Semantic Web Technologies
Public Knowledge Graphs

Heiko Paulheim
Previously on “Semantic Web Technologies”

• Linked Open Data
  – We know the principles
  – We have seen examples for some datasets

• Today
  – A closer look on actual examples
  – Some useful, large-scale resources
Growing Interest in Knowledge Graphs

**Introducing the Knowledge Graph: things, not strings**

Amit Singhal
SVP, Engineering

Published May 16, 2012

Search is a lot about discovery—the basic human need to learn and broaden your horizons. But searching still requires a lot of hard work by you, the user. So today I’m really excited to launch the Knowledge Graph, which will help you discover new information quickly and easily.

Take a query like [taj mahal]. For more than four decades, search has essentially been about matching keywords to queries. To a search engine the words [taj mahal] have been just that—two words.

But we all know that [taj mahal] has a much richer meaning. You might think of one of the world’s most beautiful monuments, or a Grammy Award-winning musician, or possibly even a casino in Atlantic City, NJ.

Sources: Google
Introduction

• Knowledge Graphs on the Web
• Everybody talks about them, but what *is* a Knowledge Graph?
  – I don’t have a definition either...

“Please define what a knowledge graph is – and what it is not.”
Definitions

- Knowledge graphs could be envisaged as a network of all kind things which are relevant to a specific domain or to an organization. They are not limited to abstract concepts and relations but can also contain instances of things like documents and datasets. (Blumauer, 2014)

- We define a Knowledge Graph as an RDF graph. (Färber and Rettinger, 2015)

- Knowledge graphs are large networks of entities, their semantic types, properties, and relationships between entities. (Kroetsch and Weikum, 2016)

- [...] systems exist, [...], which use a variety of techniques to extract new knowledge, in the form of facts, from the web. These facts are interrelated, and hence, recently this extracted knowledge has been referred to as a knowledge graph. (Pujara et al., 2013)

Ehrlinger and Wöß: Towards a Definition of Knowledge Graphs. 2016
Definitions

• My working definition: a Knowledge Graph
  – *mainly* describes instances and their relations in a graph
    • Unlike an ontology
    • Unlike, e.g., WordNet
  – Defines possible classes and relations in a *schema* or *ontology*
    • Unlike schema-free output of some IE tools
  – Allows for interlinking *arbitrary* entities with each other
    • Unlike a relational database
  – Covers *various* domains
    • Unlike, e.g., Geonames

*Paulheim: Knowledge graph refinement:
A survey of approaches and evaluation methods, 2017.*
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

Discuss "Arnold Schwarzenegger"  Show Empty Fields

- 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: Thal, Austria
- Country Of Nationality: United States
- Profession: Politician, Bodybuilder, Entrepreneur, Actor
- Religion: Roman Catholicism
- Parents: Aurelia Jadmy 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

Max. user involvement

Max. degree of formality
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)
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
  – ...with Wikidata catching up
DBpedia

Lehmann et al.: *DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia*. 2014
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, says Slipknot singer Corey Taylor, 44, in the interview. "X singer Y" ➔ band_member(X, Y)

patterns

facts

band_member(Nine_Inch_Nails, Trent_Reznor)
band_member(Filter, Richard_Patrick)
band_member(Slipknot, Corey_Taylor)
Knowledge Graph Creation: NELL
Knowledge Graph Creation

• Lesson learned no. 5:
  – Quality cannot be maximized without human intervention

Min. human intervention

Max. accuracy
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
Non-Public Knowledge Graphs

• Knowledge Graphs are used...
• ...in companies and organizations
  – collect, organize, and integrate knowledge
  – link isolated information sources
  – make information searchable and findable

Masuch, 2014
Comparison of Knowledge Graphs

• Release cycles

Instant updates: DBpedia live, Freebase, Wikidata

Days: NELL

Months: DBpedia

Years: YAGO, Cyc

Caution!

• Size and density

Table 1: Global Properties of the Knowledge Graphs compared in this paper

<table>
<thead>
<tr>
<th>Version</th>
<th>DBpedia</th>
<th>YAGO</th>
<th>Wikidata</th>
<th>OpenCyc</th>
<th>NELL</th>
</tr>
</thead>
<tbody>
<tr>
<td></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>

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

(a) Number of instances
(b) Average indegree
(c) Average outdegree

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
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

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

(a) Overlap as potential gain  
(b) Overlap relative to existing links
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 2

<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 if a URI is valid, i.e., does it point to a valid resource</td>
</tr>
<tr>
<td></td>
<td>A4</td>
<td>no misreported content types</td>
<td>detect whether the HTTP response contains the correct content type</td>
</tr>
<tr>
<td></td>
<td>A5</td>
<td>dereferenced forward links</td>
<td>dereferencedforward 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 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 coefficient, (c) centrality, (d) open sameAs chains and (e) description richness through sameAs by using network measures [25], (ii) via crowdsourcing [1,65]</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 callcounters 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 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 static-URIs</td>
<td>checking for usage of static 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 a request by the user and reception of the response from the system [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 ten requests divided by ten is not longer than the time it takes to answer one request [18]</td>
</tr>
</tbody>
</table>

### Table 14

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Metric</th>
<th>DBpedia</th>
<th>Freebase</th>
<th>OpenCyc</th>
<th>Wikidata</th>
<th>YAGO</th>
<th>Example of User Weighting $w_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>$\text{recall}$</td>
<td>0.994</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.624</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\text{precision}$</td>
<td>0.994</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.624</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\text{F-measure}$</td>
<td>0.994</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.624</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\text{transitivity}$</td>
<td>0.5</td>
<td>0.5</td>
<td>0.755</td>
<td>0.755</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>$\text{consistency}$</td>
<td>0.875</td>
<td>0.999</td>
<td>0.333</td>
<td>0.333</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>$\text{Completeness}$</td>
<td>0.991</td>
<td>0.45</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\text{Tolerance}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\text{Ease of understanding}$</td>
<td>0.965</td>
<td>0.362</td>
<td>0.955</td>
<td>0.955</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>$\text{Irreversibility}$</td>
<td>0.402</td>
<td>0.425</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\text{Accessibility}$</td>
<td>0.5</td>
<td>0</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$\text{Usability}$</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</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
New Kids on the Block

Subjective age: Measured by the fraction of the audience that understands a reference to your young days’ pop culture...
Further Sources of Knowledge in Wikipedia

- show: list pages, categories, tables, ...

### Track Listing

**Original Release**

All tracks written by Trent Reznor.

<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&quot;Mr. Self Destruct&quot;</td>
<td>4:30</td>
</tr>
<tr>
<td>2</td>
<td>&quot;Piggy&quot;</td>
<td>4:24</td>
</tr>
<tr>
<td>3</td>
<td>&quot;Heresy&quot;</td>
<td>3:54</td>
</tr>
<tr>
<td>4</td>
<td>&quot;March of the Pigs&quot;</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>&quot;Closer&quot;</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>&quot;Ruiner&quot;</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>&quot;The Becoming&quot;</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>&quot;I Do Not Want This&quot;</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>&quot;Big Man with a Gun&quot;</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>&quot;A Warm Place&quot;</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>&quot;Eraser&quot;</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>&quot;Reptile&quot;</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>&quot;The Downward Spiral&quot;</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>&quot;Hurt&quot;</td>
<td></td>
</tr>
</tbody>
</table>

### Awards

For a more comprehensive list, see List of awards and nominations received by Nine Inch Nails.

<table>
<thead>
<tr>
<th>Year</th>
<th>Nominee/work</th>
<th>Award</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>&quot;Wish&quot;</td>
<td>Best Metal Performance</td>
<td>Won</td>
</tr>
<tr>
<td>1995</td>
<td>The Downward Spiral</td>
<td>Best Alternative Music Performance</td>
<td>Nominated</td>
</tr>
<tr>
<td>1995</td>
<td>Happiness in Slavery (from Woodstock '94 compilation)</td>
<td>Best Metal Performance</td>
<td>Won</td>
</tr>
<tr>
<td>1997</td>
<td>&quot;Hurt&quot;</td>
<td>Best Rock Song</td>
<td></td>
</tr>
<tr>
<td>1997</td>
<td>&quot;The Perfect Drug&quot;</td>
<td>Best Hard Rock Performance</td>
<td>Nominated</td>
</tr>
<tr>
<td>1999</td>
<td>The Fridge</td>
<td>Best Metal Performance</td>
<td>Nominated</td>
</tr>
<tr>
<td>1999</td>
<td>&quot;Starfuckers, Inc.&quot;</td>
<td>Best Metal Performance</td>
<td>Nominated</td>
</tr>
<tr>
<td>2000</td>
<td>Info the Void</td>
<td>Best Male Rock Vocal Performance</td>
<td>Nominated</td>
</tr>
<tr>
<td>2005</td>
<td>&quot;The Hand That Feeds&quot;</td>
<td>Best Hard Rock Performance</td>
<td>Nominated</td>
</tr>
<tr>
<td>2006</td>
<td>&quot;Every Day Is Exactly the Same&quot;</td>
<td>Best Hard Rock Performance</td>
<td>Nominated</td>
</tr>
<tr>
<td>2009</td>
<td>&quot;34 Ghosts IV&quot;</td>
<td>Best Rock Instrumental Performance</td>
<td>Nominated</td>
</tr>
<tr>
<td>2009</td>
<td>Ghosts I-IV</td>
<td>Best Boxed Set or Limited Edition Package</td>
<td>Nominated</td>
</tr>
<tr>
<td>2013</td>
<td>Hesitation Marks</td>
<td>Best Alternative Music Album</td>
<td>Nominated</td>
</tr>
</tbody>
</table>

### Categories

- 1994 albums
- Albums produced by Flood (producer)
- Albums produced by Trent Reznor
- Concept albums
- Interscope Records albums
- Nine Inch Nails albums
- Nothing Records albums
- Obscenity controversies in music
CaLiGraph Idea

- Entities co-occur in surface patterns
  - e.g., enumerations, table columns, …
- Co-occurring entities share semantic patterns
  - e.g., types, relations, attribute values
- Existing entities co-occur with new entities
CaLiGraph Idea

- Surface patterns and semantic patterns also exist outside of Wikipedia
CaLiGraph – Current State

- Significant coverage enhancements of DBpedia Properties
CaLiGraph – Current State

- Significant instance set enhancements by list extraction
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
From DBpedia to DBkWik

- One (but not the only!) possible source of coverage bias
  - Articles about long-tail entities become deleted
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

Trent Reznor

- Instruments: Vocals, Guitar, Keyboards, Bass, Marimba, Saxophone, Small Percussion
- Years: 1988–present
- Tours: VIVIsectVI–present

Role: Composer

Born: May 17, 1965
Mercer, Pennsylvania, USA

1 Nomination / 1 Win

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)
From DBpedia to DBkWik

- The DBpedia Extraction Framework consumes MediaWiki dumps
- Experiment
  - Can we process dumps from arbitrary Wikis with it?
  - Are the results somewhat meaningful?
From DBpedia to DBkWik

• Example from Harry Potter Wiki

http://dbkwik.webdatacommons.org/HarryPotter/resource/Gryffindor

http://dbkwik.org/
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)
Solving the Integration Problems in DBkWik

- A new task at OAEI since 2018
  - Benchmark for schema-instance matching tools
  - Turned out to be non-trivial
WebIsALOD

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

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

http://webisa.webdatacommons.org/

10/30/20  Heiko Paulheim  72
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
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

Hertling & Paulheim: *WebIsALOD: Providing Hypernymy 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?