – Large degree of automation, occasional human feedback

• Until 2018
  – Continuously ran for ~8 years
  – New release every few days

http://rtw.ml.cmu.edu/rtw/overview
Knowledge Graph Creation: NELL

- Extraction of a Knowledge Graph from a Text Corpus

Nine Inch Nails singer Trent Reznor, born 1965...

...as stated by Filter singer Richard Patrick...

...says Slipknot singer Corey Taylor, 44, in the interview.

"X singer Y" → band_member(X,Y)

patterns

band_member(Nine_Inch_Nails, Trent_Reznor)
band_member(Filter, Richard_Patrick)
band_member(Slipknot, Corey_Taylor)

facts
Knowledge Graph Creation: NELL
Knowledge Graph Creation

- Lesson learned no. 5:
  - Quality cannot be maximized without human intervention
Summary of Trade Offs

- (Manual) effort vs. accuracy and completeness
- User involvement (or usability) vs. degree of formality
- Simplicity vs. support for plurality and provenance

→ all those decisions influence the shape of a knowledge graph!
Non-Public Knowledge Graphs

• Many companies have their own private knowledge graphs
  – Google: Knowledge Graph, Knowledge Vault
  – Yahoo!: Knowledge Graph
  – Microsoft: Satori
  – Facebook: Entities Graph
  – Thomson Reuters: permid.org (partly public)

• However, we usually know only little about them

See: Noy et al. (2019): Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it’s done
Comparison of Knowledge Graphs

- **Release cycles**

  - Instant updates: DBpedia live, Freebase, Wikidata
  - Days: NELL
  - Months: DBpedia
  - Years: YAGO, Cyc

- **Size and density**

  ![Table 1: Global Properties of the Knowledge Graphs compared in this paper](image)

<table>
<thead>
<tr>
<th></th>
<th>DBpedia</th>
<th>YAGO</th>
<th>Wikidata</th>
<th>OpenCyc</th>
<th>NELL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version</td>
<td>2016-04</td>
<td>YAGO3</td>
<td>2016-08-01</td>
<td>2016-09-05</td>
<td>08m.995</td>
</tr>
<tr>
<td># instances</td>
<td>5,109,890</td>
<td>5,130,031</td>
<td>17,581,152</td>
<td>118,125</td>
<td>1,974,297</td>
</tr>
<tr>
<td># axioms</td>
<td>397,831,457</td>
<td>1,435,808,056</td>
<td>1,633,309,138</td>
<td>2,413,894</td>
<td>3,402,971</td>
</tr>
<tr>
<td>avg. indegree</td>
<td>13.52</td>
<td>17.44</td>
<td>9.83</td>
<td>10.03</td>
<td>5.33</td>
</tr>
<tr>
<td>avg. outdegree</td>
<td>47.55</td>
<td>101.86</td>
<td>41.25</td>
<td>9.23</td>
<td>1.25</td>
</tr>
<tr>
<td># classes</td>
<td>754</td>
<td>576,331</td>
<td>30,765</td>
<td>116,822</td>
<td>290</td>
</tr>
<tr>
<td># relations</td>
<td>3,555</td>
<td>93,659</td>
<td>11,053</td>
<td>165</td>
<td>1,334</td>
</tr>
</tbody>
</table>

**Caution!**

Ringler & Paulheim: *One Knowledge Graph to Rule them All?* KI 2017
Comparison of Knowledge Graphs

• What do they actually contain?
• Experiment: pick 25 classes of interest
  – And find them in respective ontologies
• Count instances (coverage)
• Determine in and out degree (level of detail)
Comparison of Knowledge Graphs

Ringler & Paulheim: One Knowledge Graph to Rule them All? KI 2017
Comparison of Knowledge Graphs

• Summary findings:
  – Persons: more in Wikidata
    (twice as many persons as DBpedia and YAGO)
  – Countries: more details in Wikidata
  – Places: most in DBpedia
  – Organizations: most in YAGO
  – Events: most in YAGO
  – Artistic works:
    • Wikidata contains more movies and albums
    • YAGO contains more songs

Ringler & Paulheim: *One Knowledge Graph to Rule them All? KI 2017*
Caveats

• Reading the diagrams right…

• So, Wikidata contains more persons
  – but less instances of all the interesting subclasses?

• There are classes like *Actor* in Wikidata
  – but they are hardly used
  – rather: modeled using *profession* relation
Caveats

• Reading the diagrams right… (ctd.)

• So, Wikidata contains more data on countries, but less countries?
• First: Wikidata only counts current, actual countries
  – DBpedia and YAGO also count historical countries
• “KG1 contains less of X than KG2” can mean
  – it actually contains less instances of X
  – it contains equally many or more instances, but they are not typed with X (see later)
• Second: we count single facts about countries
  – Wikidata records some time indexed information, e.g., population
  – Each point in time contributes a fact
Overlap of Knowledge Graphs

• How largely do knowledge graphs overlap?
• They are interlinked, so we can simply count links
  – For NELL, we use links to Wikipedia as a proxy

Ringler & Paulheim: *One Knowledge Graph to Rule them All?* KI 2017
Overlap of Knowledge Graphs

- How largely do knowledge graphs overlap?
- They are interlinked, so we can simply count links
  - For NELL, we use links to Wikipedia as a proxy

Ringler & Paulheim: *One Knowledge Graph to Rule them All?* KI 2017
Overlap of Knowledge Graphs

• Links between Knowledge Graphs are incomplete
  – The Open World Assumption also holds for interlinks

• But we can estimate their number

• Approach:
  – find link set automatically with different heuristics
  – determine precision and recall on existing interlinks
  – estimate actual number of links

Ringler & Paulheim: One Knowledge Graph to Rule them All? KI 2017
Overlap of Knowledge Graphs

- Idea:
  - Given that the link set F is found
  - And the (unknown) actual link set would be C

- Precision P: Fraction of F which is actually correct
  - i.e., measures how much $|F|$ is over-estimating $|C|$;

- Recall R: Fraction of C which is contained in F
  - i.e., measures how much $|F|$ is under-estimating $|C|$;

- From that, we estimate $|C| = |F| \cdot P \cdot \frac{1}{R}$
Overlap of Knowledge Graphs

• Mathematical derivation:
  – Definition of recall: \( R = \frac{|F_{\text{correct}}|}{|C|} \)
  – Definition of precision: \( P = \frac{|F_{\text{correct}}|}{|F|} \)

• Resolve both to \( |F_{\text{correct}}| \), substitute, and resolve to \( |C| \)

\[
|C| = |F| \cdot P \cdot \frac{1}{R}
\]

Ringler & Paulheim: One Knowledge Graph to Rule them All? KI 2017
Overlap of Knowledge Graphs

• Experiment:
  – We use the same 25 classes as before
  – Measure 1: overlap relative to smaller KG (i.e., potential gain)
  – Measure 2: overlap relative to explicit links
    (i.e., importance of improving links)

• Link generation with 16 different metrics and thresholds
  – Intra-class correlation coefficient for |C|: 0.969
  – Intra-class correlation coefficient for |F|: 0.646

• Bottom line:
  – Despite variety in link sets generated, the overlap is estimated reliably
  – The link generation mechanisms do not need to be overly accurate

Ringler & Paulheim: *One Knowledge Graph to Rule them All?* KI 2017
Overlap of Knowledge Graphs

(a) Overlap as potential gain
(b) Overlap relative to existing links

Ringler & Paulheim: One Knowledge Graph to Rule them All? KI 2017
Overlap of Knowledge Graphs

- Summary findings:
  - DBpedia and YAGO cover roughly the same instances
    (not much surprising)
  - NELL is the most complementary to the others
  - Existing interlinks are insufficient for out-of-the-box parallel usage

Ringler & Paulheim: *One Knowledge Graph to Rule them All?* KI 2017
Intermezzo: Knowledge Graph Creation Cost

- There are quite a few metrics for evaluating KGs
  - size, degree, interlinking, quality, licensing, ...

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Alt. Metric</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability</td>
<td>A1</td>
<td>accessibility of the SPARQL endpoint and the server</td>
<td>checking whether the server responds to a SPARQL query [18]</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>accessibility of the RDF dumps</td>
<td>checking whether an RDF dump is provided and can be downloaded [18]</td>
</tr>
<tr>
<td></td>
<td>A3</td>
<td>dereferenceability of the URI</td>
<td>checking (i) for dead or broken links i.e. when an HTTP GET request is sent, the status code 404 Not Found is not be retrieved (ii) that useful data (particularly RDFS) is returned upon lookup of a URI, (iii) for changes in the URI i.e. the compliance with the recommended way of implementing redirections using the status code 303 See Other [18,30]</td>
</tr>
<tr>
<td></td>
<td>A4</td>
<td>no misreported content types</td>
<td>detect whether the HTTP response contains the header field stating the appropriate content type of the retrieved file e.g. application/rdf+xml [30]</td>
</tr>
<tr>
<td></td>
<td>A5</td>
<td>dereferenced forward links</td>
<td>dereferencability of all forward links; all available triples where the local URI is mentioned in the subject (i.e. the description of the resource) [31]</td>
</tr>
<tr>
<td>Licensing</td>
<td>L1</td>
<td>machine-readable indication of a license</td>
<td>detection of the indication of a license in the VoID description or in the dataset itself [18,31]</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>human-readable indication of a license</td>
<td>detection of a license in the documentation of the dataset [18, 31]</td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td>specifying the correct license</td>
<td>detection of whether the dataset is attributed under the same license as the original [18]</td>
</tr>
<tr>
<td>Interlinking</td>
<td>I1</td>
<td>detection of good quality interlinks</td>
<td>(i) detection of (a) interlinking degree, (b) clustering coefficients, (c) centrality, (d) open same As chains and (e) description richness through sameAs by using network measures [25], (ii) via crowdsourcing [11,65]</td>
</tr>
<tr>
<td></td>
<td>I2</td>
<td>existence of links to external data providers</td>
<td>detection of the existence and usage of external URIs (e.g. using sorial links) [31]</td>
</tr>
<tr>
<td></td>
<td>I3</td>
<td>dereferenced back-links</td>
<td>detection of all local in-links or back-links: all triples from a dataset that have the resource’s URI as the object [31]</td>
</tr>
<tr>
<td>Security</td>
<td>S1</td>
<td>usage of digital signatures</td>
<td>by signing a document containing an RDF serialization, a SPARQL result set or signing an RDF graph [13,18]</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>authenticity of the dataset</td>
<td>verifying authenticity of the dataset based on a provenance vocabulary such as author and his contributors, the publisher of the data and its sources (if present in the dataset) [18]</td>
</tr>
<tr>
<td>Performance</td>
<td>P1</td>
<td>usage of slash-URIs</td>
<td>checking for usage of slash URIs where large amounts of data is provided [18]</td>
</tr>
<tr>
<td></td>
<td>P2</td>
<td>low latency</td>
<td>(minimum) delay between submission of the request by the user and reception of the response from the server [18]</td>
</tr>
<tr>
<td></td>
<td>P3</td>
<td>high throughput</td>
<td>(maximum) no. of answered HTTP-requests per second [18]</td>
</tr>
<tr>
<td></td>
<td>P4</td>
<td>scalability of a data source</td>
<td>detection of whether the time to answer an amount of test requests divided by ten is not longer than the time it takes to answer one request [18]</td>
</tr>
</tbody>
</table>

Färber et al.: Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO SWJ 9(1), 2018
Intermezzo: Knowledge Graph Creation Cost

• ...but what is the cost of a single statement?

Some back of the envelope calculations...
Paulheim: How much is a triple?
Estimating the Cost of Knowledge Graph Creation, 2018
Intermezzo: Knowledge Graph Creation Cost

• Case 1: manual curation
  – Cyc: created by experts
    Total development cost: $120M
    Total #statements: 21M
    → $5.71 per statement
  – Freebase: created by laymen
    Assumption: adding a statement to Freebase equals adding a sentence to Wikipedia
    • English Wikipedia up to April 2011: 41M working hours
      (Geiger and Halfaker, 2013),
      size in April 2011: 3.6M pages, avg. 36.4 sentences each
    • Using US minimum wage: $2.25 per sentence
    → $2.25 per statement

(Footnote: total cost of creating Freebase would be $6.75B)
Intermezzo: Knowledge Graph Creation Cost

• Case 2: automatic/heuristic creation
  – DBpedia: 4.9M LOC, 2.2M LOC for mappings
    software project development: ~37 LOC per hour
    (Devanbu et al., 1996)
    we use German PhD salaries as a cost estimate
    → 1.85c per statement
  – YAGO: made from 1.6M LOC
    uses WordNet: 117k synsets, we treat each synset like a Wiki page
    → 0.83c per statement
  – NELL: 103k LOC
    → 14.25c per statement
• Compared to manual curation: saving factor 16-250
Intermezzo: Knowledge Graph Creation Cost

• Graph error rate against cost
  – we can pay for accuracy
  – NELL is a bit of an outlier
Enhancing the Coverage of Knowledge Graphs

- Study for KG-based Recommender Systems*
  - DBpedia (likewise: YAGO) 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/
Enhancing the Coverage of Knowledge Graphs

- 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.

A-K

- Adcock
- Acoustic
- Air Liquide
- Alarm Will Sound
- Alva Noto
- Amon Tobin
- Andy Stott
- Apparat
- Arovane
- Asypic
- Autotche
- B12
- Benn Jordan
- Biophore
- Bjork[1]
- The Black Dog
- Blanck Mass
- Boards of Canada
- Bochum Welt
- Boom Bip
- Brothomstates
- Bush
- Busub
- C418
- Cabaret Voltaire
- Casino Versus Japan
- Coephalax Acid Crew
- Coex
- Christ
- Chris Clark
- Cleoline
- Cylob
- Daedelus
- Deadbeat
- Deepchord
- Demcitke Stare
- Deru
- Richard Devine
- Dopplerreffield
- Chris Douglas
- Direcctia
- Eight Frozen Modules
- Emptyset
- Esem
- FattyDL
- Ferriesz
- The Field
- The Flashbulb
- Floating Points
- Flying Lotus
- Forest Swords
- Funkstorung
- The Future Sound Of London
- Gas
- Gecnor
- Global Communication
- Global Groon
- Golsie
- Zachary Gray[2]
- Gindflov
- Himuro Yoshiteru
- Kim Horihay
- I am Robot and Proud
- Innovaders
- Jan Jelinek
- Jega
- Jelio
- Jin
- John Tejada
- Jon Hopkins[3]
- Kettel
- Kevin Blechdom
- Kidos
- Kodomo
- Koreless[4]
Entity Extraction from List Pages

- Lists form (shallow) hierarchies

Lists of writers

From Wikipedia, the free encyclopedia

The following are lists of writers:

- Bestsellers
- Biographies
- Buddhism
- Business theorists
- Catholicism
- Children's literature
- Christian fiction

Lists by language (non-English)

- Ancient Greek
- Arabic
- Bengali
- Catalan
- Chechews

Lists by ethnicity or nationality

- African writers
- African-American writers
- Albanian writers
- Algerian writers
- American authors
- Arab American writers
- Armenian authors

Entity Extraction from List Pages

- Lists form (shallow) hierarchies
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
- 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
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

Cat2Ax: Axiomatizing Wikipedia Categories

\[
\subseteq \text{dbo:genre.\{dbr:Rock\_Music\}} \\
\subseteq \text{dbo:artist.\{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 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?
<table>
<thead>
<tr>
<th>rdfs:label</th>
<th>▪ Tiamat</th>
</tr>
</thead>
<tbody>
<tr>
<td>owl:sameAs</td>
<td>▪ dbr:Tiamat_(band)</td>
</tr>
<tr>
<td>clgo:activeYearsStartYear</td>
<td>▪ 1987</td>
</tr>
<tr>
<td>clgo:genre</td>
<td>▪ Symphonic metal</td>
</tr>
<tr>
<td>clgo:hometown</td>
<td>▪ Sweden</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
Pushing Entity Coverage Further

- Beyond red links (2020)

<table>
<thead>
<tr>
<th>Cinematic films</th>
<th>Running time</th>
<th>Year released</th>
<th>Notes</th>
</tr>
</thead>
</table>
| Amra Edda Cinema Banabo (The Innocence)                                       | 1205 min (21 hr, 5 min) | 2019 | ![|](
| Rosan (The Journey)                                                          | 873 min (14 hr, 33 min)   | 1987 | ![|](
| La Fier                                                                  | 893 min (13 hr, 22 min)   | 2018 | ![|](
| Out 1 (Not me tangere)                                                      | 775 min (12 hr, 55 min)   | 1971 | ![|](
| Evolution of a Filippo Family                                               | 593 min (9 hr, 53 min)    | 2004 | ![|](
| Shoah                                                                      | 566 min (9 hr, 28 min)    | 1986 | ![|](
| Tie Xi Qu: West of the Tracks                                                | 551 min (9 hr, 11 min)    | 2003 | ![|](
| Death in the Land of Encantos                                               | 538 min (8 hr, 58 min)    | 2007 | ![|](
| Oedred Souls                                                                | 490 min (8 hr, 15 min)    | 2016 | ![|](
| A Lullaby to the Sorrowful Mystery                                           | 485 min (8 hr, 5 min)     | 2016 | ![|](
| O.J. Made in America                                                        | 463 min (7 hr, 43 min)    | 2016 | ![|](
| Melancholia                                                                 | 420 min (7 hr, 30 min)    | 2006 | ![|](
| Sättintangó                                                                | 419 min (6 hr, 59 min)    | 1994 | ![|](
| La Roue                                                                    | 412 min (6 hr, 53 min)    | 1923 (Restoration, 2010) | ![|](
| The Best of Youth                                                          | 366 min (6 hr, 6 min)     | 2003 | ![|](
| Century of Birthing                                                        | 360 min (6 hr)            | 2011 | ![|](
| Hears Death                                                                | 350 min (5 hr, 58 min)    | 1989 | ![|](
| Karamay                                                                    | 356 min (5 hr, 56 min)    | 2011 | ![|](
| Little Dorrit                                                               | 350 min (5 hr, 50 min)    | 1987 | ![|](
| Cartos                                                                     | 339 min (5 hr, 39 min)    | 2010 | ![|](
| Mula sa Kung Aning Noon                                                    | 330 min (5 hr, 35 min)    | 2014 | ![|](
| Napoleon                                                                   | 332 min (5 hr, 32 min)    | 1927 (Restoration, 2016) | ![|](
| 1900                                                                       | 317 min (5 hr, 17 min)    | 1976 | ![|](
| Happy Hour                                                                 | 317 min (5 hr, 17 min)    | 2015 | ![|](
| Batang West Side                                                            | 310 min (5 hr, 15 min)    | 2001 | ![|](
| The Deluge                                                                 | 315 min (5 hr, 15 min)    | 1974 | ![|](
| Fanny and Alexander                                                        | 312 min (5 hr, 12 min)    | 1962 | ![|](
| Tschelef                                                                   | 304 min (5 hr, 4 min)     | 1994 | ![|](
| ![|](

- Beyond explicit lists (2021)

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

Former members [edit]
- Christoph "Hook" Micheliott – drums, electronic percussion
- Frank Kölligs – drums
- Christina Schnekenburger – keyboards
- Walter Jäger – ?
- Darren Minter – drums (1993)
- George Lewis – drums (1997)
- Oliver Röhl – drums
- Achim Farber – drums

Discography [edit]

Albums [edit]
- Samba/Khyntone (1981)
- Völe Kraft Voxba! (1982)
- Entering the Arena (1985)
- [ ] (1986)
- A – The Final Option (1993)
- The Final Remixes (1994)
- Ill – Odyssey of the Mind (1995)
- Paradise Now (1997)
- The Machinists of Joy (2013)
- Stahleiferquen (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
Beyond List Pages

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

Axl Rose

<table>
<thead>
<tr>
<th>Discography [edit]</th>
</tr>
</thead>
<tbody>
<tr>
<td>with Guns N' Roses [edit]</td>
</tr>
<tr>
<td>• Appetite for Destruction (1987)</td>
</tr>
<tr>
<td>• G N' R Lies (1988)</td>
</tr>
<tr>
<td>• Use Your Illusion I (1991)</td>
</tr>
<tr>
<td>• Use Your Illusion II (1991)</td>
</tr>
<tr>
<td>• &quot;The Spaghetti Incident?&quot; (1990)</td>
</tr>
<tr>
<td>• Chinese Democracy (2008)</td>
</tr>
<tr>
<td>with Hollywood Rose [edit]</td>
</tr>
<tr>
<td>• The Roots of Guns N' Roses (2004)</td>
</tr>
<tr>
<td>with Rapidfire [edit]</td>
</tr>
<tr>
<td>• Ready to Rumble EP (2014)</td>
</tr>
<tr>
<td>Guest appearances [edit]</td>
</tr>
<tr>
<td>• The Decline of Western Civilization Part II: The Metal Years – Original Motion Picture Soundtrack by various artists (1988; &quot;Under My Wheels&quot; ft. Alice Cooper, Slash and Izzy Stradlin)</td>
</tr>
<tr>
<td>• The End of the Innocence by Don Henley (1988; &quot;I Will Not Go Quietly&quot;)</td>
</tr>
<tr>
<td>• Fire and Gasoline by Steve Jones (1989; &quot;I Did U No Wrong&quot;)</td>
</tr>
<tr>
<td>• Pawnshop Guitar by Gilby Clarke (1994; &quot;Dead Flowers&quot;)</td>
</tr>
<tr>
<td>• Aroused Disease by The Outpatients (1996; &quot;Anxious Disease&quot; ft. Slash)</td>
</tr>
<tr>
<td>• Angel Down by Sebastian Bach (2007; &quot;Back in the Saddle;&quot; &quot;Love Is a Bitchslap;&quot; &quot;Stuck Inside&quot;)</td>
</tr>
<tr>
<td>• New Looney Tunes (2016; &quot;Roll the Rock&quot;)</td>
</tr>
</tbody>
</table>

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>1985</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-OST 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)[123]</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

(artist, Axl_Rose)
∈ topSection."Discography"
∈ artist.{->PageEntity-}

(type, MusicalWork)
∈ topSection."Discography"
∈ MusicalWork

(musicalBand, Guns_N_Roses)
∈ topSection."Discography"
∈ sectionEntityType.{Band}
∈ musicalBand.{->SectionEntity-}
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
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*
CaLiGraph Glitches

About: clgr:Mannheim

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdfs:label</td>
<td>Mannheim</td>
</tr>
<tr>
<td>clgo:country</td>
<td>Moldova, Germany</td>
</tr>
<tr>
<td>rdf:type</td>
<td>Planned capital</td>
</tr>
<tr>
<td></td>
<td>City in Baden-Württemberg</td>
</tr>
<tr>
<td></td>
<td>Twin town or sister city</td>
</tr>
<tr>
<td></td>
<td>Coat of arms with the Palatine Lion</td>
</tr>
<tr>
<td></td>
<td>French exonym for German toponyms</td>
</tr>
<tr>
<td></td>
<td>Twin town or sister city in Lithuania</td>
</tr>
<tr>
<td></td>
<td>University town in Germany</td>
</tr>
<tr>
<td></td>
<td>clgo:NamedIndividual</td>
</tr>
<tr>
<td></td>
<td>City or town in Germany</td>
</tr>
<tr>
<td></td>
<td>Most polluted city in the world</td>
</tr>
</tbody>
</table>

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 in this list are townes, the list also comprises villages, cities, districts, counties, etc. with similar links.
From DBpedia to DBkWik

• Wikipedia-based Knowledge Graphs will remain an essential building block of Semantic Web applications
• But they suffer from...
  – ...a coverage bias
  – ...limitations of the creating heuristics
• One (but not the only!) possible source of coverage bias
  – Articles about long-tail entities become deleted

Wikipedia:Deletion policy

From Wikipedia, the free encyclopedia


This page documents an English Wikipedia policy.
It describes a widely accepted standard that all editors should normally
This page in a nutshell: Administrators have the ability to delete
in accordance with established policies and guidelines, and comm
Deletion of a Wikipedia article removes the current version and all pr
reverse ("undelete") any deletion. All such actions (other than visibl
administrators normally will not delete it.

Contents [hide]
1 Reasons for deletion
2 Alternatives to deletion
  2.1 Editing and discussion
  2.2 Tagging
2.3 Mergers
2.4 Redirection
2.5 Incubation
2.6 Other projects
2.7 Archiving
3 Processes
  3.1 Copyright violations
  3.2 speedy deletion
  3.3 Proposed deletion
    3.3.1 Proposed deletion of biographies of living people
  3.4 Deletion discussion
  3.5 Page deletion
  3.6 Deletion of biographies and BLPs
  3.7 Deletion review

http://speedydeletion.wikia.com
From DBpedia to DBkWik

• Why stop at Wikipedia?
• Wikipedia is based on the MediaWiki software
  – ...and so are thousands of Wikis
  – Fandom by Wikia: >385,000 Wikis on special topics
  – WikiApiary: reports >20,000 installations of MediaWiki on the Web
From DBpedia to DBkWik

• Collecting Data from a Multitude of Wikis
From DBpedia to DBkWik

• The DBpedia Extraction Framework consumes MediaWiki dumps
• Experiment (started as team project 2017)
  – Can we process dumps from arbitrary Wikis with it?
  – Are the results somewhat meaningful?
From DBpedia to DBkWik

- Example from Harry Potter Wiki

DBkWik


<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>owl:sameAs</td>
<td>dor:Gryffindor</td>
</tr>
<tr>
<td>dcterms:subject</td>
<td>dbkwik:HarryPotter/resource/Category/Gryffindor_House</td>
</tr>
<tr>
<td></td>
<td>dbkwik:HarryPotter/resource/Category/Hogwarts_Houses</td>
</tr>
<tr>
<td>skos:altLabel</td>
<td>Gryffindor House, Gryffindor Students, Gryffindor, Gryffindors, Gryffindor, Gryffindor Students, Gryffindor Girl, The Gryffindors</td>
</tr>
<tr>
<td>skos:prefLabel</td>
<td>Gryffindor</td>
</tr>
<tr>
<td>doiwik:HarryPotter/property/animal</td>
<td>doiwik:HarryPotter/resource/Lion</td>
</tr>
</tbody>
</table>

From DBpedia to DBkWik

• Differences to DBpedia
  – DBpedia has manually created mappings to an ontology
  – Wikipedia has one page per subject
  – Wikipedia has global infobox conventions (more or less)

• Challenges
  – On-the-fly ontology creation
  – Instance matching
  – Schema matching

Hertling & Paulheim: *DBkWik: A Consolidated Knowledge Graph from Thousands of Wikis*. ICBK 2018
From DBpedia to DBkWik

- Heuristics
  - Ontology induction
  - Instance/Schema Matching

Hertling & Paulheim: *DBkWik: A Consolidated Knowledge Graph from Thousands of Wikis*. ICBK 2018
From DBpedia to DBkWik

• Downloaded ~15k Wiki dumps from Fandom
  – 52.4GB of data, roughly the size of the English Wikipedia

• Prototype: extracted data for ~250 Wikis
  – 4.3M instances, ~750k linked to DBpedia
  – 7k classes, ~1k linked to DBpedia
  – 43k properties, ~20k linked to DBpedia
  – ...including duplicates!

• Link quality
  – Good for classes, OK for properties (F1 of .957 and .852)
  – Needs improvement for instances (F1 of .641)
From DBpedia to DBkWik

- Scalability of matching:
  - Pairwise matching does not scale
  - 300k Wikis, 1 minute for each pair → 171k years

- Iteratively match and merge
  - 300k Wikis, 1 minute for each match&merge run → 200 days

- Tree-shaped execution plan
  - Parallelizable
  - Hierarchical clustering by topic
  - Whole run under a week
Hertling & Paulheim: *WebIsALOD: Providing Hypernymy Relations extracted from the Web as Linked Open Data*. ISWC 2017

- Background: Web table interpretation
- Most approaches need typing information
  - DBpedia etc. have too little coverage on the long tail
  - Wanted: extensive type database
WebIsALOD

• Extraction of type information using Hearst-like patterns, e.g.,
  – T, such as X
  – X, Y, and other T

• Text corpus: common crawl
  – ~2 TB crawled web pages
  – Fast implementation: regex over text
  – “Expensive” operations only applied once regex has fired

• Resulting database
  – 400M hypernymy relations

Seitner et al.: A large DataBase of hypernymy relations extracted from the Web.
LREC 2016
WebIsALOD

• Example:

About: fiction writer

Premodifier: fiction
Head noun: writer

<table>
<thead>
<tr>
<th>label</th>
<th>provenance</th>
<th>confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>writer</td>
<td>isap:391280092</td>
<td>0.799331</td>
</tr>
<tr>
<td>great idea</td>
<td>isap:493244047</td>
<td>0.672180</td>
</tr>
<tr>
<td>several magazine</td>
<td>isap:104101164</td>
<td>0.655684</td>
</tr>
<tr>
<td>category</td>
<td>isap:104762336</td>
<td>0.477191</td>
</tr>
<tr>
<td>artist</td>
<td>isap:387107910</td>
<td>0.471280</td>
</tr>
<tr>
<td>blog</td>
<td>isap:492616662</td>
<td>0.458511</td>
</tr>
<tr>
<td>writers</td>
<td>isap:439622913</td>
<td>0.427701</td>
</tr>
<tr>
<td>story</td>
<td>isap:122402598</td>
<td>0.306667</td>
</tr>
<tr>
<td>group</td>
<td>isap:115379219</td>
<td>0.299656</td>
</tr>
<tr>
<td>poet</td>
<td>isap:49228497</td>
<td>0.287619</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>label</th>
<th>provenance</th>
<th>confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>george orwell</td>
<td>isap:386458501</td>
<td>0.662121</td>
</tr>
<tr>
<td>science fiction</td>
<td>isap:275668279</td>
<td>0.635865</td>
</tr>
<tr>
<td>franz kafka</td>
<td>isap:159147340</td>
<td>0.602015</td>
</tr>
<tr>
<td>steve almond</td>
<td>isap:392552636</td>
<td>0.581515</td>
</tr>
<tr>
<td>dan brown</td>
<td>isap:157209267</td>
<td>0.574564</td>
</tr>
<tr>
<td>james joyce</td>
<td>isap:159394667</td>
<td>0.561794</td>
</tr>
<tr>
<td>stephen king</td>
<td>isap:306753456</td>
<td>0.567364</td>
</tr>
<tr>
<td>flannery oconnor</td>
<td>isap:266705231</td>
<td>0.555096</td>
</tr>
<tr>
<td>alicia munro</td>
<td>isap:162537618</td>
<td>0.552668</td>
</tr>
<tr>
<td>ayn rand</td>
<td>isap:301402665</td>
<td>0.528867</td>
</tr>
</tbody>
</table>

http://webisa.webdatacommons.org/
WebIsALOD

- Initial effort: transformation to a LOD dataset
  - including rich provenance information

Hertling & Paulheim: **WebIsALOD: Providing Hypernymy Relations extracted from the Web as Linked Open Data. ISWC 2017**
WebIsALOD

• Estimated contents breakdown

Hertling & Paulheim: WebIsALOD: Providing Hypernymy Relations extracted from the Web as Linked Open Data. ISWC 2017
WebIsALOD

• Main challenge
  – Original dataset is quite noisy (<10% correct statements)
  – Recap: coverage vs. accuracy
  – Simple thresholding removes too much knowledge

• Approach
  – Train RandomForest model for predicting correct vs. wrong statements
  – Using all the provenance information we have
  – Use model to compute confidence scores

• Current ongoing research
  – Using transformers and a larger training set

Hertling & Paulheim: WebIsALOD: Providing Hyponymy Relations extracted from the Web as Linked Open Data. ISWC 2017
WebIsALOD

• Current challenges and works in progress
  – Distinguishing instances and classes
    • i.e.: subclass vs. instance of relations
  – Splitting instances
    • *Bauhaus is a goth band*
    • *Bauhaus is a German school*
  – Knowledge extraction from pre and post modifiers
    • *Bauhaus is a goth band* → genre(Bauhaus, Goth)
    • *Bauhaus is a German school* → location(Bauhaus, Germany)

Hertling & Paulheim: *WebIsALOD: Providing Hypernymy Relations extracted from the Web as Linked Open Data*. ISWC 2017
Summary

- We have seen a couple of Knowledge Graphs
  - How they are built
  - What they contain

- For your project
  - Have a look at the fit for your domain
  - Try different options

- For a master’s thesis later
  - Work on recent developments in our group
Questions?