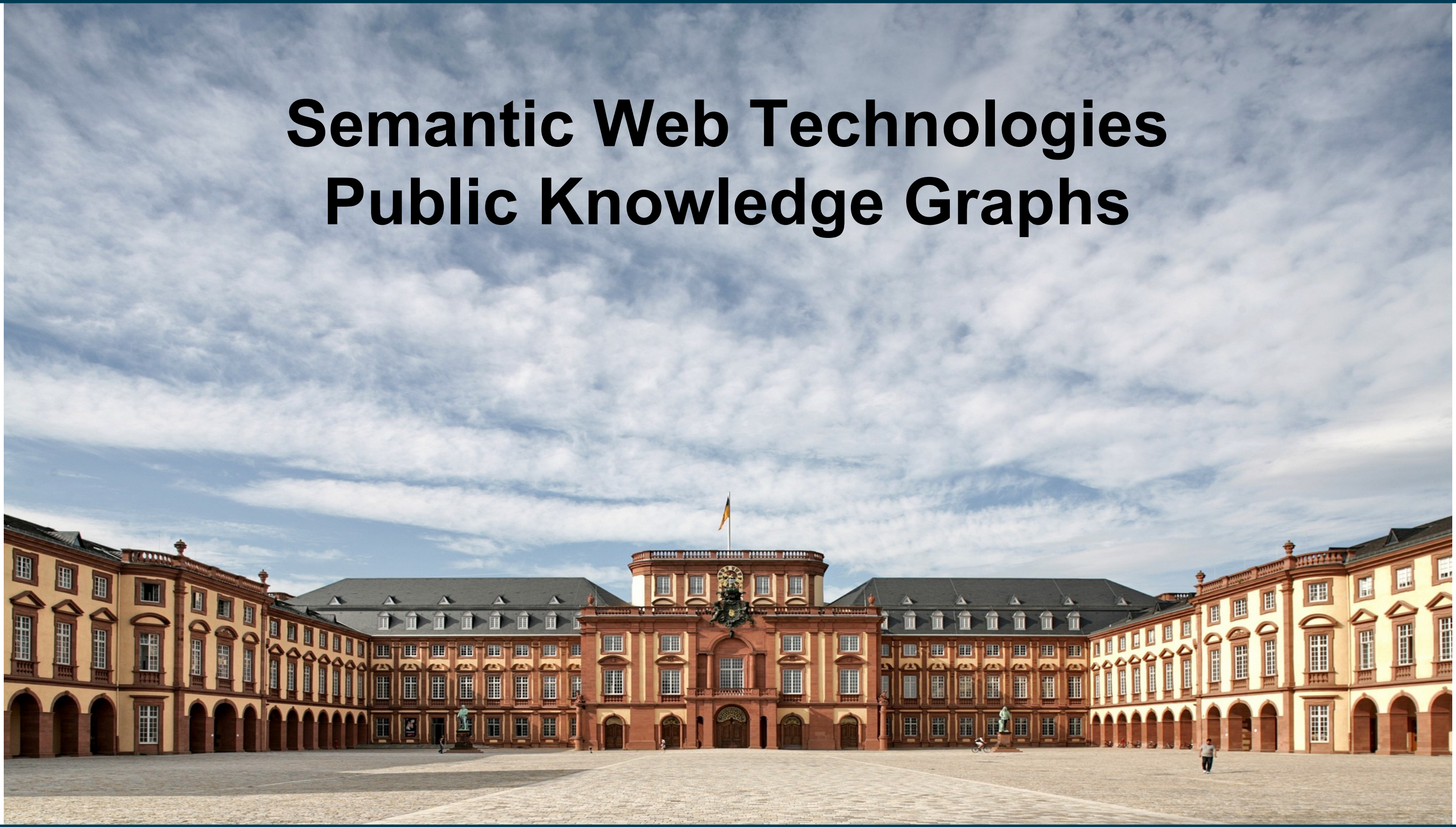
UNIVERSITÄT
MANNHEIM



# Semantic Web Technologies
# Public Knowledge Graphs

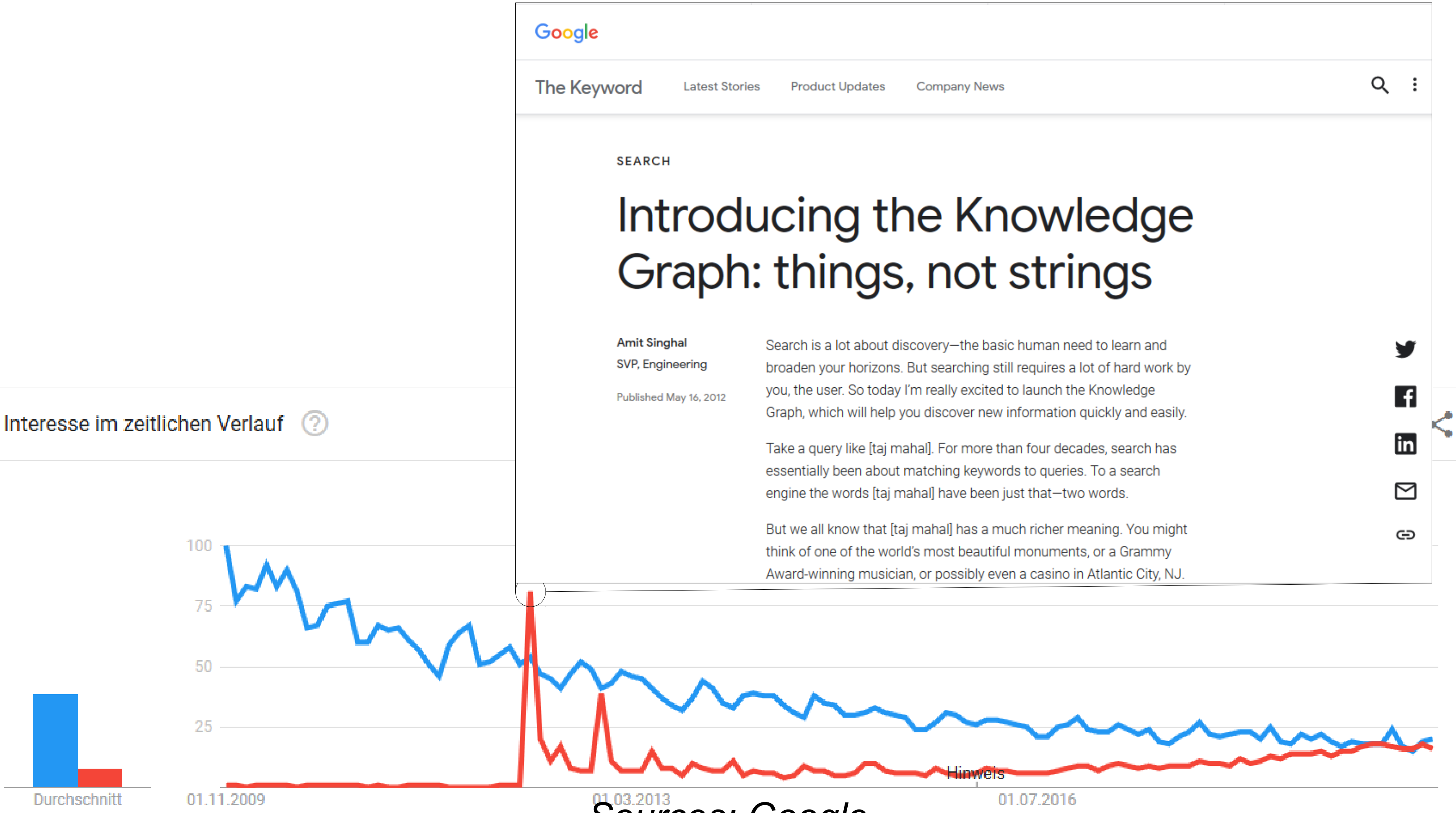
Heiko Paulheim

# 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



Google

The Keyword

Latest Stories

Product Updates

Company News

SEARCH

## 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...

Journal Paper Review, (Natasha Noy, Google, June 2015):
*“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*

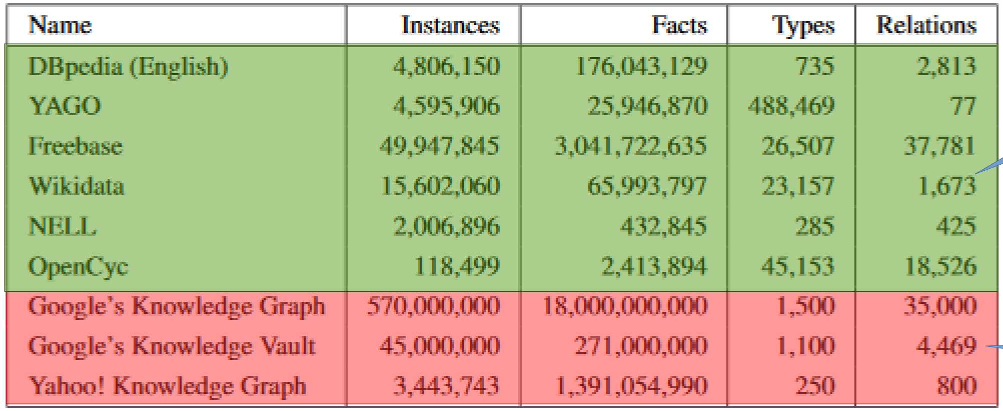
# Introduction

- 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)



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

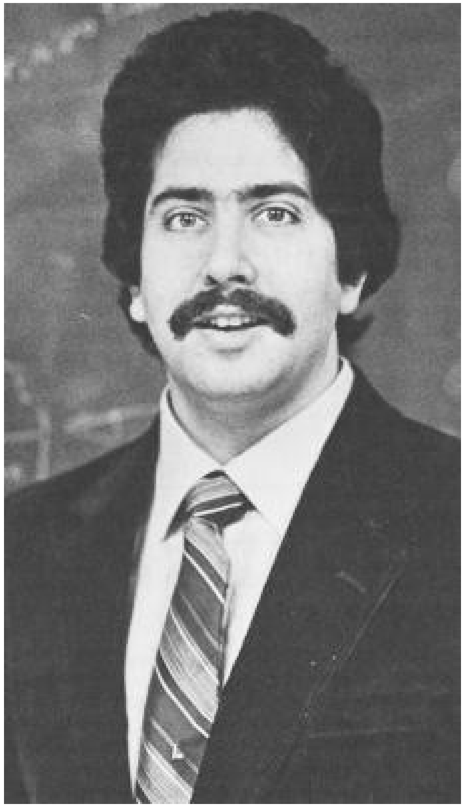
public

private

Paulheim: *Knowledge graph refinement: A survey of approaches and evaluation methods.* Semantic Web 8:3 (2017), pp. 489-508

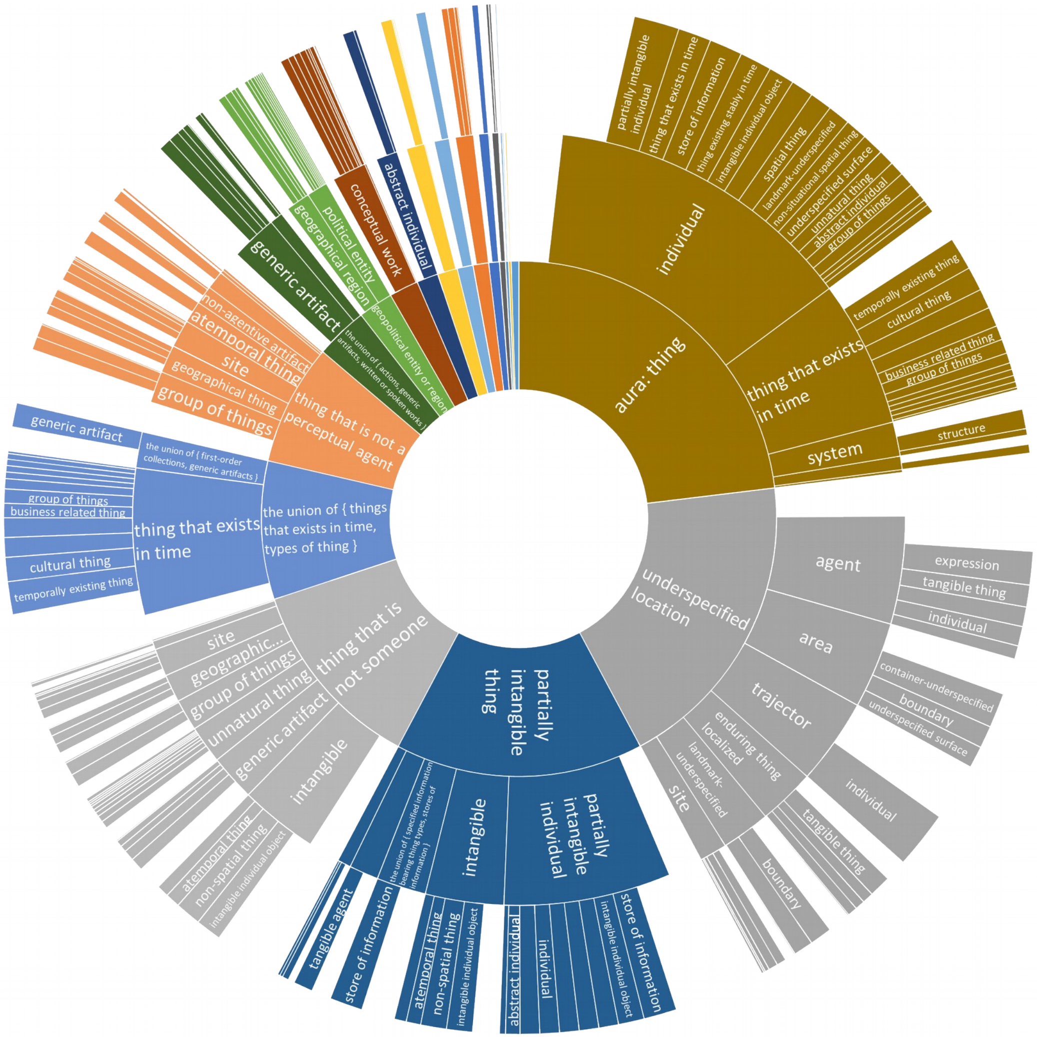
# Knowledge Graph Creation: CyC

- 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



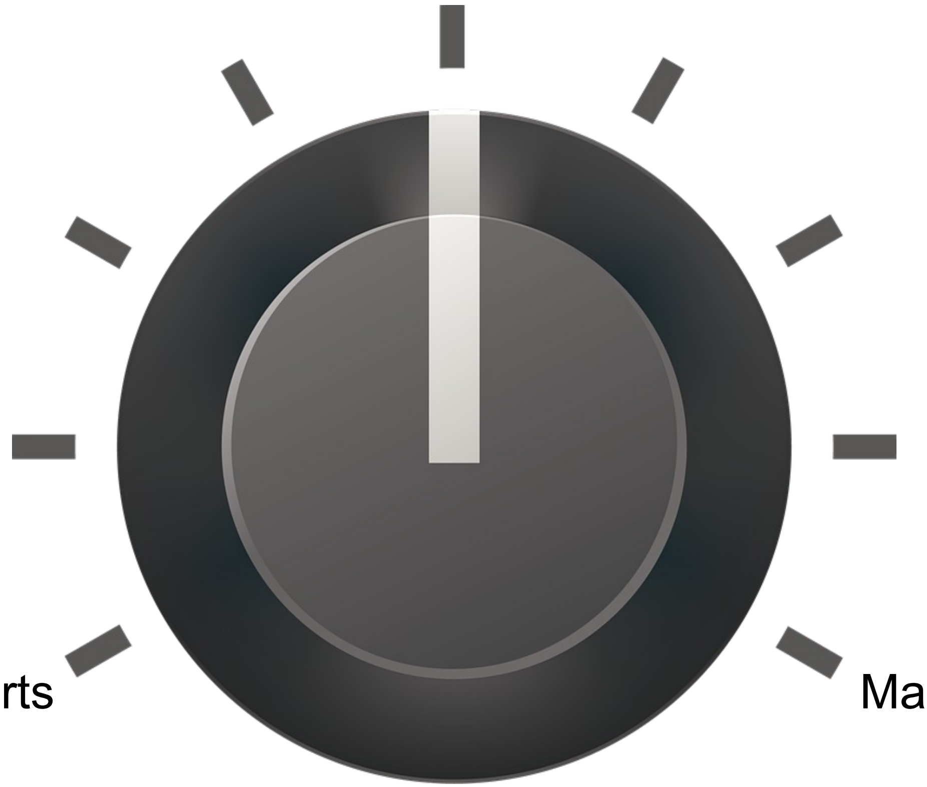
- The present (as of June 2017)
  - ~1,000 person years, $120M total development cost
  - 21M axioms and rules
  - Used to exist until 2017

# Knowledge Graph Creation: CyC



# Knowledge Graph Creation

- Lesson learned no. 1:
  - Trading efforts against accuracy



Min. efforts

Max. accuracy

# Knowledge Graph Creation: Freebase

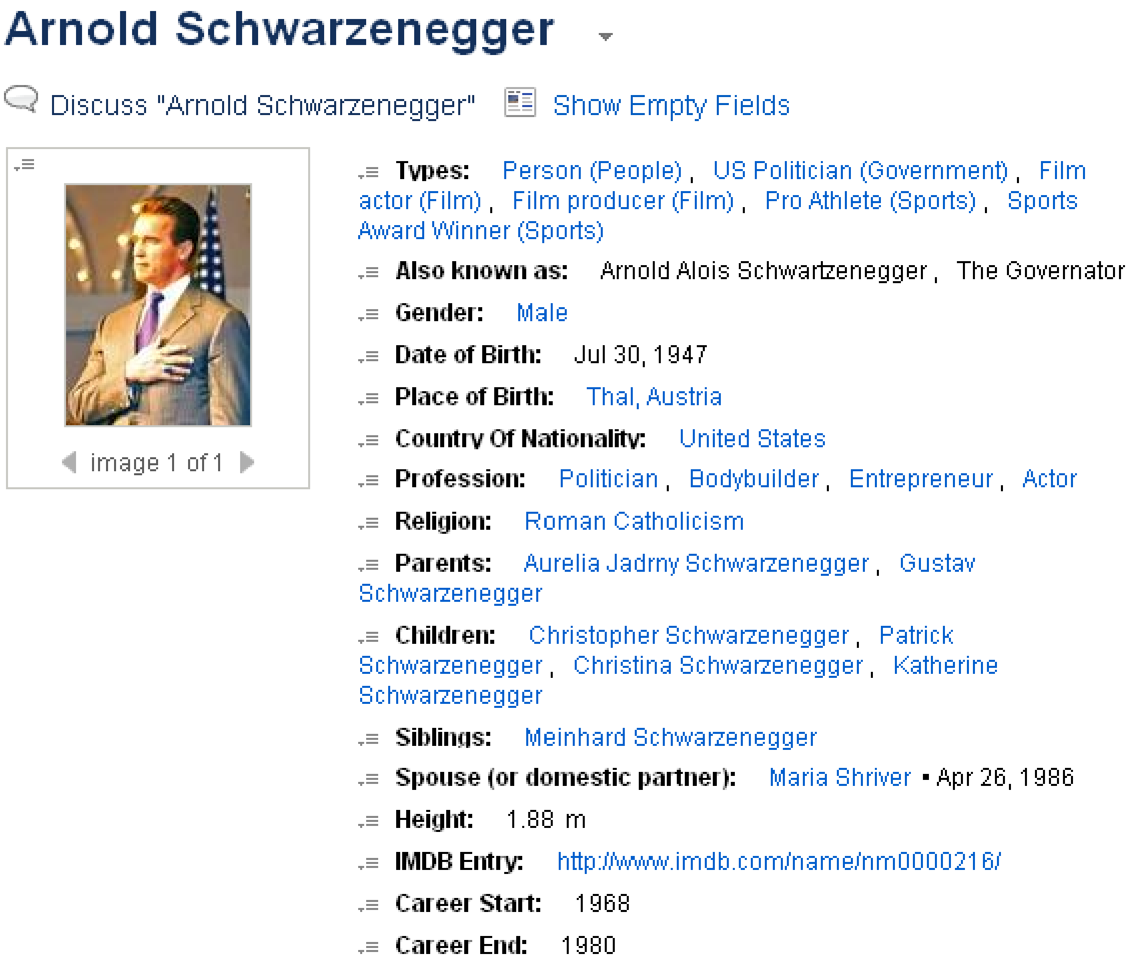
- The 2000s
  - Freebase: collaborative editing
  - Schema not fixed

- 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 Schwartzenegger, The Governator

**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 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

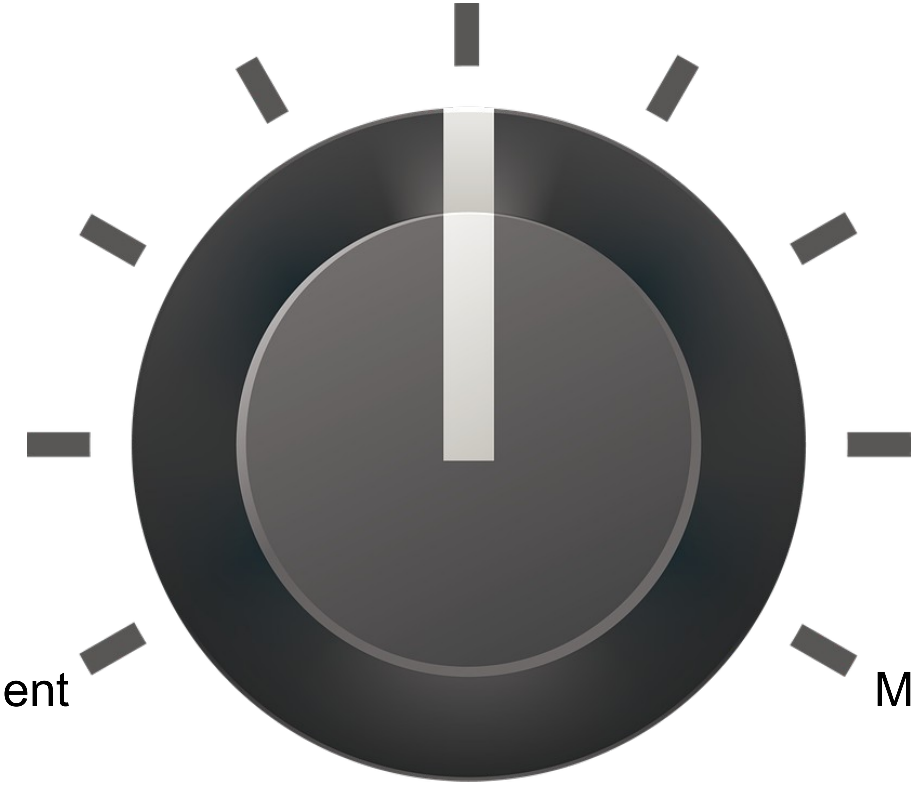
**IMDB Entry:** http://www.imdb.com/name/nm0000216/

**Career Start:** 1968

**Career End:** 1980

# Knowledge Graph Creation

- Lesson learned no. 2:
  - Trading formality against number of users



Max. user involvement

Max. degree of formality

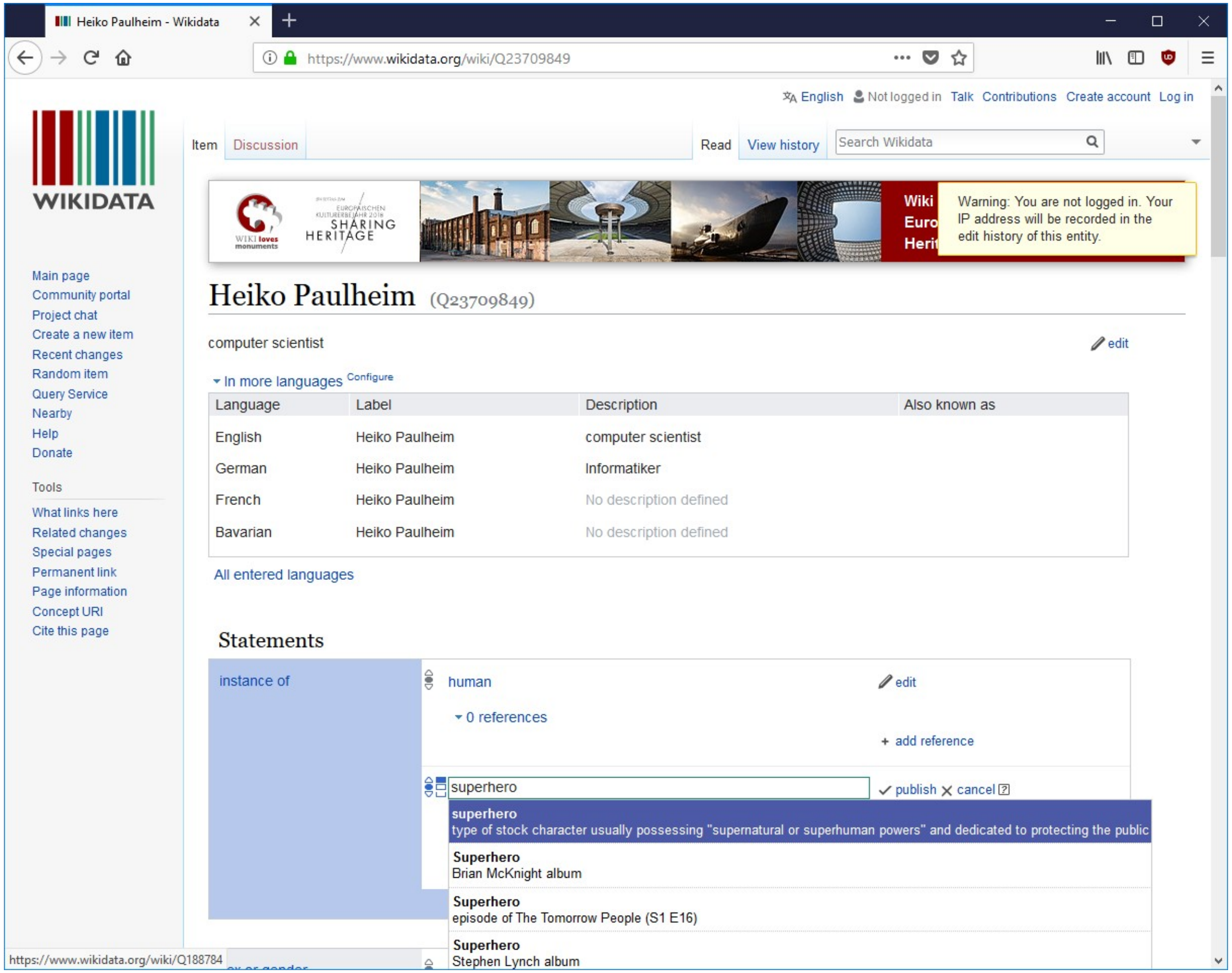
# Knowledge Graph Creation: Wikidata

- 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



Warning: You are not logged in. Your IP address will be recorded in the edit history of this entity.

Heiko Paulheim (Q23709849)

computer scientist

In more languages

| Language | Label | Description | Also known as |
|---|---|---|---|
| English | Heiko Paulheim | computer scientist | |
| German | Heiko Paulheim | Informatiker | |
| French | Heiko Paulheim | No description defined | |
| Bavarian | Heiko Paulheim | No description defined | |

All entered languages

Statements

instance of: human

0 references

add reference

superhero

**superhero**
type of stock character usually possessing "supernatural or superhuman powers" and dedicated to protecting the public

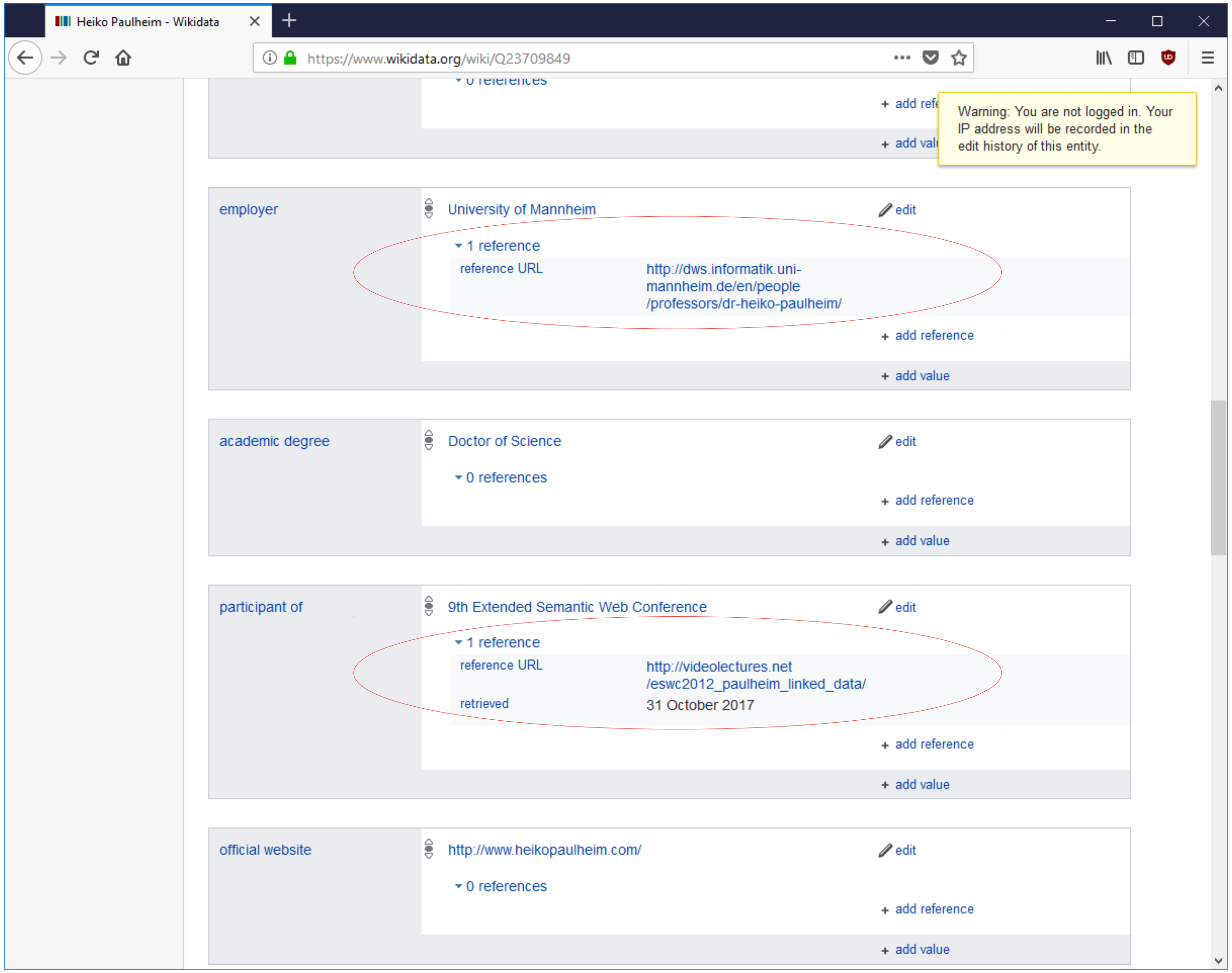
**Superhero**
Brian McKnight album

**Superhero**
episode of The Tomorrow People (S1 E16)

**Superhero**
Stephen Lynch album

# Knowledge Graph Creation: Wikidata

- Provenance



Heiko Paulheim - Wikidata
https://www.wikidata.org/wiki/Q23709849

Warning: You are not logged in. Your IP address will be recorded in the edit history of this entity.

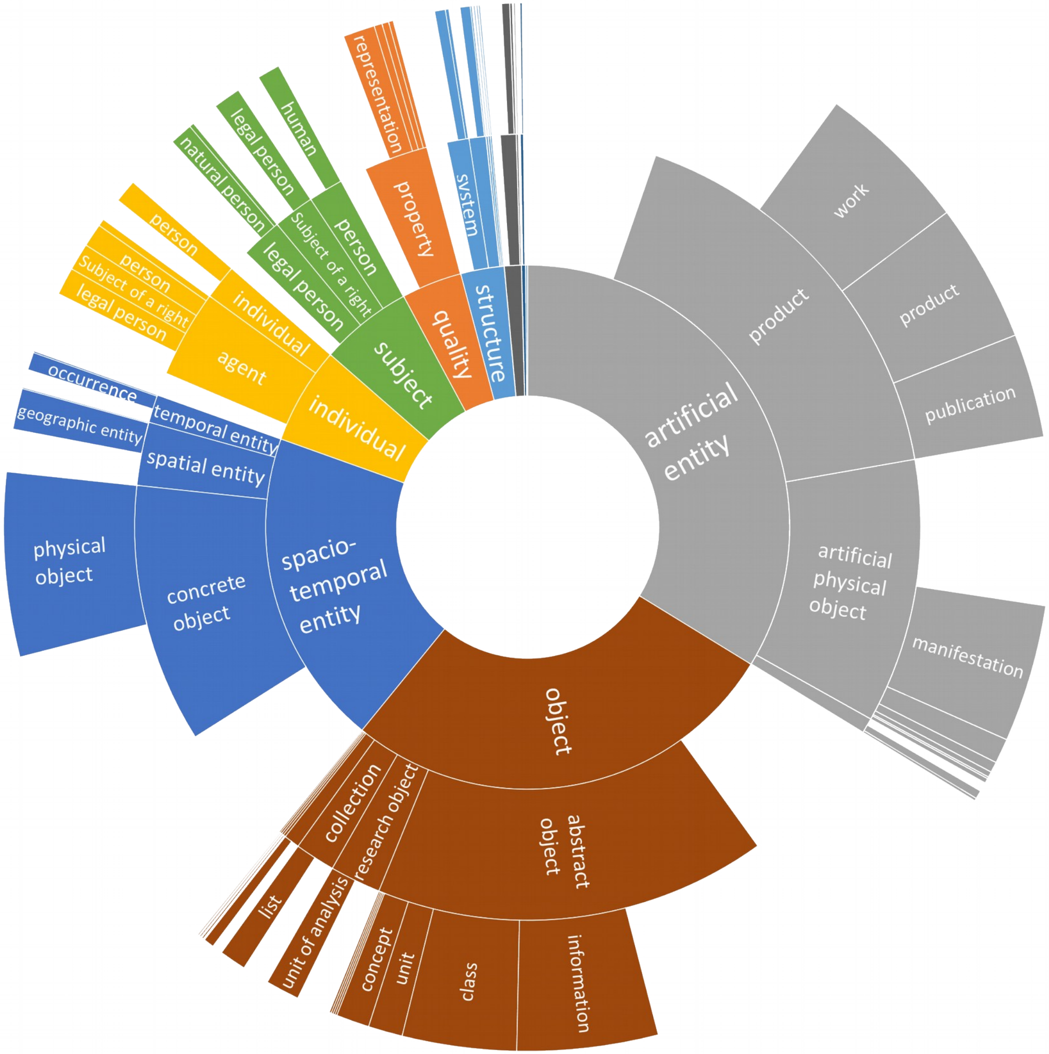
employer — University of Mannheim
1 reference
reference URL: http://dws.informatik.uni-mannheim.de/en/people/professors/dr-heiko-paulheim/

academic degree — Doctor of Science
0 references

participant of — 9th Extended Semantic Web Conference
1 reference
reference URL: http://videolectures.net/eswc2012_paulheim_linked_data/
retrieved: 31 October 2017

official website — http://www.heikopaulheim.com/
0 references

# Wikidata


artificial entity
product
work
product
publication
artificial physical object
manifestation
object
abstract object
information
class
unit
concept
research object
unit of analysis
collection
list
spacio-temporal entity
concrete object
physical object
spatial entity
geographic entity
temporal entity
occurrence
individual
agent
individual
person
person
Subject of a right
legal person
subject
legal person
Subject of a right
person
natural person
legal person
human
quality
property
representation
structure
system

# Knowledge Graph Creation

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

Max. simplicity

Max. support for plurality

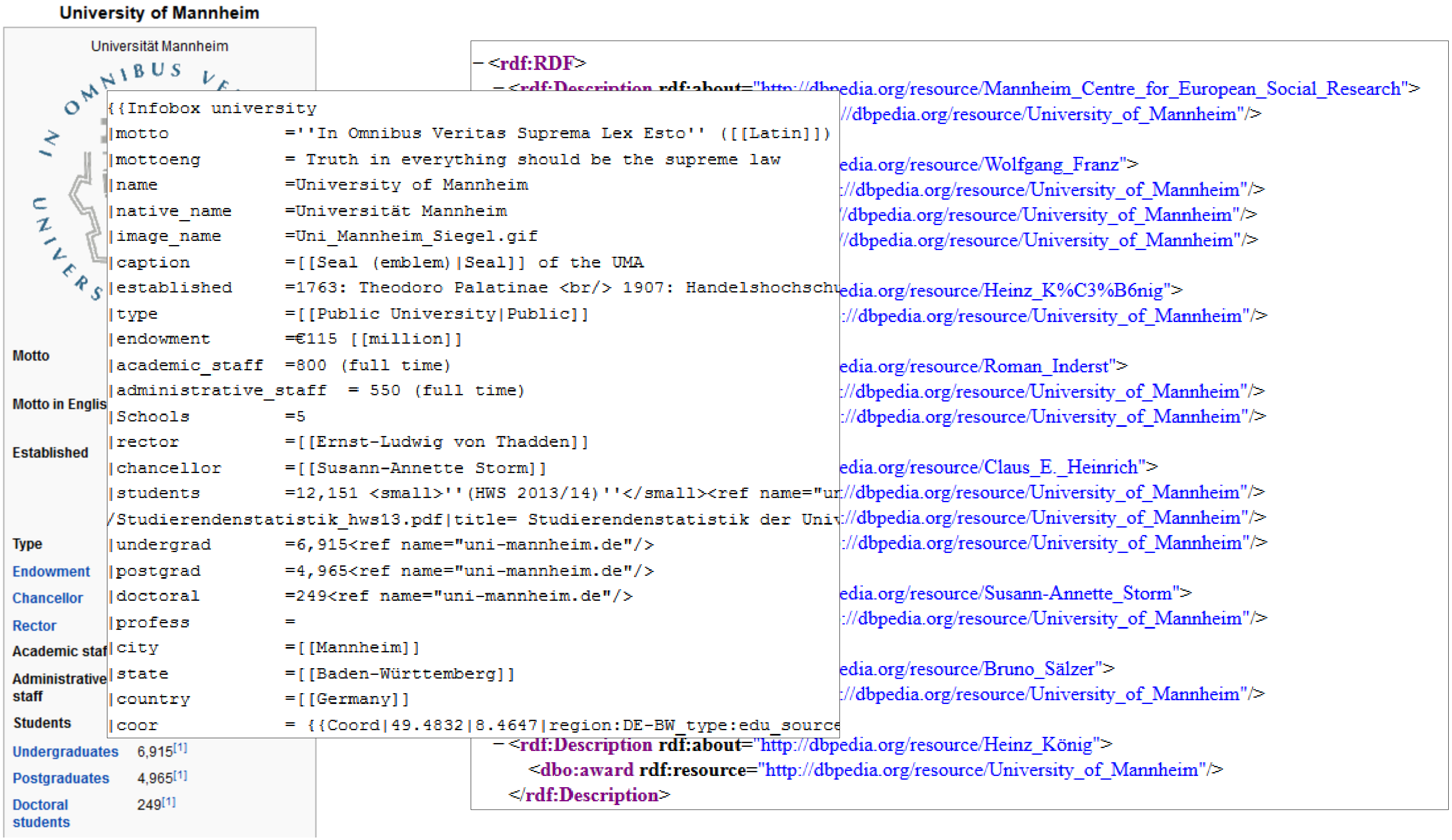
# Knowledge Graph Creation: DBpedia & YAGO

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

- Present
  - Two of the most used knowledge graphs
  - ...with Wikidata catching up

yago
select knowledge
DBpedia

# DBpedia



University of Mannheim

Universität Mannheim

Motto

Motto in Englis

Established

Type

Endowment

Chancellor

Rector

Academic staf

Administrative staff

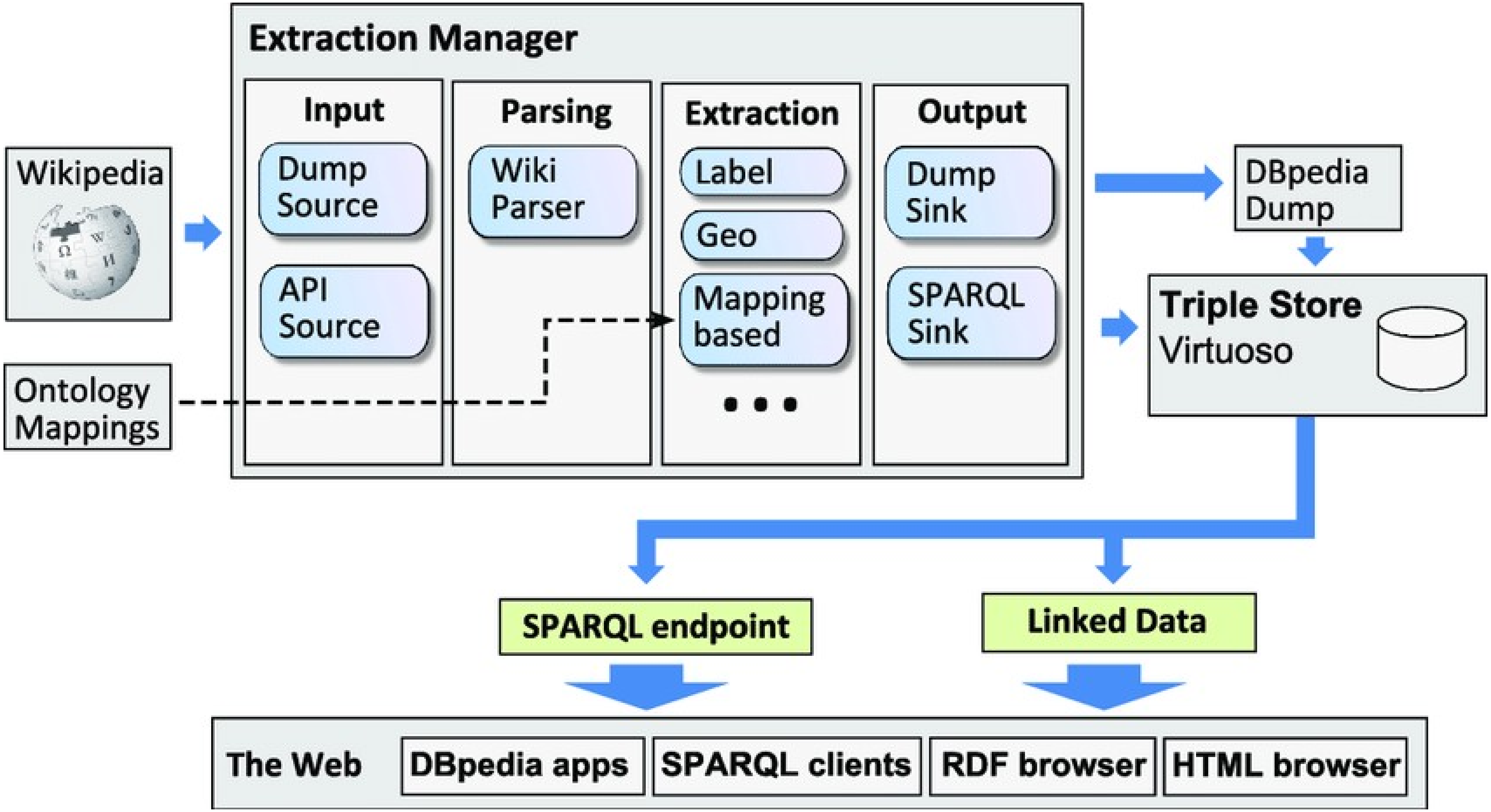
Students

| | |
|---|---|
| Undergraduates | 6,915[1] |
| Postgraduates | 4,965[1] |
| Doctoral students | 249[1] |

```
{{Infobox university
|motto            =''In Omnibus Veritas Suprema Lex Esto'' ([[Latin]])
|mottoeng         = Truth in everything should be the supreme law
|name             =University of Mannheim
|native_name      =Universität Mannheim
|image_name       =Uni_Mannheim_Siegel.gif
|caption          =[[Seal (emblem)|Seal]] of the UMA
|established      =1763: Theodoro Palatinae <br/> 1907: Handelshochschu
|type             =[[Public University|Public]]
|endowment        =€115 [[million]]
|academic_staff  =800 (full time)
|administrative_staff  = 550 (full time)
|Schools          =5
|rector           =[[Ernst-Ludwig von Thadden]]
|chancellor       =[[Susann-Annette Storm]]
|students         =12,151 <small>''(HWS 2013/14)''</small><ref name="ur
/Studierendenstatistik_hws13.pdf|title= Studierendenstatistik der Univ
|undergrad        =6,915<ref name="uni-mannheim.de"/>
|postgrad         =4,965<ref name="uni-mannheim.de"/>
|doctoral         =249<ref name="uni-mannheim.de"/>
|profess          =
|city             =[[Mannheim]]
|state            =[[Baden-Württemberg]]
|country          =[[Germany]]
|coor             = {{Coord|49.4832|8.4647|region:DE-BW_type:edu_source
```

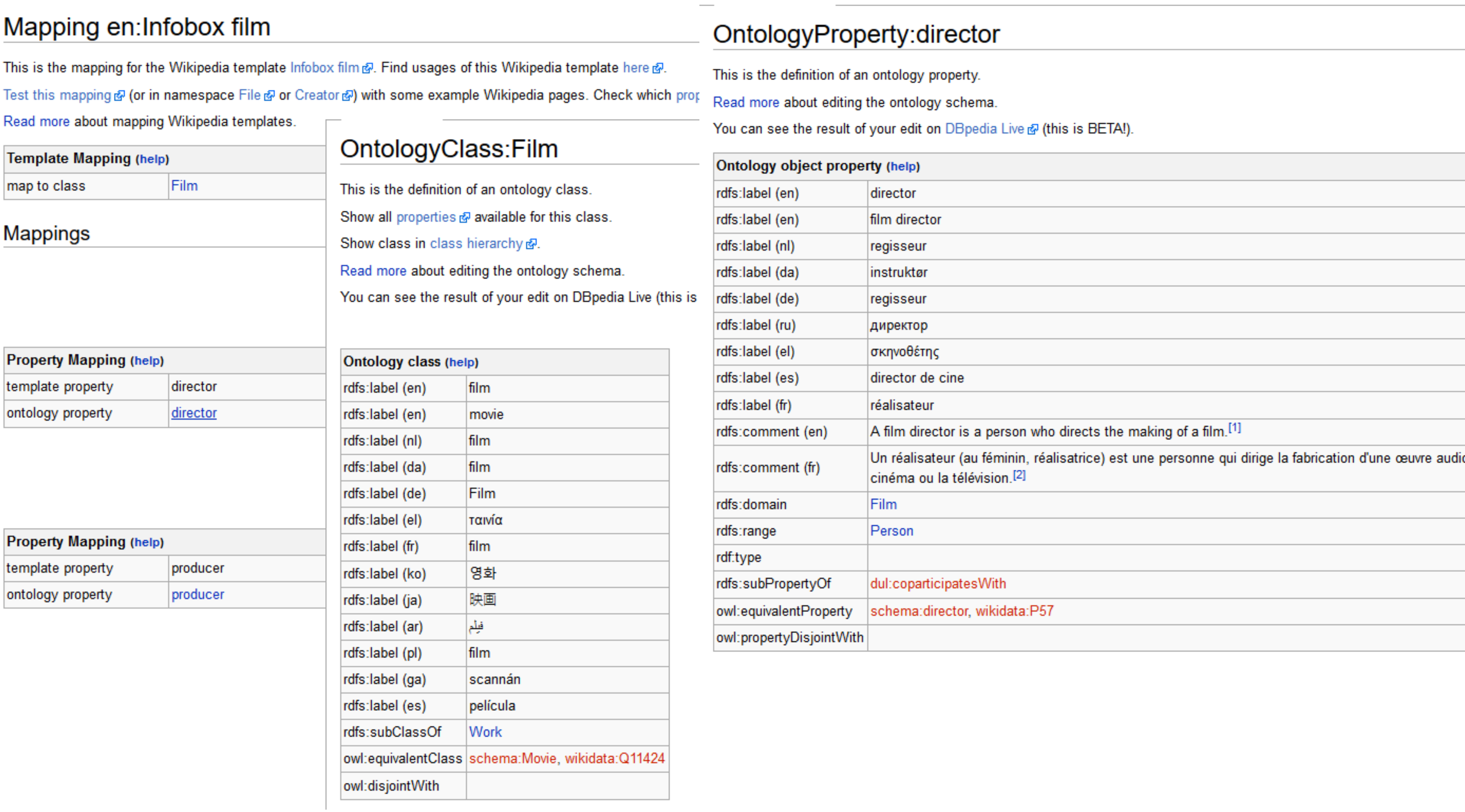
```
– <rdf:RDF>
  – <rdf:Description rdf:about="http://dbpedia.org/resource/Mannheim_Centre_for_European_Social_Research">
                    //dbpedia.org/resource/University_of_Mannheim"/>

                    edia.org/resource/Wolfgang_Franz">
                    ://dbpedia.org/resource/University_of_Mannheim"/>
                    //dbpedia.org/resource/University_of_Mannheim"/>
                    //dbpedia.org/resource/University_of_Mannheim"/>

                    edia.org/resource/Heinz_K%C3%B6nig">
                    ://dbpedia.org/resource/University_of_Mannheim"/>

                    edia.org/resource/Roman_Inderst">
                    ://dbpedia.org/resource/University_of_Mannheim"/>
                    ://dbpedia.org/resource/University_of_Mannheim"/>

                    edia.org/resource/Claus_E._Heinrich">
                    ://dbpedia.org/resource/University_of_Mannheim"/>
                    ://dbpedia.org/resource/University_of_Mannheim"/>
                    ://dbpedia.org/resource/University_of_Mannheim"/>

                    edia.org/resource/Susann-Annette_Storm">
                    ://dbpedia.org/resource/University_of_Mannheim"/>

                    edia.org/resource/Bruno_Sälzer">
                    ://dbpedia.org/resource/University_of_Mannheim"/>
  – <rdf:Description rdf:about="http://dbpedia.org/resource/Heinz_König">
      <dbo:award rdf:resource="http://dbpedia.org/resource/University_of_Mannheim"/>
    </rdf:Description>
```

# DBpedia



Lehmann et al.: *DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia.* 2014

# DBpedia



## 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 prop

Read more about mapping Wikipedia templates.

| Template Mapping (help) | |
|---|---|
| map to class | Film |

### Mappings

| Property Mapping (help) | |
|---|---|
| template property | director |
| ontology property | director |

| Property Mapping (help) | |
|---|---|
| template property | producer |
| ontology property | producer |

## 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

| Ontology class (help) | |
|---|---|
| rdfs:label (en) | film |
| rdfs:label (en) | movie |
| rdfs:label (nl) | film |
| rdfs:label (da) | film |
| rdfs:label (de) | Film |
| rdfs:label (el) | ταινία |
| rdfs:label (fr) | film |
| rdfs:label (ko) | 영화 |
| rdfs:label (ja) | 映画 |
| rdfs:label (ar) | فيلم |
| rdfs:label (pl) | film |
| rdfs:label (ga) | scannán |
| rdfs:label (es) | película |
| rdfs:subClassOf | Work |
| owl:equivalentClass | schema:Movie, wikidata:Q11424 |
| owl:disjointWith | |

## OntologyProperty:director

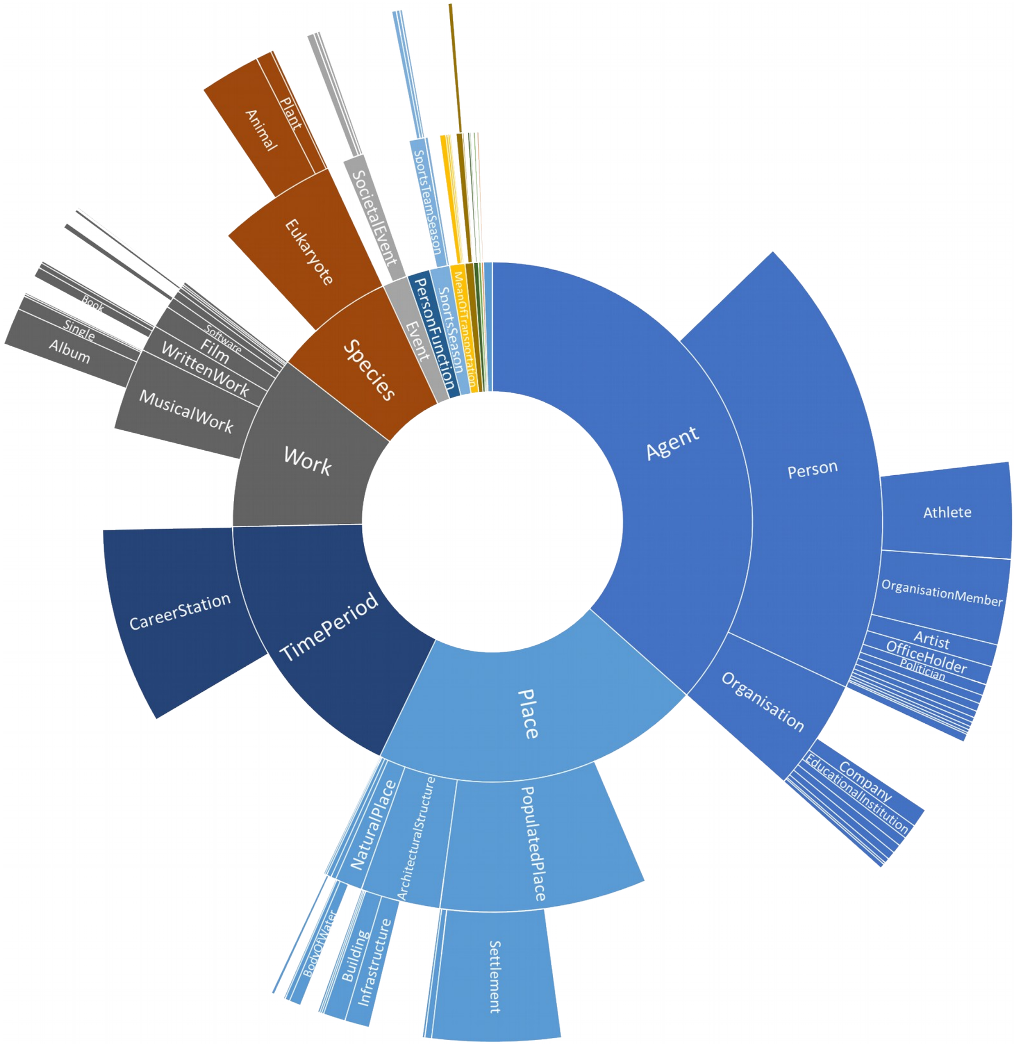
This is the definition of an ontology property.

Read more about editing the ontology schema.

You can see the result of your edit on DBpedia Live (this is BETA!).

| Ontology object property (help) | |
|---|---|
| rdfs:label (en) | director |
| rdfs:label (en) | film director |
| rdfs:label (nl) | regisseur |
| rdfs:label (da) | instruktør |
| rdfs:label (de) | regisseur |
| rdfs:label (ru) | директор |
| rdfs:label (el) | σκηνοθέτης |
| rdfs:label (es) | director de cine |
| rdfs:label (fr) | réalisateur |
| rdfs:comment (en) | A film director is a person who directs the making of a film.[1] |
| rdfs:comment (fr) | Un réalisateur (au féminin, réalisatrice) est une personne qui dirige la fabrication d'une œuvre audi cinéma ou la télévision.[2] |
| rdfs:domain | Film |
| rdfs:range | Person |
| rdf:type | |
| rdfs:subPropertyOf | dul:coparticipatesWith |
| owl:equivalentProperty | schema:director, wikidata:P57 |
| owl:propertyDisjointWith | |

# DBpedia


Agent
Person
Athlete
OrganisationMember
Artist
OfficeHolder
Politician
Organisation
Company
EducationalInstitution
Place
PopulatedPlace
Settlement
ArchitecturalStructure
NaturalPlace
BodyOfWater
Building
Infrastructure
TimePeriod
CareerStation
Work
MusicalWork
Album
Single
WrittenWork
Book
Film
Software
Species
Eukaryote
Animal
Plant
Event
SocietalEvent
PersonFunction
SportsSeason
SportsTeamSeason
MeanOfTransportation

# YAGO

- Wikipedia categories for types
  - Plus WordNet as upper structure

- Manual mappings for properties

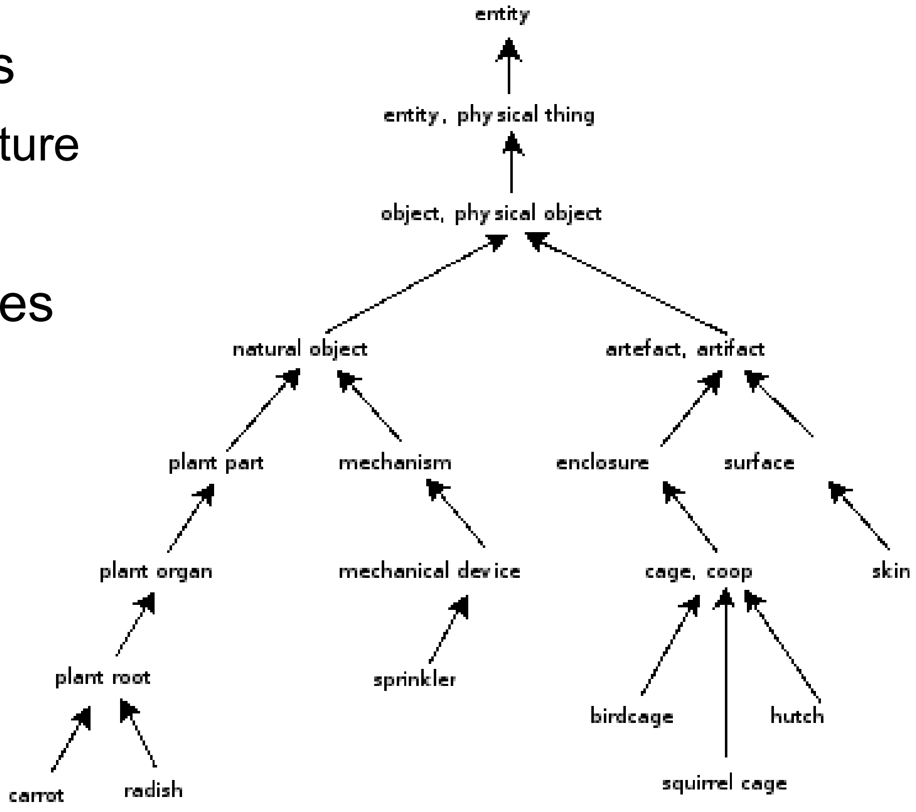
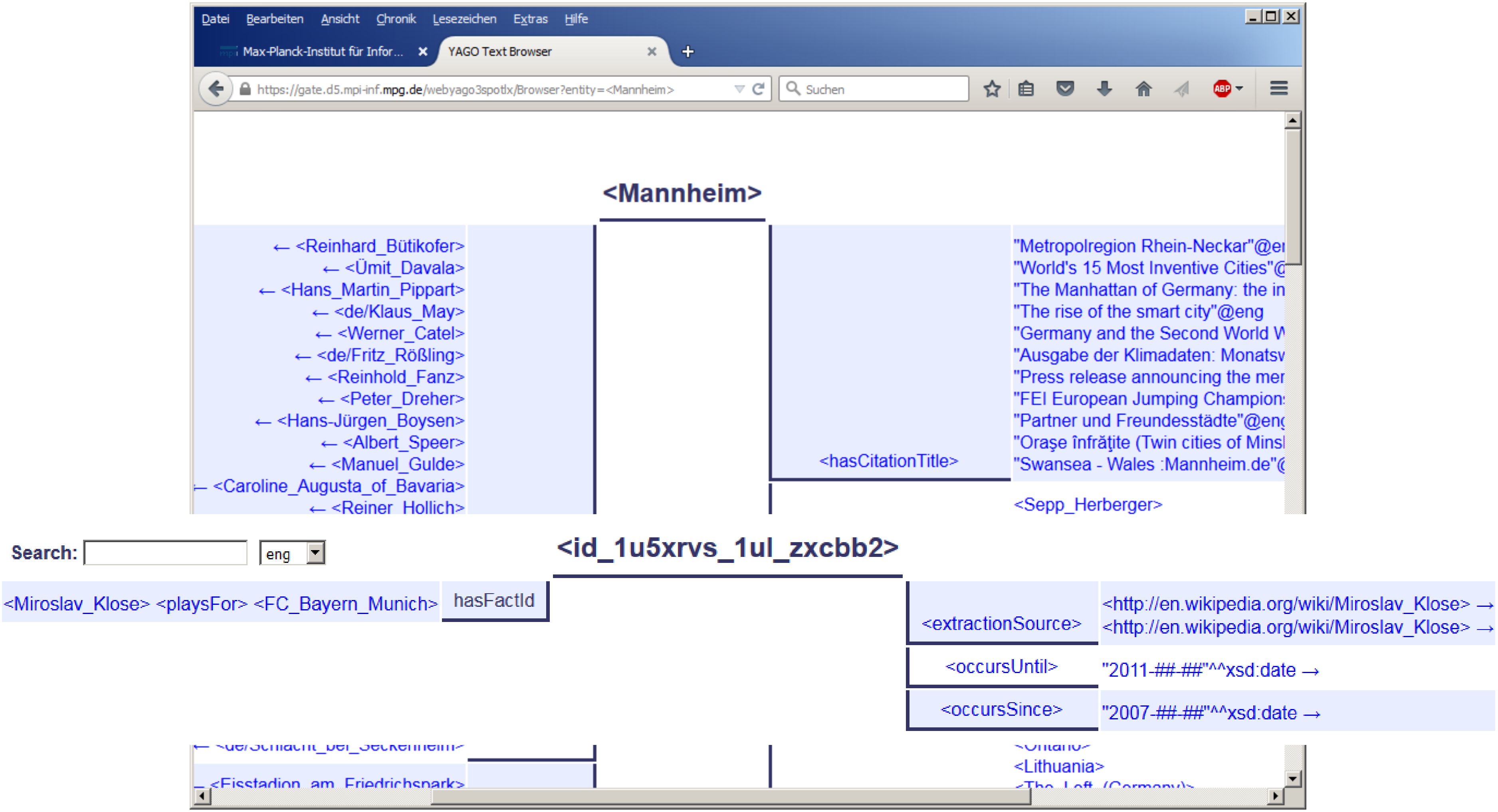


Figure 1. "is a" relation example

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

# YAGO


Datei
Bearbeiten
Ansicht
Chronik
Lesezeichen
Extras
Hilfe
Max-Planck-Institut für Infor...
YAGO Text Browser
https://gate.d5.mpi-inf.mpg.de/webyago3spotlx/Browser?entity=<Mannheim>
Suchen

<Mannheim>

← <Reinhard_Bütikofer>
← <Ümit_Davala>
← <Hans_Martin_Pippart>
← <de/Klaus_May>
← <Werner_Catel>
← <de/Fritz_Rößling>
← <Reinhold_Fanz>
← <Peter_Dreher>
← <Hans-Jürgen_Boysen>
← <Albert_Speer>
← <Manuel_Gulde>
← <Caroline_Augusta_of_Bavaria>
← <Reiner_Hollich>

<hasCitationTitle>

"Metropolregion Rhein-Neckar"@er
"World's 15 Most Inventive Cities"@
"The Manhattan of Germany: the in
"The rise of the smart city"@eng
"Germany and the Second World W
"Ausgabe der Klimadaten: Monatsv
"Press release announcing the mer
"FEI European Jumping Champion:
"Partner und Freundesstädte"@eng
"Oraşe înfrăţite (Twin cities of Minsl
"Swansea - Wales :Mannheim.de"@

<Sepp_Herberger>

Search: eng

<id_1u5xrvs_1ul_zxcbb2>

<Miroslav_Klose> <playsFor> <FC_Bayern_Munich> hasFactId

<extractionSource>
<http://en.wikipedia.org/wiki/Miroslav_Klose> →
<http://en.wikipedia.org/wiki/Miroslav_Klose> →

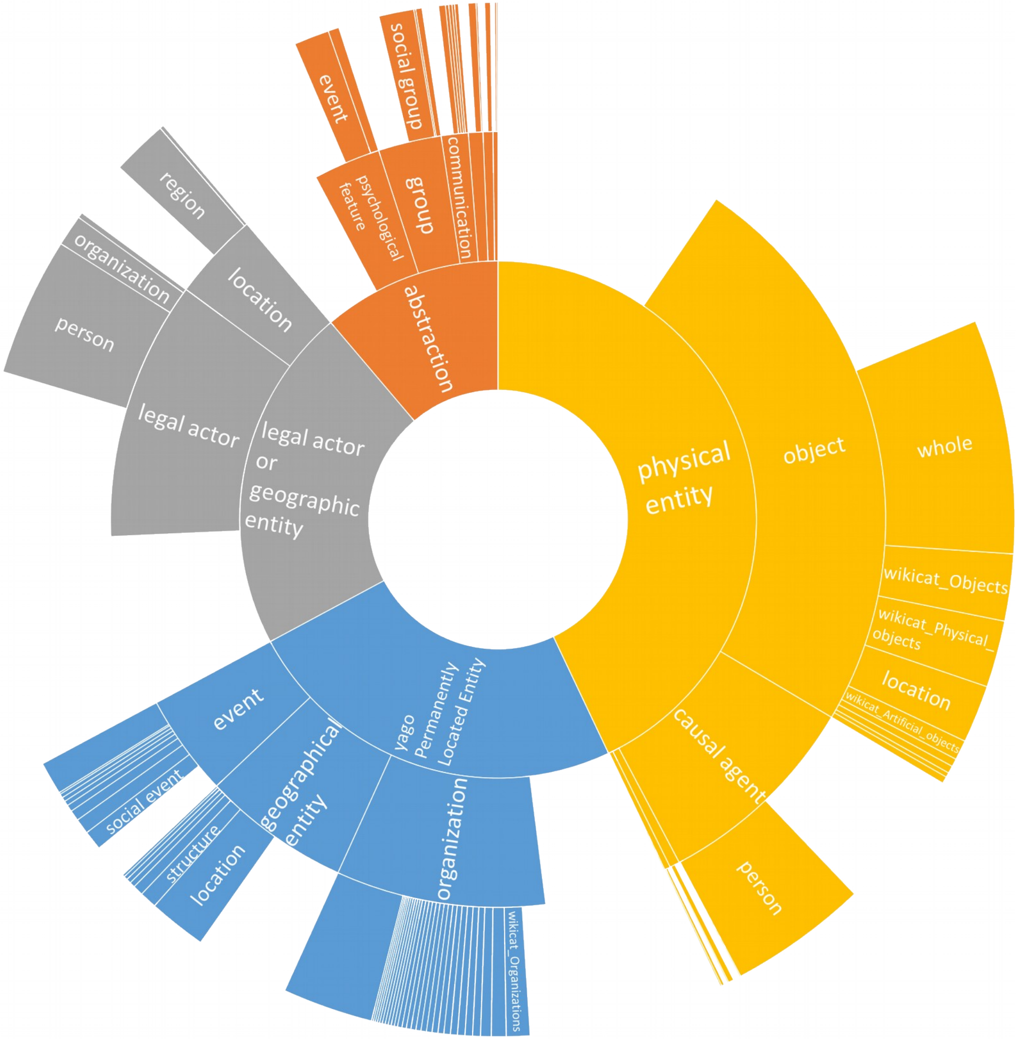
<occursUntil> "2011-##-##"^^xsd:date →

<occursSince> "2007-##-##"^^xsd:date →

← <de/Schlacht_bei_Seckenheim>
← <Eisstadion_am_Friedrichspark>

<Ontario>
<Lithuania>
<The_Loft_(Germany)>

# YAGO


event
social group
psychological feature
group
communication
abstraction
region
organization
location
person
legal actor
legal actor or geographic entity
physical entity
object
whole
wikicat_Objects
wikicat_Physical_objects
location
wikicat_Artificial_objects
causal agent
person
yago Permanently Located Entity
event
social event
geographical entity
structure
location
organization
wikicat_Organizations

# Knowledge Graph Creation

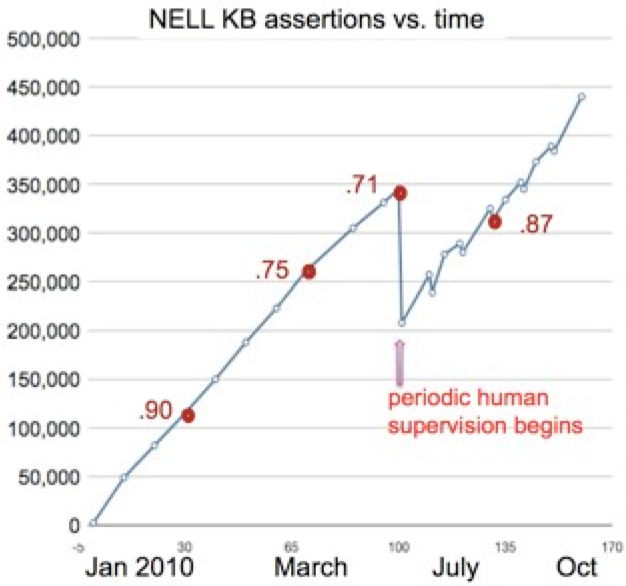
- Lesson learned no. 4:
  - Heuristics help increasing coverage (at the cost of accuracy)

Max. accuracy

Max. coverage

# 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

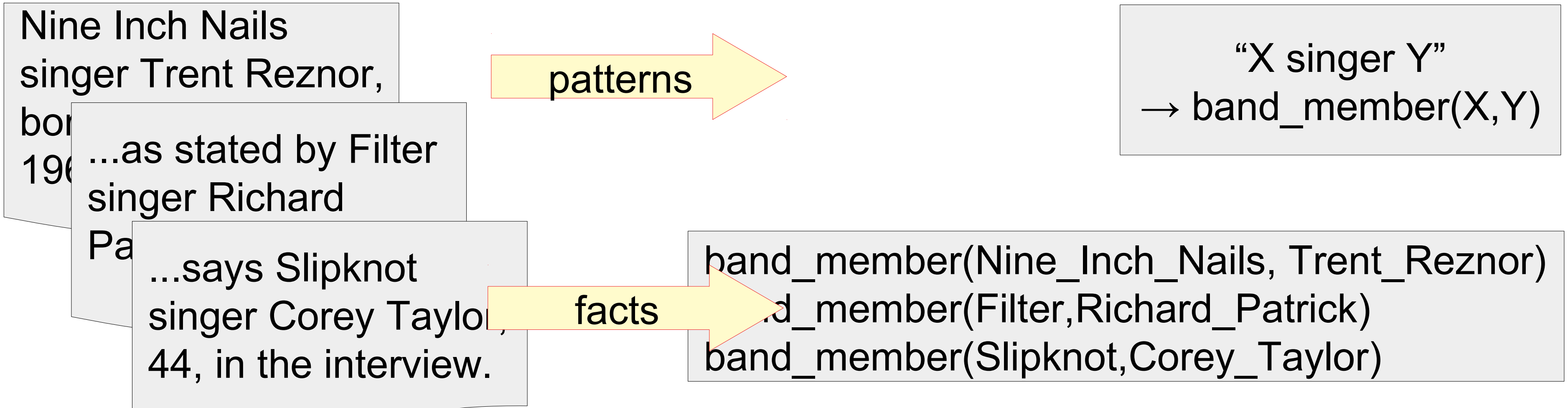


- Until ~one year ago
  - Still running
  - New release every few days

http://rtw.ml.cmu.edu/rtw/overview

# Knowledge Graph Creation: NELL

- Extraction of a Knowledge Graph from a Text Corpus



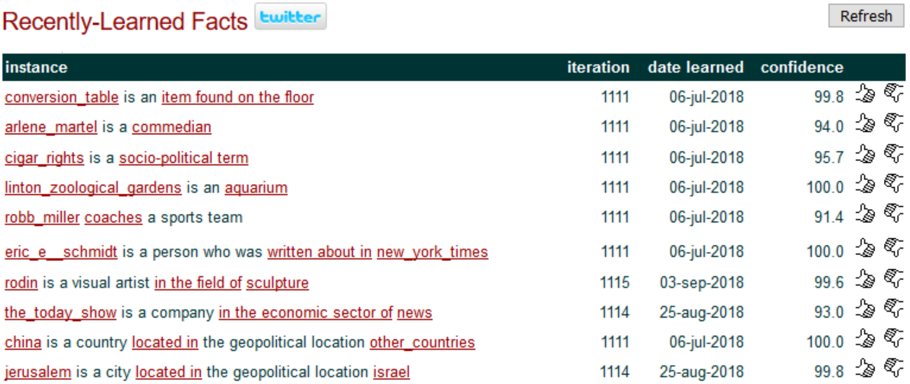
Nine Inch Nails singer Trent Reznor, bor 196

...as stated by Filter singer Richard Pa

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

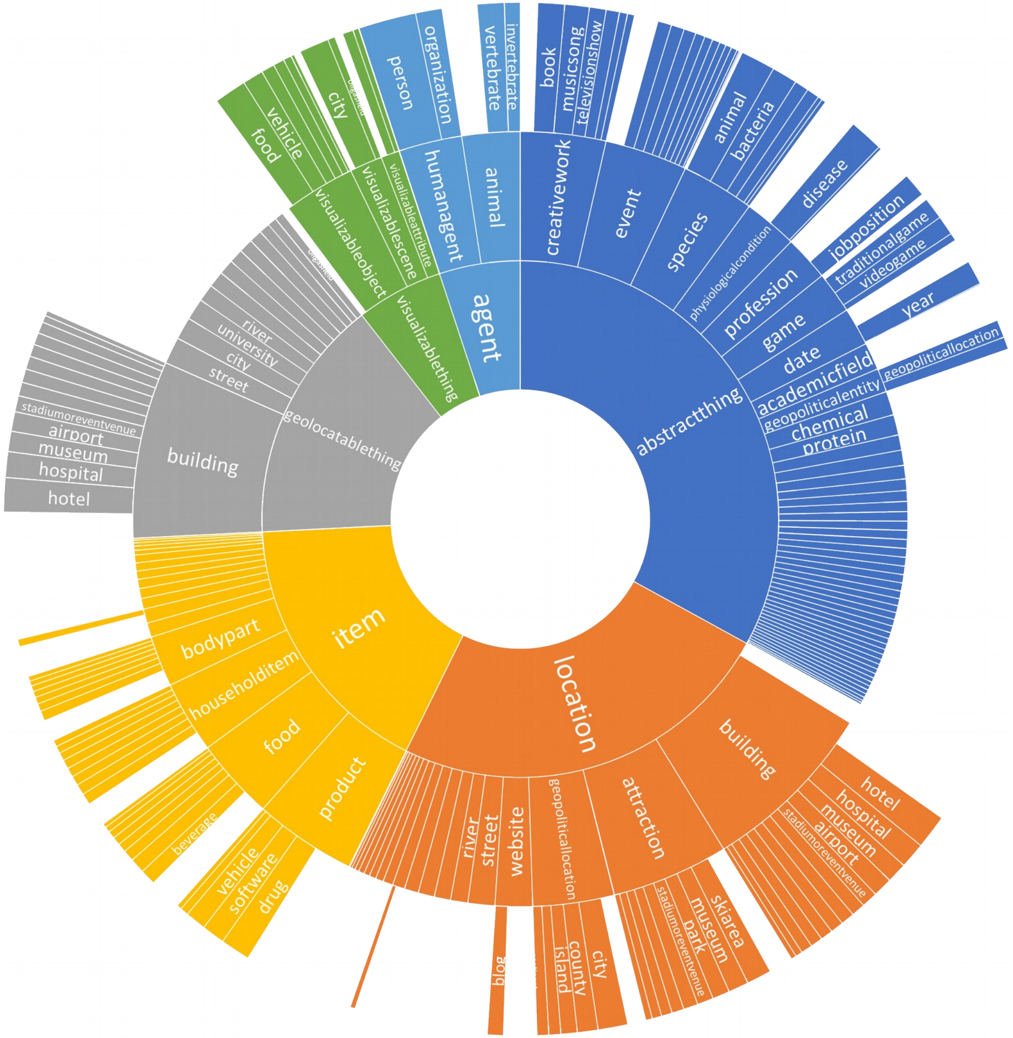
"X singer Y"
→ band_member(X,Y)

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



| instance | iteration | date learned | confidence |
|---|---|---|---|
| conversion_table is an item found on the floor | 1111 | 06-jul-2018 | 99.8 |
| arlene_martel is a commedian | 1111 | 06-jul-2018 | 94.0 |
| cigar_rights is a socio-political term | 1111 | 06-jul-2018 | 95.7 |
| linton_zoological_gardens is an aquarium | 1111 | 06-jul-2018 | 100.0 |
| robb_miller coaches a sports team | 1111 | 06-jul-2018 | 91.4 |
| eric_e__schmidt is a person who was written about in new_york_times | 1111 | 06-jul-2018 | 100.0 |
| rodin is a visual artist in the field of sculpture | 1115 | 03-sep-2018 | 99.6 |
| the_today_show is a company in the economic sector of news | 1114 | 25-aug-2018 | 93.0 |
| china is a country located in the geopolitical location other_countries | 1111 | 06-jul-2018 | 100.0 |
| jerusalem is a city located in the geopolitical location israel | 1114 | 25-aug-2018 | 99.8 |

# 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!

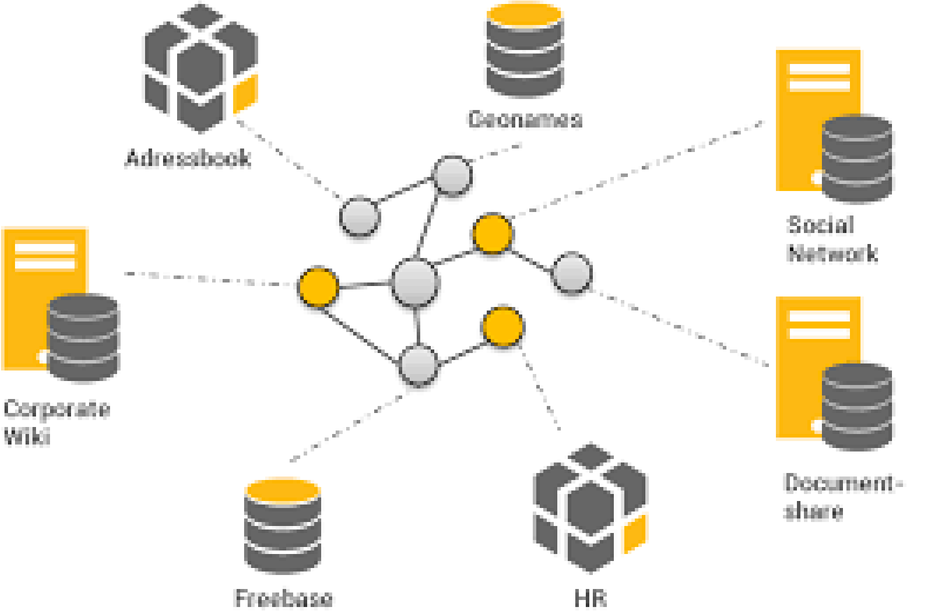
# Non-Public Knowledge Graphs

- 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

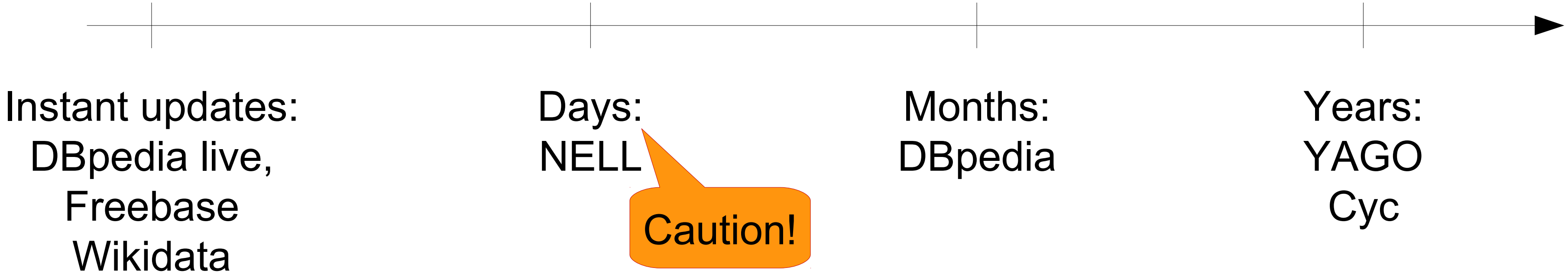
- Knowledge Graphs are used…
- …in companies and organizations
  - collect, organize, and integrate knowledge
  - link isolated information sources
  - make information searchable and findable



Masuch, 2014

# Comparison of Knowledge Graphs

- Release cycles



Instant updates:
DBpedia live,
Freebase
Wikidata

Days:
NELL

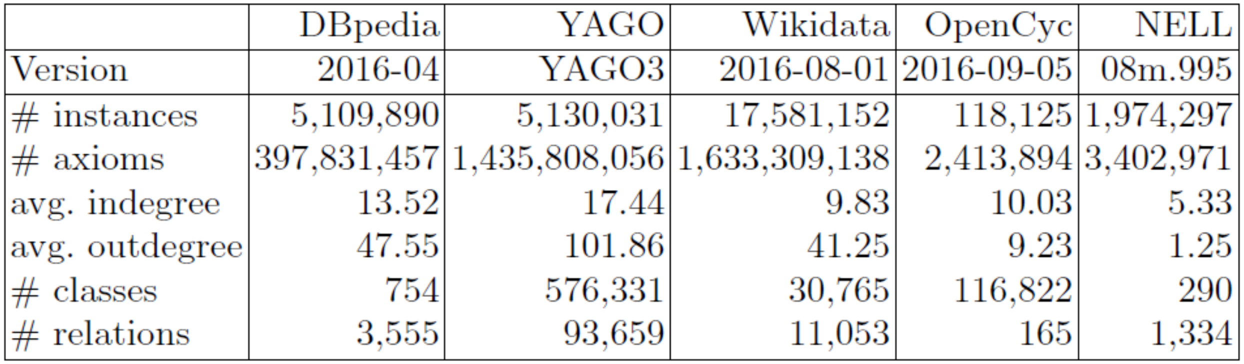
Caution!

Months:
DBpedia

Years:
YAGO
Cyc

- Size and density

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



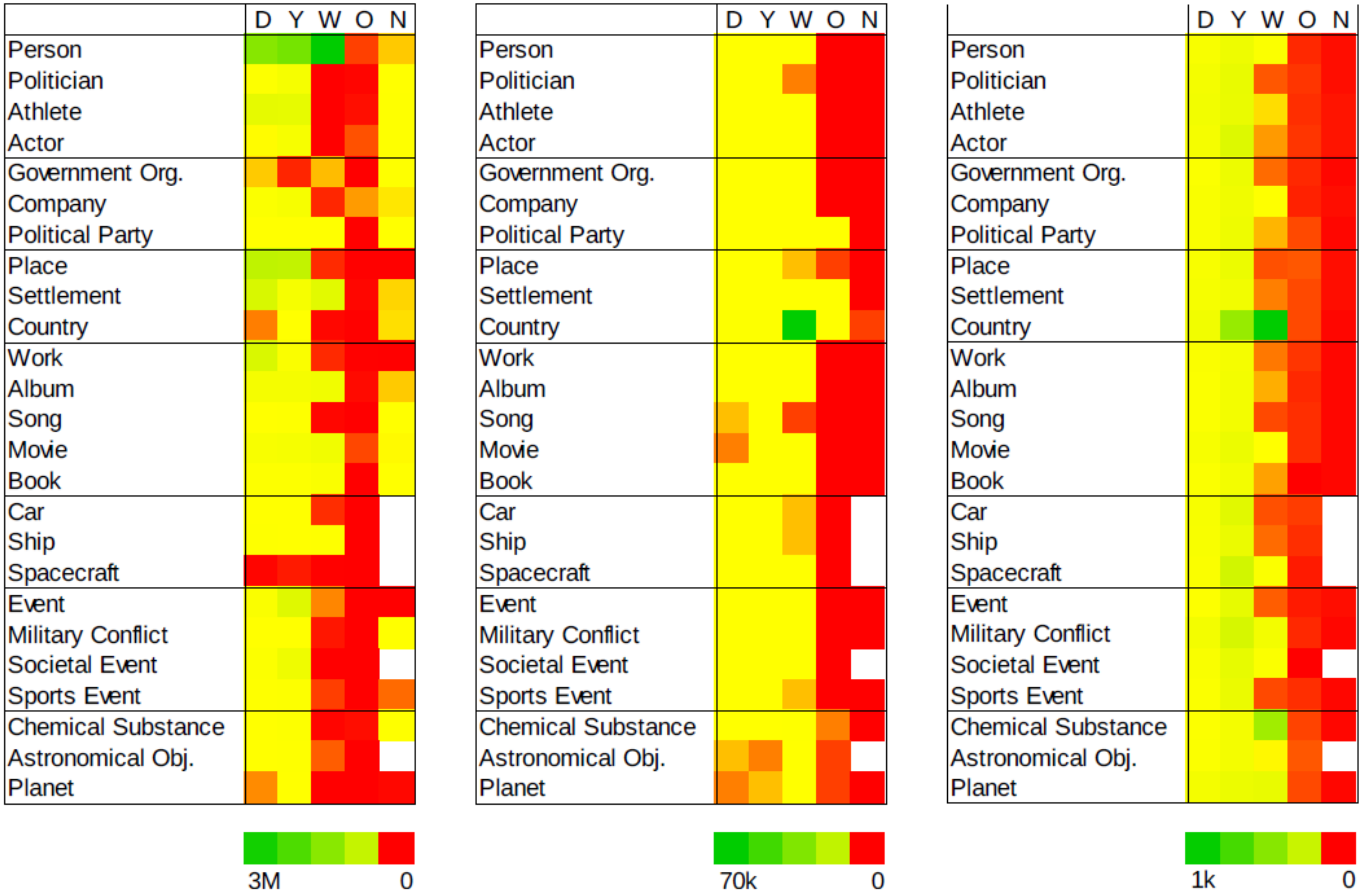
| | DBpedia | YAGO | Wikidata | OpenCyc | NELL |
|---|---|---|---|---|---|
| Version | 2016-04 | YAGO3 | 2016-08-01 | 2016-09-05 | 08m.995 |
| # instances | 5,109,890 | 5,130,031 | 17,581,152 | 118,125 | 1,974,297 |
| # axioms | 397,831,457 | 1,435,808,056 | 1,633,309,138 | 2,413,894 | 3,402,971 |
|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

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

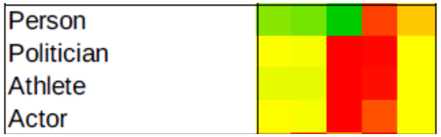
# Comparison of Knowledge Graphs

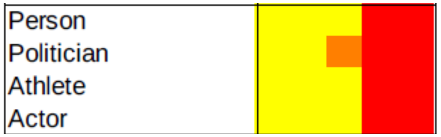
- Summary findings:
  - Persons: more in Wikidata
    (twice as many persons as DBpedia and YAGO)
  - Countries: more details in Wikidata
  - Places: most in DBpedia
  - Organizations: most in YAGO
  - Events: most in YAGO
  - Artistic works:
    - Wikidata contains more movies and albums
    - YAGO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

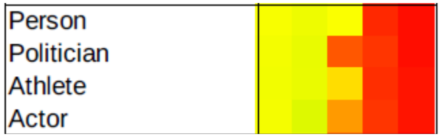
Ringler & Paulheim: *One Knowledge Graph to Rule them All?* KI 2017

# Caveats

- Reading the diagrams right…





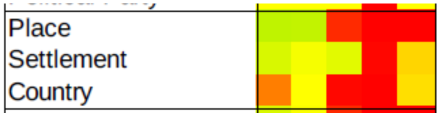


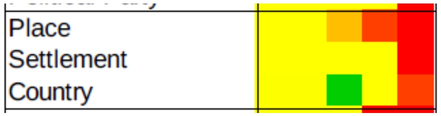
- So, Wikidata contains more persons
  - but less instances of all the interesting subclasses?

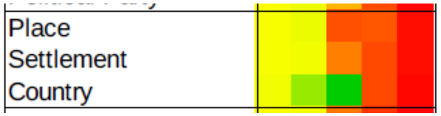
- There are classes like *Actor* in Wikidata
  - but they are hardly used
  - rather: modeled using *profession* relation

# Caveats

- Reading the diagrams right… (ctd.)



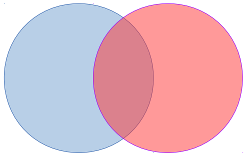


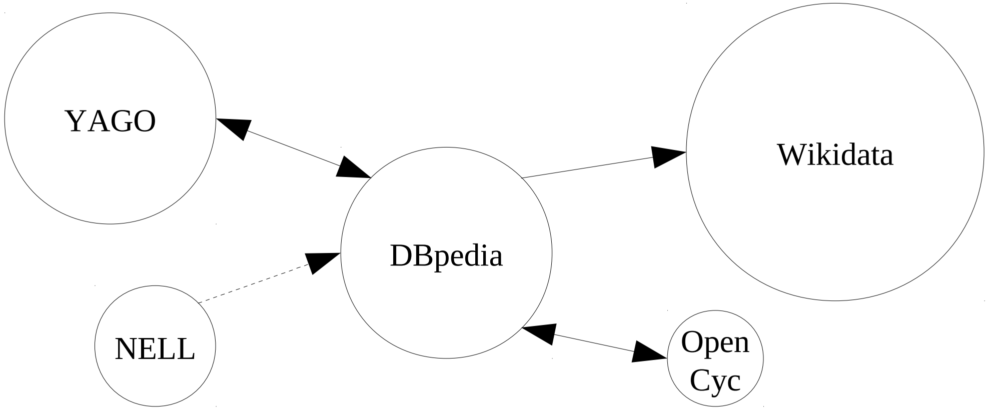


- So, Wikidata contains more data on countries, but less countries?
- First: Wikidata only counts current, actual countries
  - DBpedia and YAGO also count historical countries
- “KG1 contains less of X than KG2” can mean
  - it actually contains less instances of X
  - it contains equally many or more 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

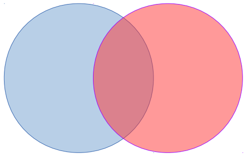




Ringler & Paulheim: *One Knowledge Graph to Rule them All?* KI 2017

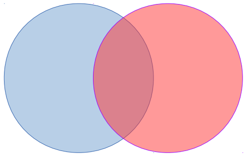
# Overlap of Knowledge Graphs

- 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

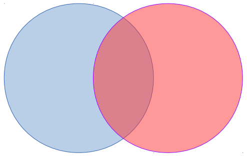


- Links between Knowledge Graphs are incomplete
  - The Open World Assumption also holds for interlinks

- 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



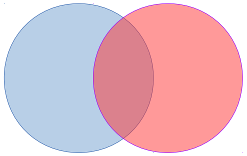
- Idea:
  - Given that the link set F is found
  - And the (unknown) actual link set would be C

- 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 $|C| = |F| \cdot P \cdot \frac{1}{R}$

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

# Overlap of Knowledge Graphs

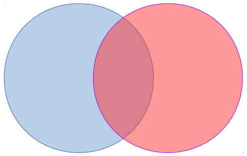


- Mathematical derivation:
  - Definition of recall: $R=\frac{|F_{correct}|}{|C|}$
  - Definition of precision: $P=\frac{|F_{correct}|}{|F|}$
- Resolve both to $|F_{correct}|$ , substitute, and resolve to $|C|$

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

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

# Overlap of Knowledge Graphs

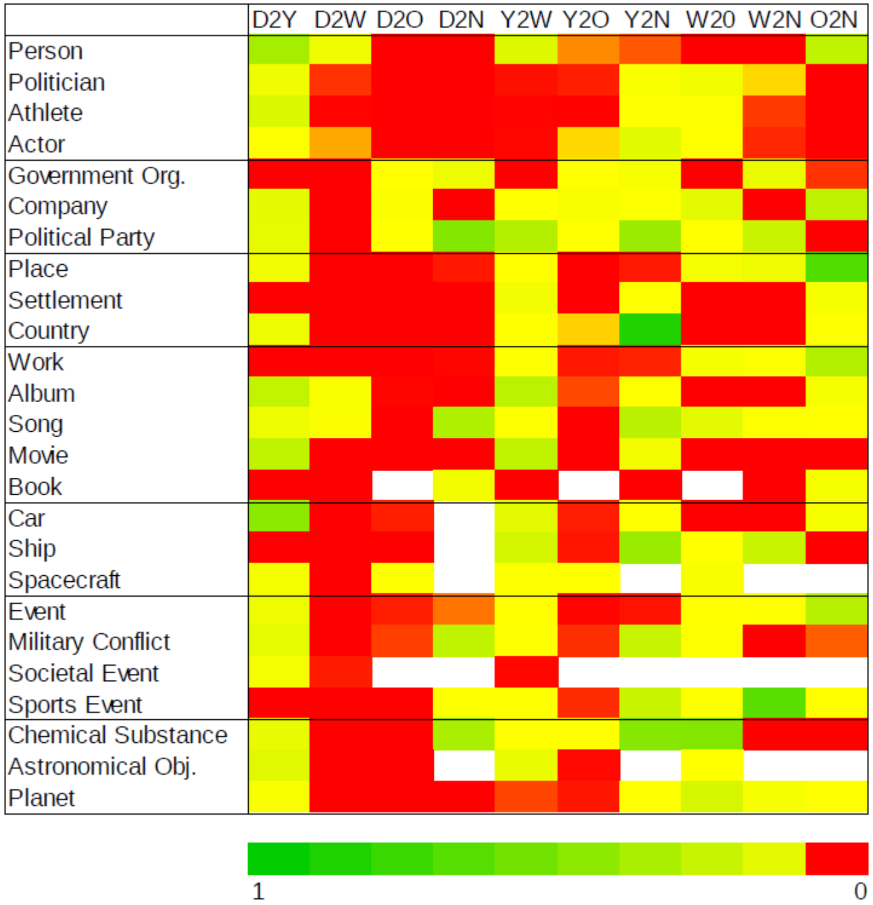


- Experiment:
  - We use the same 25 classes as before
  - Measure 1: overlap relative to smaller KG (i.e., potential gain)
  - Measure 2: overlap relative to explicit links (i.e., importance of improving links)

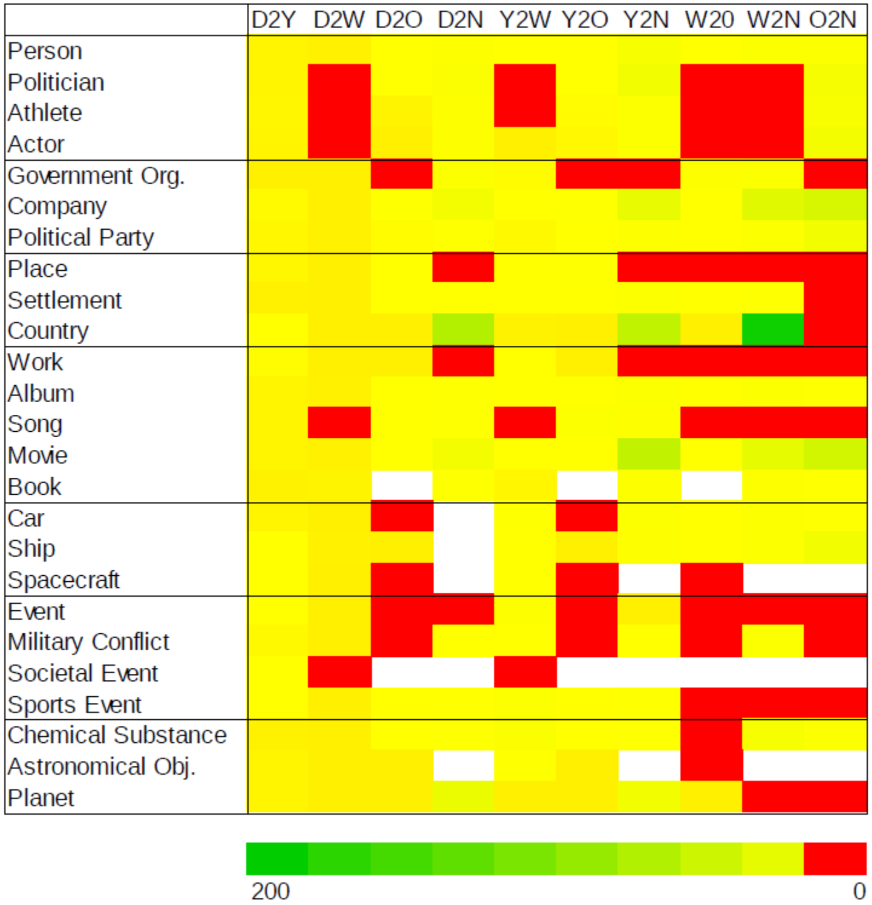
- Link generation with 16 different metrics and thresholds
  - Intra-class correlation coefficient for |C|: 0.969
  - Intra-class correlation coefficient for |F|: 0.646
- Bottom line:
  - Despite variety in link sets generated, the overlap is estimated reliably
  - The link generation mechanisms do not need to be overly accurate

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

# Overlap of Knowledge Graphs



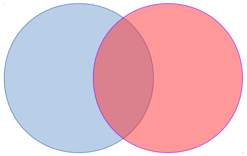
(a) Overlap as potential gain



(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, ...

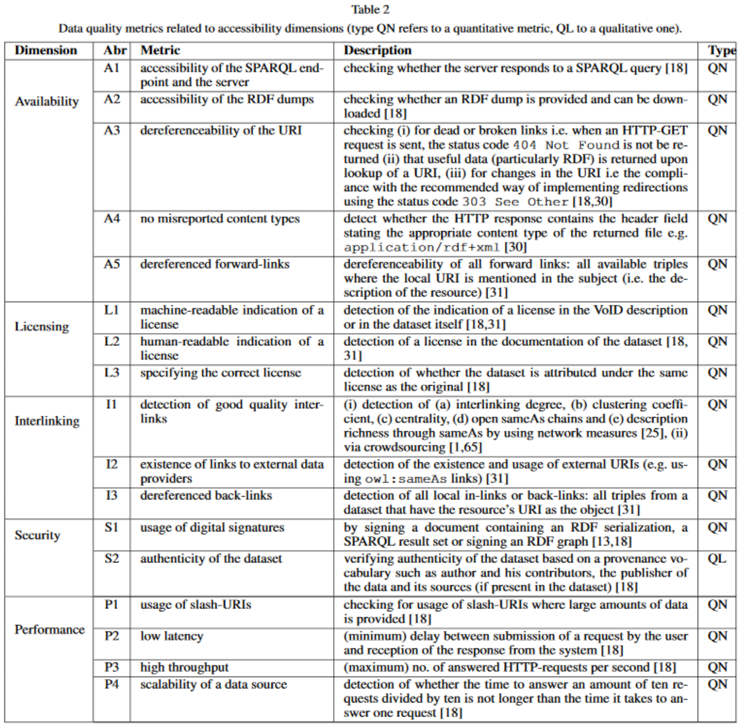


Table 2

Data quality metrics related to accessibility dimensions (type QN refers to a quantitative metric, QL to a qualitative one).

| Dimension | Abr | Metric | Description | Type |
|---|---|---|---|---|
| Availability | A1 | accessibility of the SPARQL end-point and the server | checking whether the server responds to a SPARQL query [18] | QN |
| | A2 | accessibility of the RDF dumps | checking whether an RDF dump is provided and can be downloaded [18] | QN |
| | A3 | dereferenceability of the URI | 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 returned (ii) that useful data (particularly RDF) 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] | QN |
| | A4 | no misreported content types | detect whether the HTTP response contains the header field stating the appropriate content type of the returned file e.g. `application/rdf+xml` [30] | QN |
| | A5 | dereferenced forward-links | dereferenceability of all forward links: all available triples where the local URI is mentioned in the subject (i.e. the description of the resource) [31] | QN |
| Licensing | L1 | machine-readable indication of a license | detection of the indication of a license in the VoID description or in the dataset itself [18,31] | QN |
| | L2 | human-readable indication of a license | detection of a license in the documentation of the dataset [18, 31] | QN |
| | L3 | specifying the correct license | detection of whether the dataset is attributed under the same license as the original [18] | QN |
| Interlinking | I1 | detection of good quality inter-links | (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] | QN |
| | I2 | existence of links to external data providers | detection of the existence and usage of external URIs (e.g. using `owl:sameAs` links) [31] | QN |
| | I3 | dereferenced back-links | detection of all local in-links or back-links: all triples from a dataset that have the resource's URI as the object [31] | QN |
| Security | S1 | usage of digital signatures | by signing a document containing an RDF serialization, a SPARQL result set or signing an RDF graph [13,18] | QN |
| | S2 | authenticity of the dataset | 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] | QL |
| Performance | P1 | usage of slash-URIs | checking for usage of slash-URIs where large amounts of data is provided [18] | QN |
| | P2 | low latency | (minimum) delay between submission of a request by the user and reception of the response from the system [18] | QN |
| | P3 | high throughput | (maximum) no. of answered HTTP-requests per second [18] | QN |
| | P4 | scalability of a data source | 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] | QN |

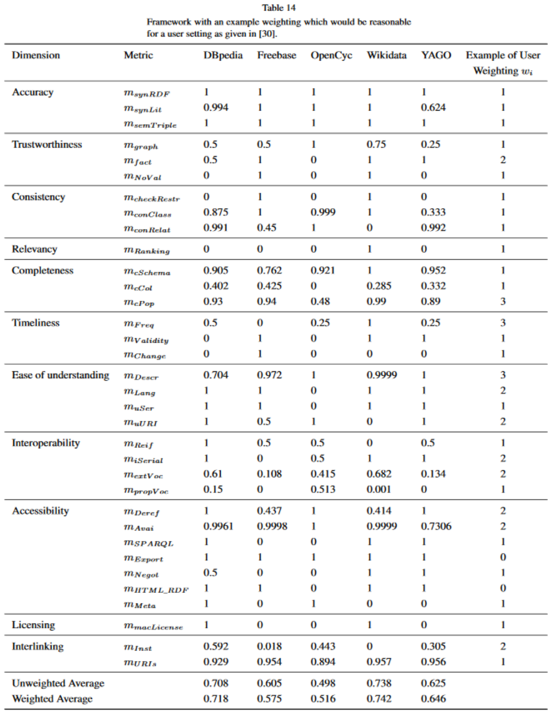


Table 14

Framework with an example weighting which would be reasonable for a user setting as given in [30].

| Dimension | Metric | DBpedia | Freebase | OpenCyc | Wikidata | YAGO | Example of User Weighting $w_i$ |
|---|---|---|---|---|---|---|---|
| Accuracy | $m_{synRDF}$ | 1 | 1 | 1 | 1 | 1 | 1 |
| | $m_{synLit}$ | 0.994 | 1 | 1 | 1 | 0.624 | 1 |
| | $m_{semTriple}$ | 1 | 1 | 1 | 1 | 1 | 1 |
| Trustworthiness | $m_{graph}$ | 0.5 | 0.5 | 1 | 0.75 | 0.25 | 1 |
| | $m_{fact}$ | 0.5 | 1 | 0 | 1 | 1 | 2 |
| | $m_{NoVal}$ | 0 | 1 | 0 | 1 | 0 | 1 |
| Consistency | $m_{checkRestr}$ | 0 | 1 | 0 | 1 | 0 | 1 |
| | $m_{conClass}$ | 0.875 | 1 | 0.999 | 1 | 0.333 | 1 |
| | $m_{conRelat}$ | 0.991 | 0.45 | 1 | 0 | 0.992 | 1 |
| Relevancy | $m_{Ranking}$ | 0 | 0 | 0 | 1 | 0 | 1 |
| Completeness | $m_{cSchema}$ | 0.905 | 0.762 | 0.921 | 1 | 0.952 | 1 |
| | $m_{cCol}$ | 0.402 | 0.425 | 0 | 0.285 | 0.332 | 1 |
| | $m_{cPop}$ | 0.93 | 0.94 | 0.48 | 0.99 | 0.89 | 3 |
| Timeliness | $m_{Freq}$ | 0.5 | 0 | 0.25 | 1 | 0.25 | 3 |
| | $m_{Validity}$ | 0 | 1 | 0 | 1 | 1 | 1 |
| | $m_{Change}$ | 0 | 1 | 0 | 0 | 0 | 1 |
| Ease of understanding | $m_{Descr}$ | 0.704 | 0.972 | 1 | 0.9999 | 1 | 3 |
| | $m_{Lang}$ | 1 | 1 | 0 | 1 | 1 | 2 |
| | $m_{uSer}$ | 1 | 1 | 0 | 1 | 1 | 1 |
| | $m_{uURI}$ | 1 | 0.5 | 1 | 0 | 1 | 2 |
| Interoperability | $m_{Reif}$ | 1 | 0.5 | 0.5 | 0 | 0.5 | 1 |
| | $m_{iSerial}$ | 1 | 0 | 0.5 | 1 | 1 | 2 |
| | $m_{extVoc}$ | 0.61 | 0.108 | 0.415 | 0.682 | 0.134 | 2 |
| | $m_{propVoc}$ | 0.15 | 0 | 0.513 | 0.001 | 0 | 1 |
| Accessibility | $m_{Deref}$ | 1 | 0.437 | 1 | 0.414 | 1 | 2 |
| | $m_{Avai}$ | 0.9961 | 0.9998 | 1 | 0.9999 | 0.7306 | 2 |
| | $m_{SPARQL}$ | 1 | 0 | 0 | 1 | 1 | 1 |
| | $m_{Export}$ | 1 | 1 | 1 | 1 | 1 | 0 |
| | $m_{Negot}$ | 0.5 | 0 | 0 | 1 | 1 | 1 |
| | $m_{HTML\_RDF}$ | 1 | 1 | 0 | 1 | 1 | 0 |
| | $m_{Meta}$ | 1 | 0 | 1 | 0 | 0 | 1 |
| Licensing | $m_{macLicense}$ | 1 | 0 | 0 | 1 | 0 | 1 |
| Interlinking | $m_{Inst}$ | 0.592 | 0.018 | 0.443 | 0 | 0.305 | 2 |
| | $m_{URIs}$ | 0.929 | 0.954 | 0.894 | 0.957 | 0.956 | 1 |
| Unweighted Average | | 0.708 | 0.605 | 0.498 | 0.738 | 0.625 | |
| Weighted Average | | 0.718 | 0.575 | 0.516 | 0.742 | 0.646 | |

Zaveri et al.: *Quality Assessment for Linked Open Data: A Survey.* SWJ 7(1), 2016

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),

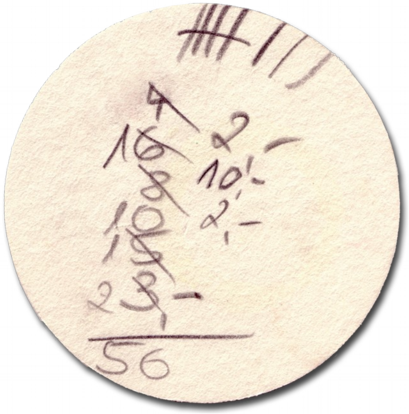
      size in April 2011: 3.6M pages, avg. 36.4 sentences each
    - Using US minimum wage: $2.25 per sentence

    → **$2.25 per statement**

    (Footnote: total cost of creating Freebase would be $6.75B)

acquisition by Google
estimated as $60-300M

# 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

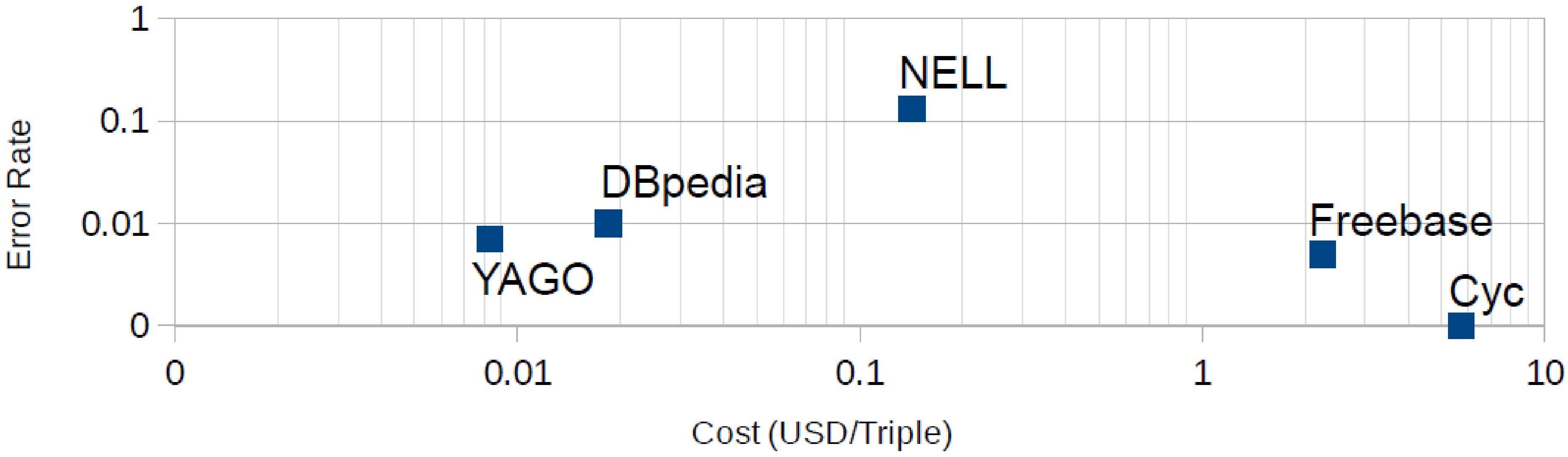
    uses WordNet: 117k synsets, we treat each synset like a Wiki page

    → **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

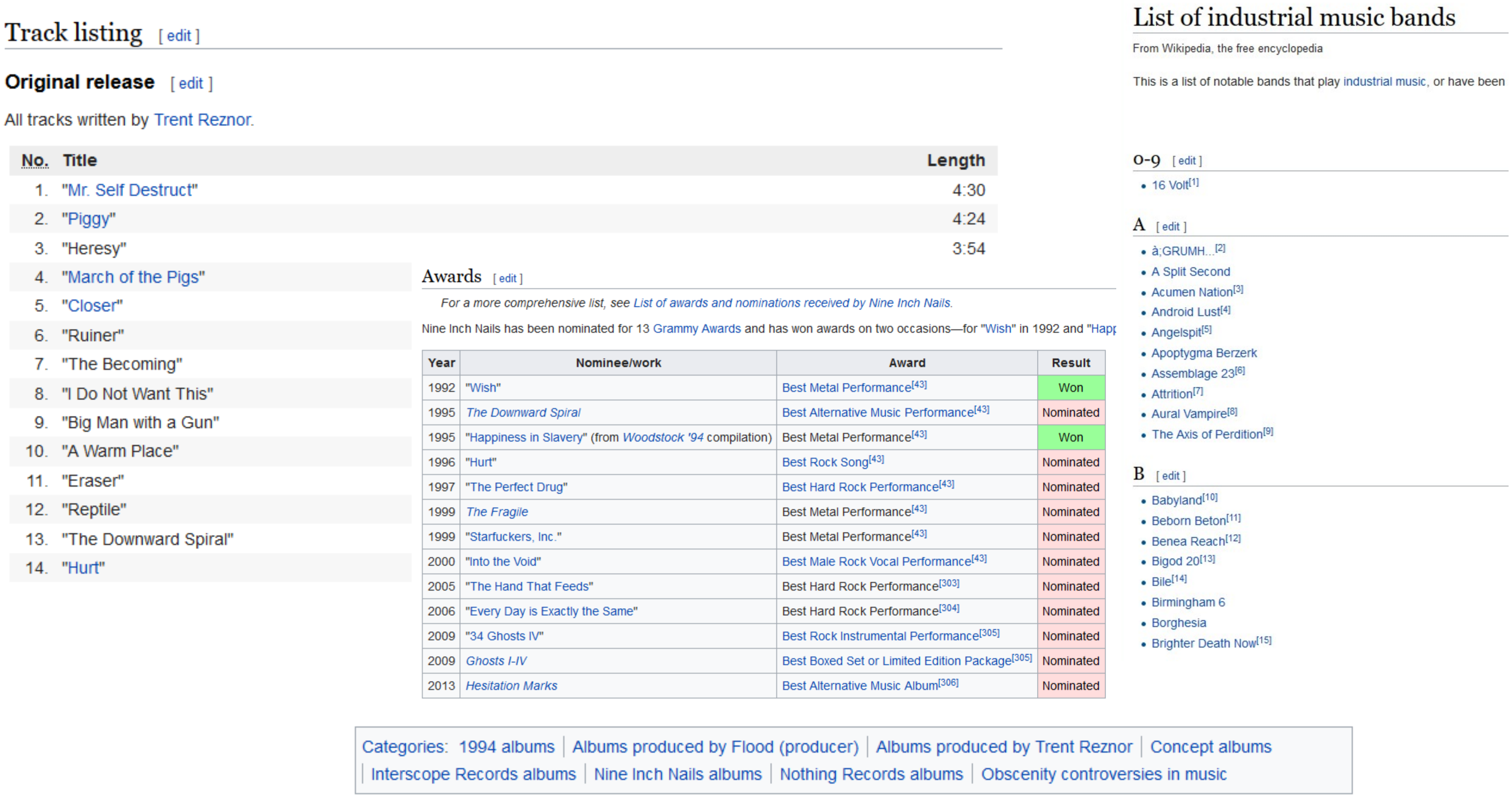


# New Kids on the Block

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 [ edit ]

### Original release [ edit ]

All tracks written by Trent Reznor.

| No. | Title | Length |
|---|---|---|
| 1. | "Mr. Self Destruct" | 4:30 |
| 2. | "Piggy" | 4:24 |
| 3. | "Heresy" | 3:54 |
| 4. | "March of the Pigs" | |
| 5. | "Closer" | |
| 6. | "Ruiner" | |
| 7. | "The Becoming" | |
| 8. | "I Do Not Want This" | |
| 9. | "Big Man with a Gun" | |
| 10. | "A Warm Place" | |
| 11. | "Eraser" | |
| 12. | "Reptile" | |
| 13. | "The Downward Spiral" | |
| 14. | "Hurt" | |

## Awards [ edit ]

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

Nine Inch Nails has been nominated for 13 Grammy Awards and has won awards on two occasions—for "Wish" in 1992 and "Happ

| Year | Nominee/work | Award | Result |
|---|---|---|---|
| 1992 | "Wish" | Best Metal Performance[43] | Won |
| 1995 | *The Downward Spiral* | Best Alternative Music Performance[43] | Nominated |
| 1995 | "Happiness in Slavery" (from *Woodstock '94* compilation) | Best Metal Performance[43] | Won |
| 1996 | "Hurt" | Best Rock Song[43] | Nominated |
| 1997 | "The Perfect Drug" | Best Hard Rock Performance[43] | Nominated |
| 1999 | *The Fragile* | Best Metal Performance[43] | Nominated |
| 1999 | "Starfuckers, Inc." | Best Metal Performance[43] | Nominated |
| 2000 | "Into the Void" | Best Male Rock Vocal Performance[43] | Nominated |
| 2005 | "The Hand That Feeds" | Best Hard Rock Performance[303] | Nominated |
| 2006 | "Every Day is Exactly the Same" | Best Hard Rock Performance[304] | Nominated |
| 2009 | "34 Ghosts IV" | Best Rock Instrumental Performance[305] | Nominated |
| 2009 | *Ghosts I-IV* | Best Boxed Set or Limited Edition Package[305] | Nominated |
| 2013 | *Hesitation Marks* | Best Alternative Music Album[306] | Nominated |

## List of industrial music bands

From Wikipedia, the free encyclopedia

This is a list of notable bands that play industrial music, or have been

### 0-9 [ edit ]

- 16 Volt[1]

### A [ edit ]

- à;GRUMH...[2]
- A Split Second
- Acumen Nation[3]
- Android Lust[4]
- Angelspit[5]
- Apoptygma Berzerk
- Assemblage 23[6]
- Attrition[7]
- Aural Vampire[8]
- The Axis of Perdition[9]

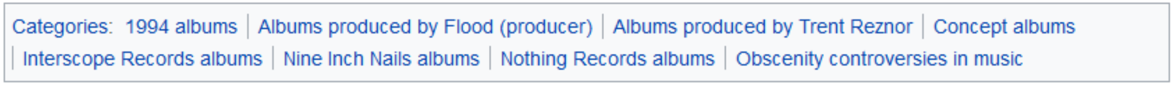
### B [ edit ]

- Babyland[10]
- Beborn Beton[11]
- Benea Reach[12]
- Bigod 20[13]
- Bile[14]
- Birmingham 6
- Borghesia
- Brighter Death Now[15]

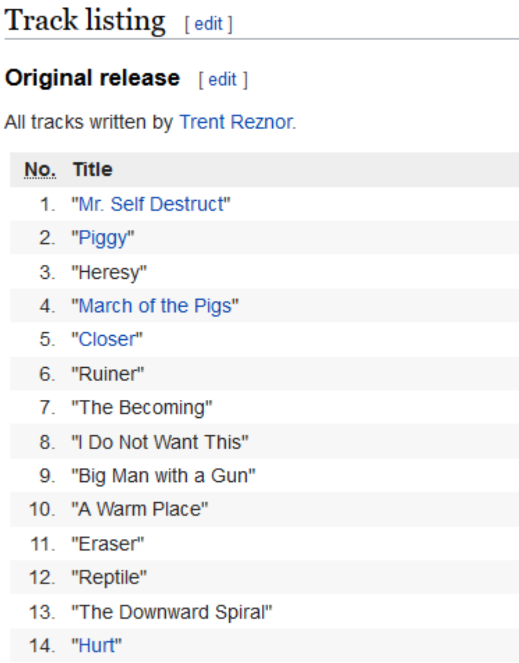
Categories: 1994 albums | Albums produced by Flood (producer) | Albums produced by Trent Reznor | Concept albums | Interscope Records albums | Nine Inch Nails albums | Nothing Records albums | Obscenity controversies in music

# 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



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



## Track listing [ edit ]

### Original release [ edit ]

All tracks written by Trent Reznor.

| No. | Title |
|---|---|
| 1. | "Mr. Self Destruct" |
| 2. | "Piggy" |
| 3. | "Heresy" |
| 4. | "March of the Pigs" |
| 5. | "Closer" |
| 6. | "Ruiner" |
| 7. | "The Becoming" |
| 8. | "I Do Not Want This" |
| 9. | "Big Man with a Gun" |
| 10. | "A Warm Place" |
| 11. | "Eraser" |
| 12. | "Reptile" |
| 13. | "The Downward Spiral" |
| 14. | "Hurt" |

# CaLiGraph Idea

- Surface patterns and semantic patterns also exist outside of Wikipedia

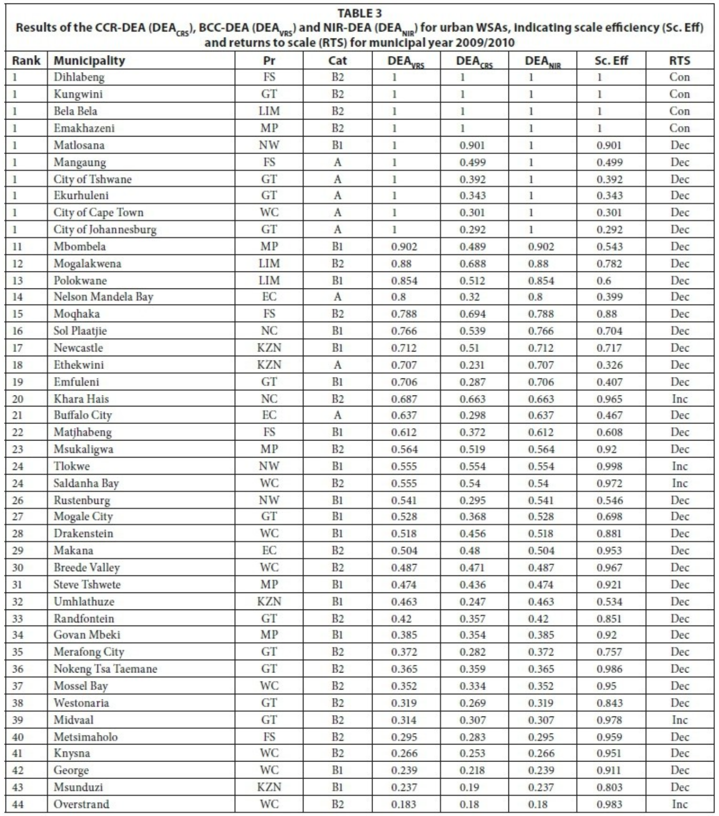
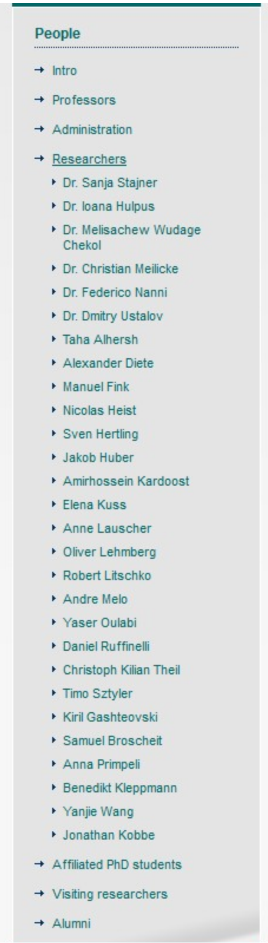


TABLE 3
Results of the CCR-DEA (DEA$_{CRS}$), BCC-DEA (DEA$_{VRS}$) and NIR-DEA (DEA$_{NIR}$) for urban WSAs, indicating scale efficiency (Sc. Eff) and returns to scale (RTS) for municipal year 2009/2010

| Rank | Municipality | Pr | Cat | DEA$_{VRS}$ | DEA$_{CRS}$ | DEA$_{NIR}$ | Sc. Eff | RTS |
|---|---|---|---|---|---|---|---|---|
| 1 | Dihlabeng | FS | B2 | 1 | 1 | 1 | 1 | Con |
| 1 | Kungwini | GT | B2 | 1 | 1 | 1 | 1 | Con |
| 1 | Bela Bela | LIM | B2 | 1 | 1 | 1 | 1 | Con |
| 1 | Emakhazeni | MP | B2 | 1 | 1 | 1 | 1 | Con |
| 1 | Matlosana | NW | B1 | 1 | 0.901 | 1 | 0.901 | Dec |
| 1 | Mangaung | FS | A | 1 | 0.499 | 1 | 0.499 | Dec |
| 1 | City of Tshwane | GT | A | 1 | 0.392 | 1 | 0.392 | Dec |
| 1 | Ekurhuleni | GT | A | 1 | 0.343 | 1 | 0.343 | Dec |
| 1 | City of Cape Town | WC | A | 1 | 0.301 | 1 | 0.301 | Dec |
| 1 | City of Johannesburg | GT | A | 1 | 0.292 | 1 | 0.292 | Dec |
| 11 | Mbombela | MP | B1 | 0.902 | 0.489 | 0.902 | 0.543 | Dec |
| 12 | Mogalakwena | LIM | B2 | 0.88 | 0.688 | 0.88 | 0.782 | Dec |
| 13 | Polokwane | LIM | B1 | 0.854 | 0.512 | 0.854 | 0.6 | Dec |
| 14 | Nelson Mandela Bay | EC | A | 0.8 | 0.32 | 0.8 | 0.399 | Dec |
| 15 | Moqhaka | FS | B2 | 0.788 | 0.694 | 0.788 | 0.88 | Dec |
| 16 | Sol Plaatjie | NC | B1 | 0.766 | 0.539 | 0.766 | 0.704 | Dec |
| 17 | Newcastle | KZN | B1 | 0.712 | 0.51 | 0.712 | 0.717 | Dec |
| 18 | Ethekwini | KZN | A | 0.707 | 0.231 | 0.707 | 0.326 | Dec |
| 19 | Emfuleni | GT | B1 | 0.706 | 0.287 | 0.706 | 0.407 | Dec |
| 20 | Khara Hais | NC | B2 | 0.687 | 0.663 | 0.663 | 0.965 | Inc |
| 21 | Buffalo City | EC | A | 0.637 | 0.298 | 0.637 | 0.467 | Dec |
| 22 | Matjhabeng | FS | B1 | 0.612 | 0.372 | 0.612 | 0.608 | Dec |
| 23 | Msukaligwa | MP | B2 | 0.564 | 0.519 | 0.564 | 0.92 | Dec |
| 24 | Tlokwe | NW | B1 | 0.555 | 0.554 | 0.554 | 0.998 | Inc |
| 24 | Saldanha Bay | WC | B2 | 0.555 | 0.54 | 0.54 | 0.972 | Inc |
| 26 | Rustenburg | NW | B1 | 0.541 | 0.295 | 0.541 | 0.546 | Dec |
| 27 | Mogale City | GT | B1 | 0.528 | 0.368 | 0.528 | 0.698 | Dec |
| 28 | Drakenstein | WC | B1 | 0.518 | 0.456 | 0.518 | 0.881 | Dec |
| 29 | Makana | EC | B2 | 0.504 | 0.48 | 0.504 | 0.953 | Dec |
| 30 | Breede Valley | WC | B2 | 0.487 | 0.471 | 0.487 | 0.967 | Dec |
| 31 | Steve Tshwete | MP | B1 | 0.474 | 0.436 | 0.474 | 0.921 | Dec |
| 32 | Umhlathuze | KZN | B1 | 0.463 | 0.247 | 0.463 | 0.534 | Dec |
| 33 | Randfontein | GT | B2 | 0.42 | 0.357 | 0.42 | 0.851 | Dec |
| 34 | Govan Mbeki | MP | B1 | 0.385 | 0.354 | 0.385 | 0.92 | Dec |
| 35 | Merafong City | GT | B2 | 0.372 | 0.282 | 0.372 | 0.757 | Dec |
| 36 | Nokeng Tsa Taemane | GT | B2 | 0.365 | 0.359 | 0.365 | 0.986 | Dec |
| 37 | Mossel Bay | WC | B2 | 0.352 | 0.334 | 0.352 | 0.95 | Dec |
| 38 | Westonaria | GT | B2 | 0.319 | 0.269 | 0.319 | 0.843 | Dec |
| 39 | Midvaal | GT | B2 | 0.314 | 0.307 | 0.307 | 0.978 | Inc |
| 40 | Metsimaholo | FS | B2 | 0.295 | 0.283 | 0.295 | 0.959 | Dec |
| 41 | Knysna | WC | B2 | 0.266 | 0.253 | 0.266 | 0.951 | Dec |
| 42 | George | WC | B1 | 0.239 | 0.218 | 0.239 | 0.911 | Dec |
| 43 | Msunduzi | KZN | B1 | 0.237 | 0.19 | 0.237 | 0.803 | Dec |
| 44 | Overstrand | WC | B2 | 0.183 | 0.18 | 0.18 | 0.983 | Inc |



People

→ Intro

→ Professors

→ Administration

→ Researchers

- Dr. Sanja Stajner
- Dr. Ioana Hulpus
- Dr. Melisachew Wudage Chekol
- Dr. Christian Meilicke
- Dr. Federico Nanni
- Dr. Dmitry Ustalov
- Taha Alhersh
- Alexander Diete
- Manuel Fink
- Nicolas Heist
- Sven Hertling
- Jakob Huber
- Amirhossein Kardoost
- Elena Kuss
- Anne Lauscher
- Oliver Lehmberg
- Robert Litschko
- Andre Melo
- Yaser Oulabi
- Daniel Ruffinelli
- Christoph Kilian Theil
- Timo Sztyler
- Kiril Gashteovski
- Samuel Broscheit
- Anna Primpeli
- Benedikt Kleppmann
- Yanjie Wang
- Jonathan Kobbe

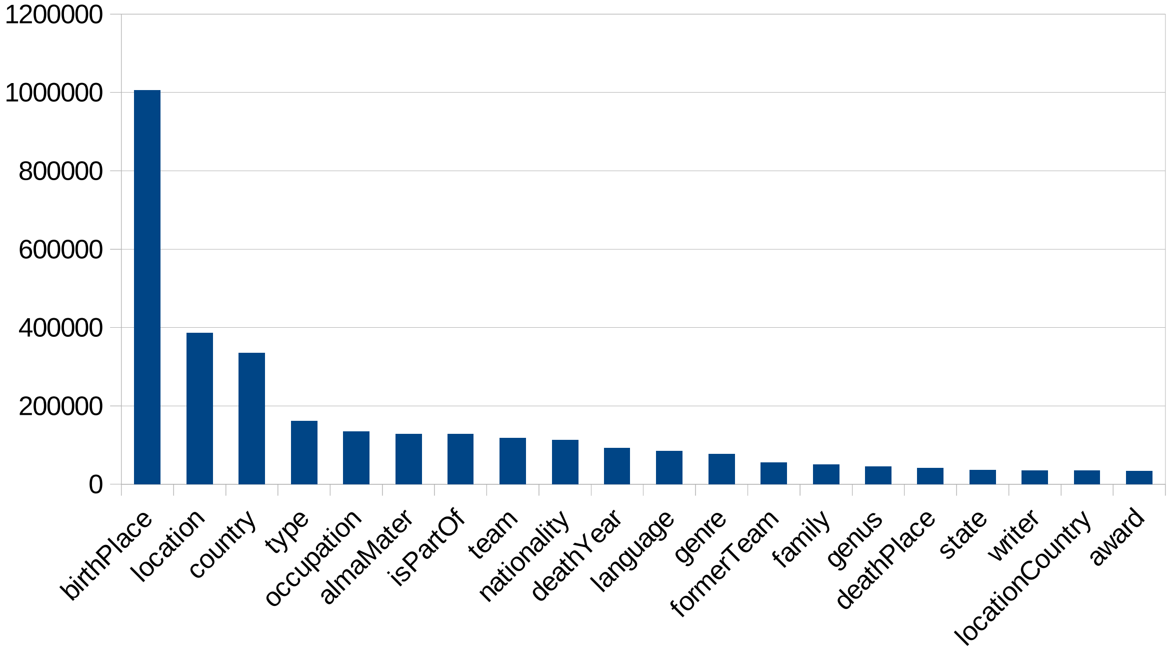
→ Affiliated PhD students

→ Visiting researchers

→ Alumni

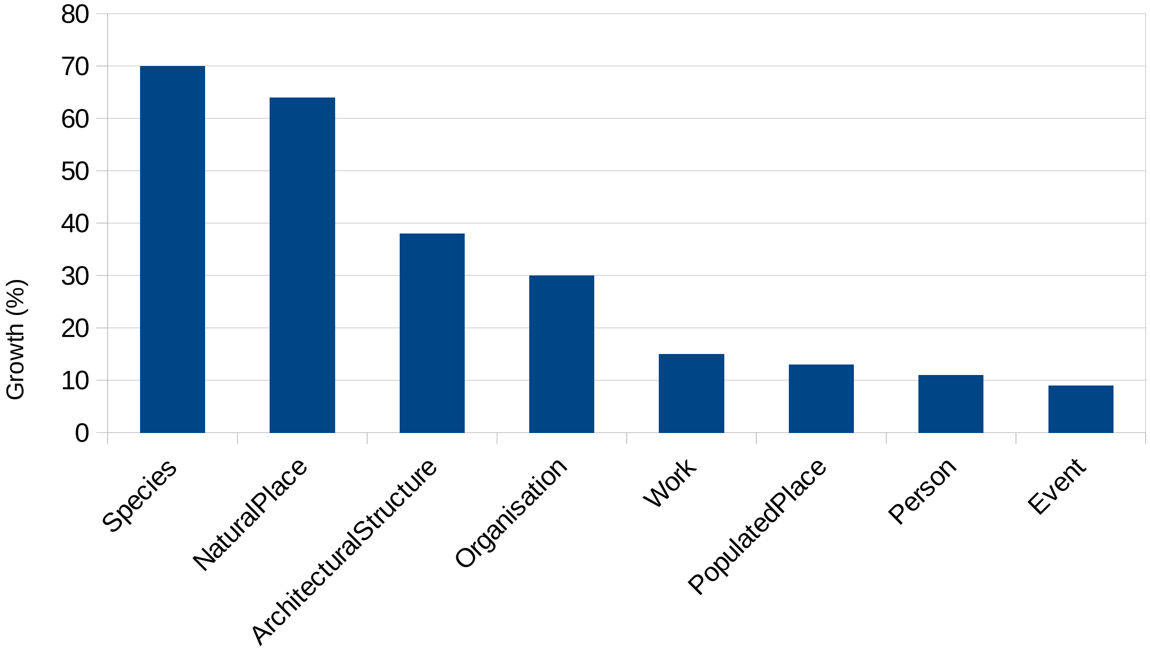
# CaLiGraph – Current State

- Significant coverage enhancements of DBpedia Properties



# CaLiGraph – Current State

- Significant instance set enhancements by list extraction

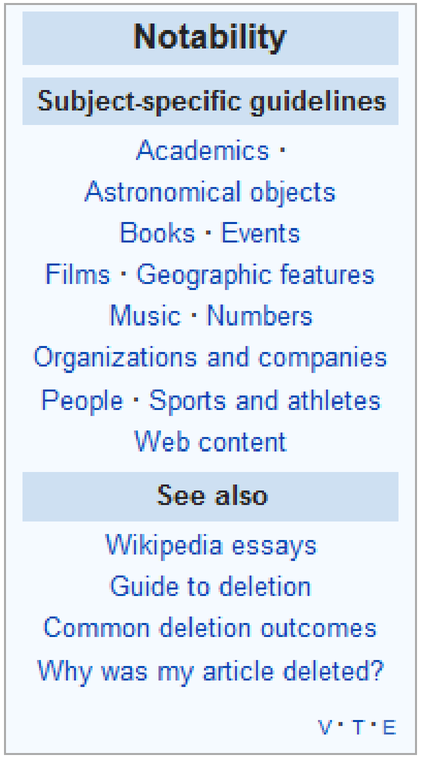


# From DBpedia to DBkWik

- 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



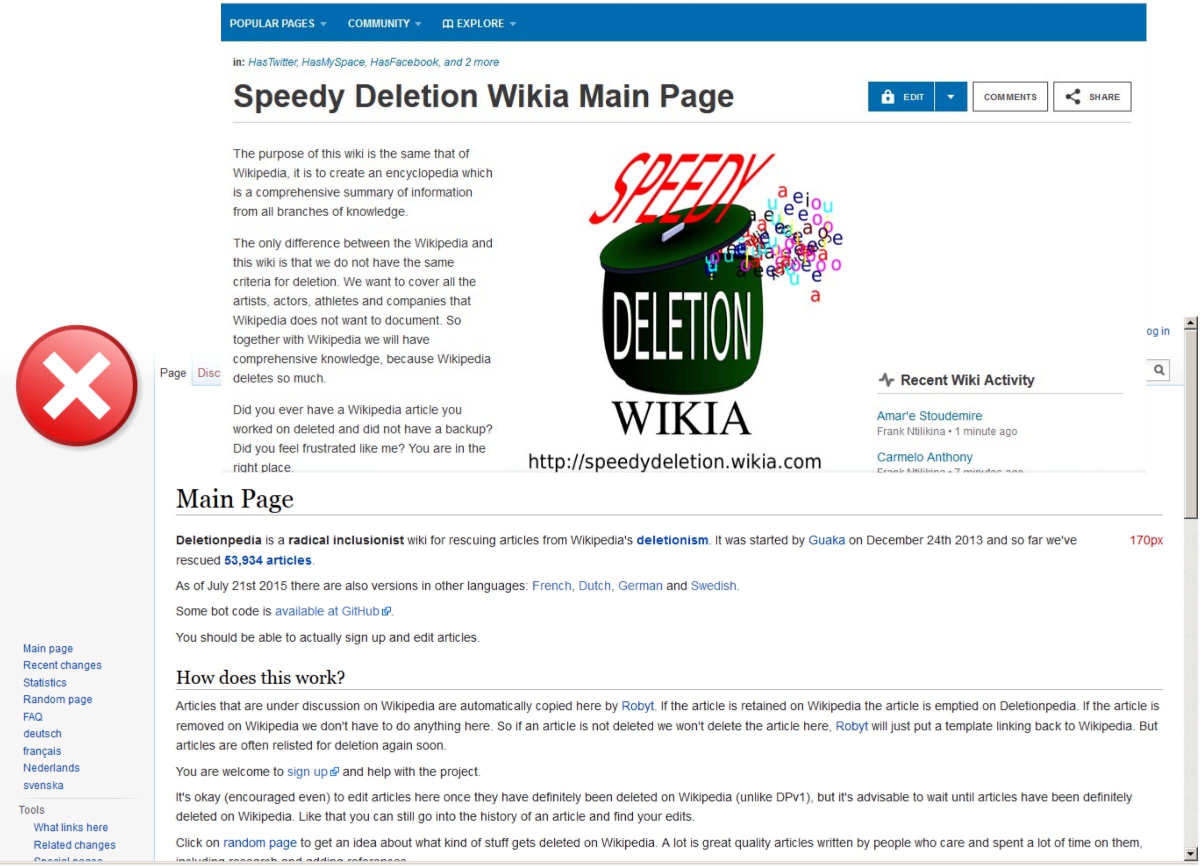
**Notability**

**Subject-specific guidelines**

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

**See also**

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



in: HasTwitter, HasMySpace, HasFacebook, and 2 more

## Speedy Deletion Wikia Main Page

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

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

Did you ever have a Wikipedia article you worked on deleted and did not have a backup? Did you feel frustrated like me? You are in the right place.

http://speedydeletion.wikia.com

**Recent Wiki Activity**

Amar'e Stoudemire
Frank Ntilikina • 1 minute ago

Carmelo Anthony
Frank Ntilikina • 7 minutes ago

Main page
Recent changes
Statistics
Random page
FAQ
deutsch
français
Nederlands
svenska

Tools
What links here
Related changes
Special pages

## Main Page

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

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

Some bot code is available at GitHub.

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

## How does this work?

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

You are welcome to sign up and help with the project.

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

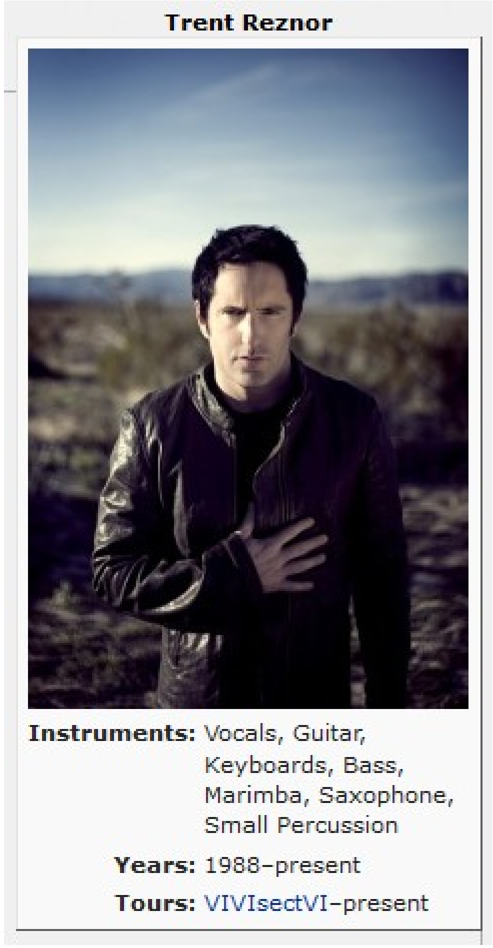
Click on random page to get an idea about what kind of stuff gets deleted on Wikipedia. A lot is great quality articles written by people who care and spent a lot of time on them, including research and adding references.

# 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

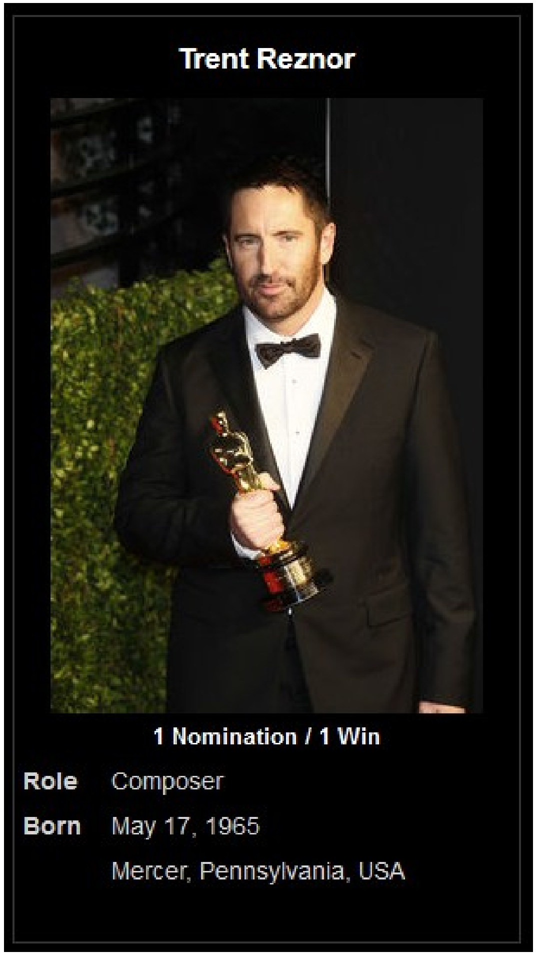


Trent Reznor

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

**Years:** 1988–present

**Tours:** VIVIsectVI–present



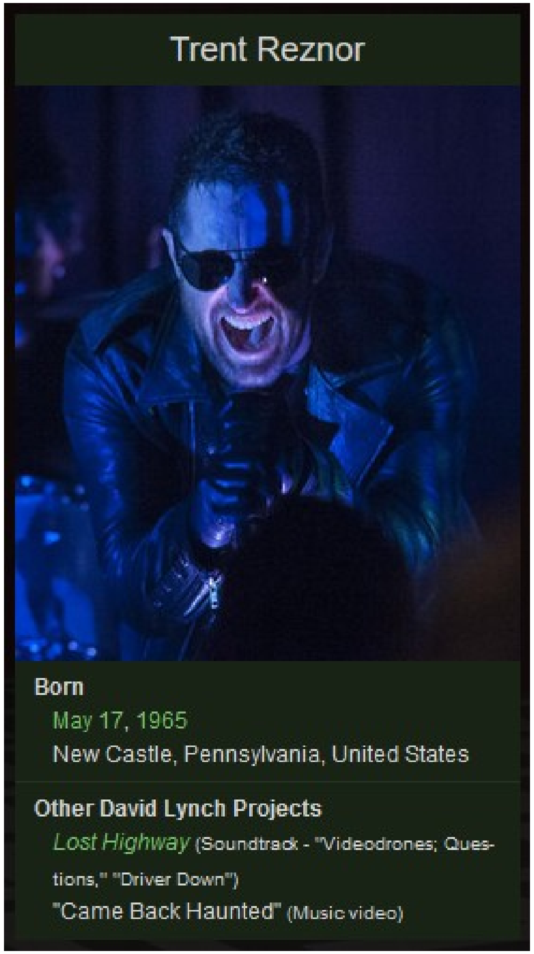
Trent Reznor

1 Nomination / 1 Win

**Role** Composer

**Born** May 17, 1965

Mercer, Pennsylvania, USA



Trent Reznor

Born

May 17, 1965

New Castle, Pennsylvania, United States

Other David Lynch Projects

*Lost Highway* (Soundtrack - "Videodrones; Questions," "Driver Down")

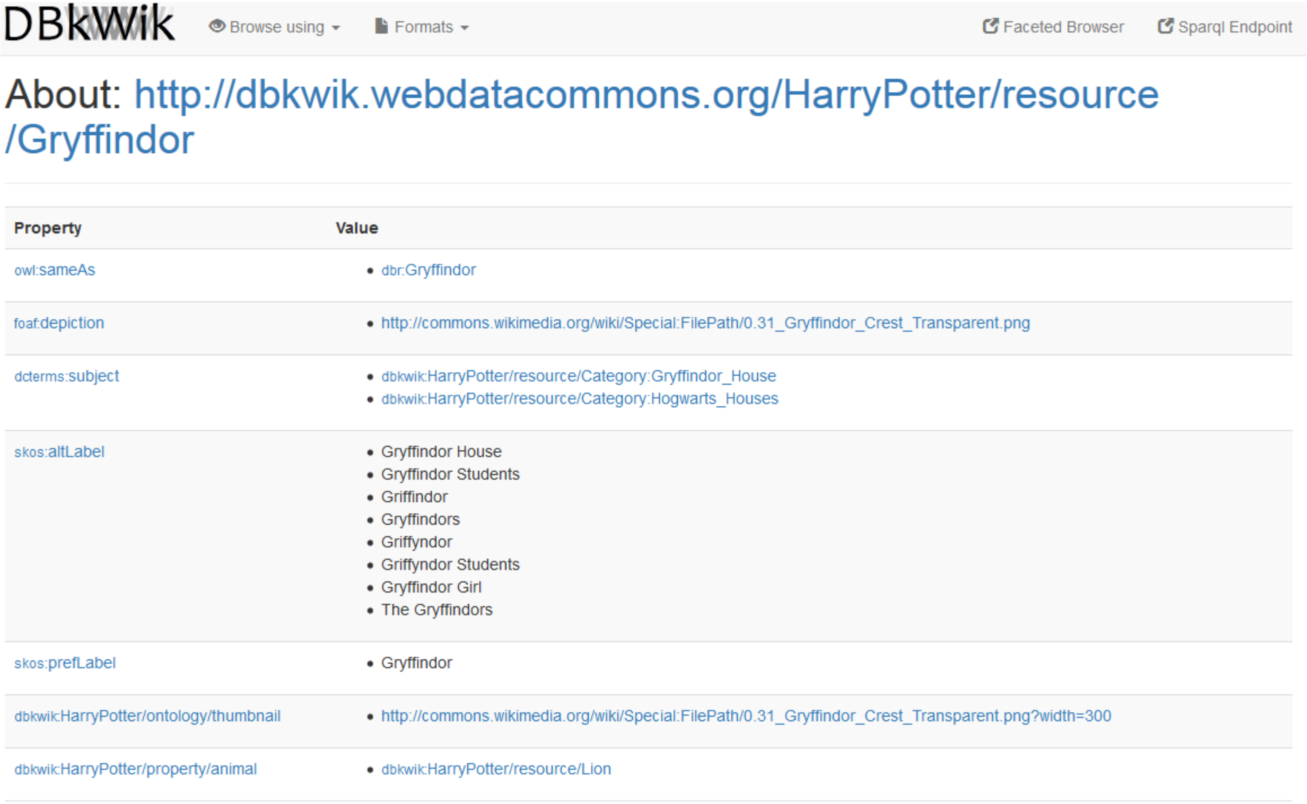
"Came Back Haunted" (Music video)

# 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



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

| Property | Value |
|---|---|
| owl:sameAs | dbr:Gryffindor |
| foaf:depiction | http://commons.wikimedia.org/wiki/Special:FilePath/0.31_Gryffindor_Crest_Transparent.png |
| dcterms:subject | dbkwik:HarryPotter/resource/Category:Gryffindor_House<br>dbkwik:HarryPotter/resource/Category:Hogwarts_Houses |
| skos:altLabel | Gryffindor House<br>Gryffindor Students<br>Griffindor<br>Gryffindors<br>Griffyndor<br>Griffyndor Students<br>Gryffindor Girl<br>The Gryffindors |
| skos:prefLabel | Gryffindor |
| dbkwik:HarryPotter/ontology/thumbnail | http://commons.wikimedia.org/wiki/Special:FilePath/0.31_Gryffindor_Crest_Transparent.png?width=300 |
| dbkwik:HarryPotter/property/animal | dbkwik:HarryPotter/resource/Lion |

http://dbkwik.org/

# From DBpedia to DBkWik

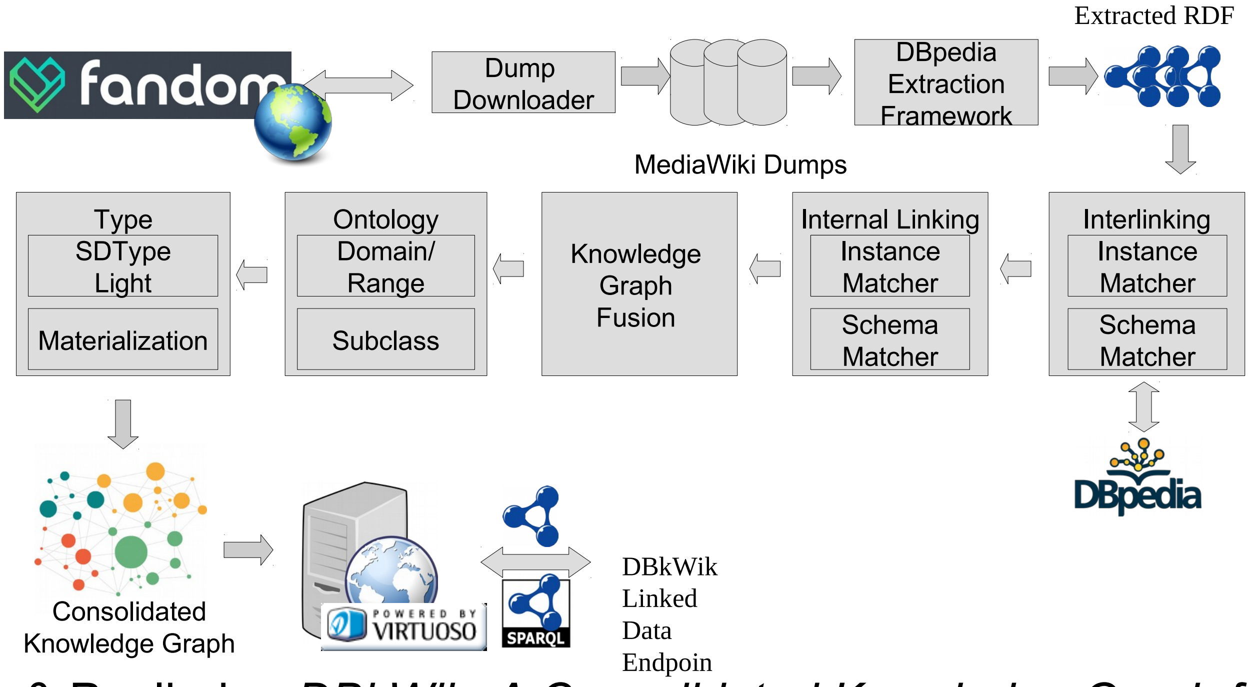
- Differences to DBpedia
  - DBpedia has manually created mappings to an ontology
  - Wikipedia has one page per subject
  - Wikipedia has global infobox conventions (more or less)

- 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

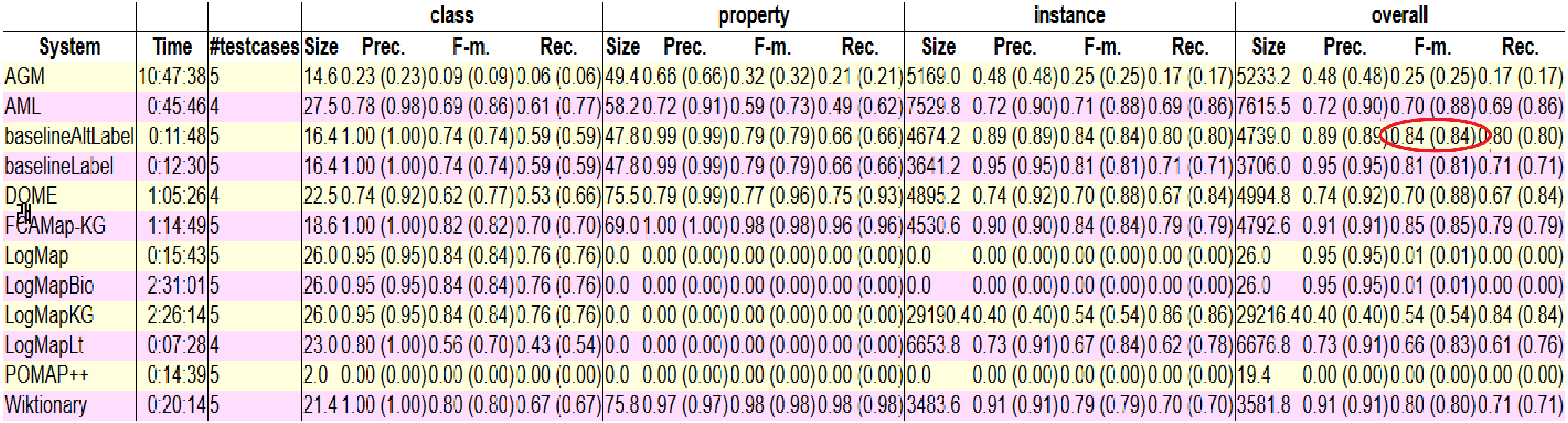
- Downloaded ~15k Wiki dumps from Fandom
  - 52.4GB of data, roughly the size of the English Wikipedia

- 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

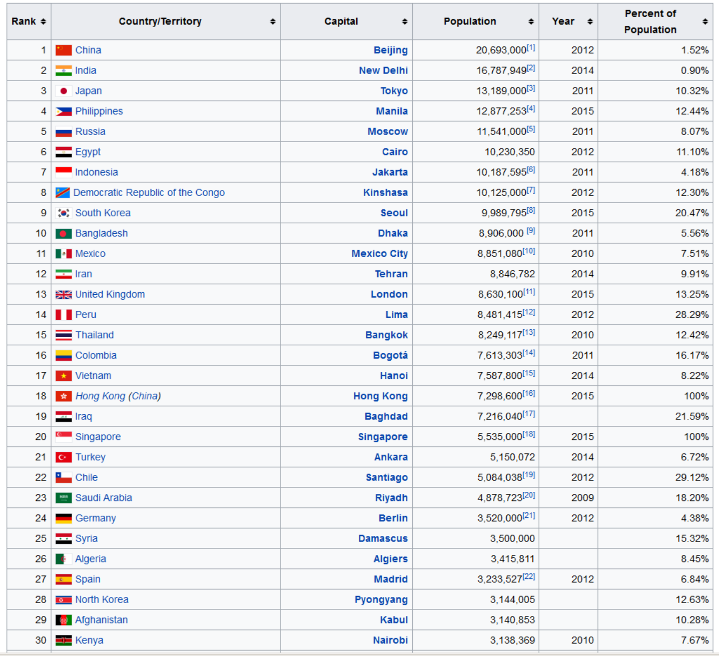
- A new task at OAEI since 2018
  - Benchmark for schema/instance matching tools
  - Turned out to be non-trivial



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

# WebIsALOD

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



| Rank | Country/Territory | Capital | Population | Year | Percent of Population |
|---|---|---|---|---|---|
| 1 | China | Beijing | 20,693,000[1] | 2012 | 1.52% |
| 2 | India | New Delhi | 16,787,949[2] | 2014 | 0.90% |
| 3 | Japan | Tokyo | 13,189,000[3] | 2011 | 10.32% |
| 4 | Philippines | Manila | 12,877,253[4] | 2015 | 12.44% |
| 5 | Russia | Moscow | 11,541,000[5] | 2011 | 8.07% |
| 6 | Egypt | Cairo | 10,230,350 | 2012 | 11.10% |
| 7 | Indonesia | Jakarta | 10,187,595[6] | 2011 | 4.18% |
| 8 | Democratic Republic of the Congo | Kinshasa | 10,125,000[7] | 2012 | 12.30% |
| 9 | South Korea | Seoul | 9,989,795[8] | 2015 | 20.47% |
| 10 | Bangladesh | Dhaka | 8,906,000 [9] | 2011 | 5.56% |
| 11 | Mexico | Mexico City | 8,851,080[10] | 2010 | 7.51% |
| 12 | Iran | Tehran | 8,846,782 | 2014 | 9.91% |
| 13 | United Kingdom | London | 8,630,100[11] | 2015 | 13.25% |
| 14 | Peru | Lima | 8,481,415[12] | 2012 | 28.29% |
| 15 | Thailand | Bangkok | 8,249,117[13] | 2010 | 12.42% |
| 16 | Colombia | Bogotá | 7,613,303[14] | 2011 | 16.17% |
| 17 | Vietnam | Hanoi | 7,587,800[15] | 2014 | 8.22% |
| 18 | *Hong Kong* (*China*) | Hong Kong | 7,298,600[16] | 2015 | 100% |
| 19 | Iraq | Baghdad | 7,216,040[17] |  | 21.59% |
| 20 | Singapore | Singapore | 5,535,000[18] | 2015 | 100% |
| 21 | Turkey | Ankara | 5,150,072 | 2014 | 6.72% |
| 22 | Chile | Santiago | 5,084,038[19] | 2012 | 29.12% |
| 23 | Saudi Arabia | Riyadh | 4,878,723[20] | 2009 | 18.20% |
| 24 | Germany | Berlin | 3,520,000[21] | 2012 | 4.38% |
| 25 | Syria | Damascus | 3,500,000 |  | 15.32% |
| 26 | Algeria | Algiers | 3,415,811 |  | 8.45% |
| 27 | Spain | Madrid | 3,233,527[22] | 2012 | 6.84% |
| 28 | North Korea | Pyongyang | 3,144,005 |  | 12.63% |
| 29 | Afghanistan | Kabul | 3,140,853 |  | 10.28% |
| 30 | Kenya | Nairobi | 3,138,369 | 2010 | 7.67% |

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

# WebIsALOD

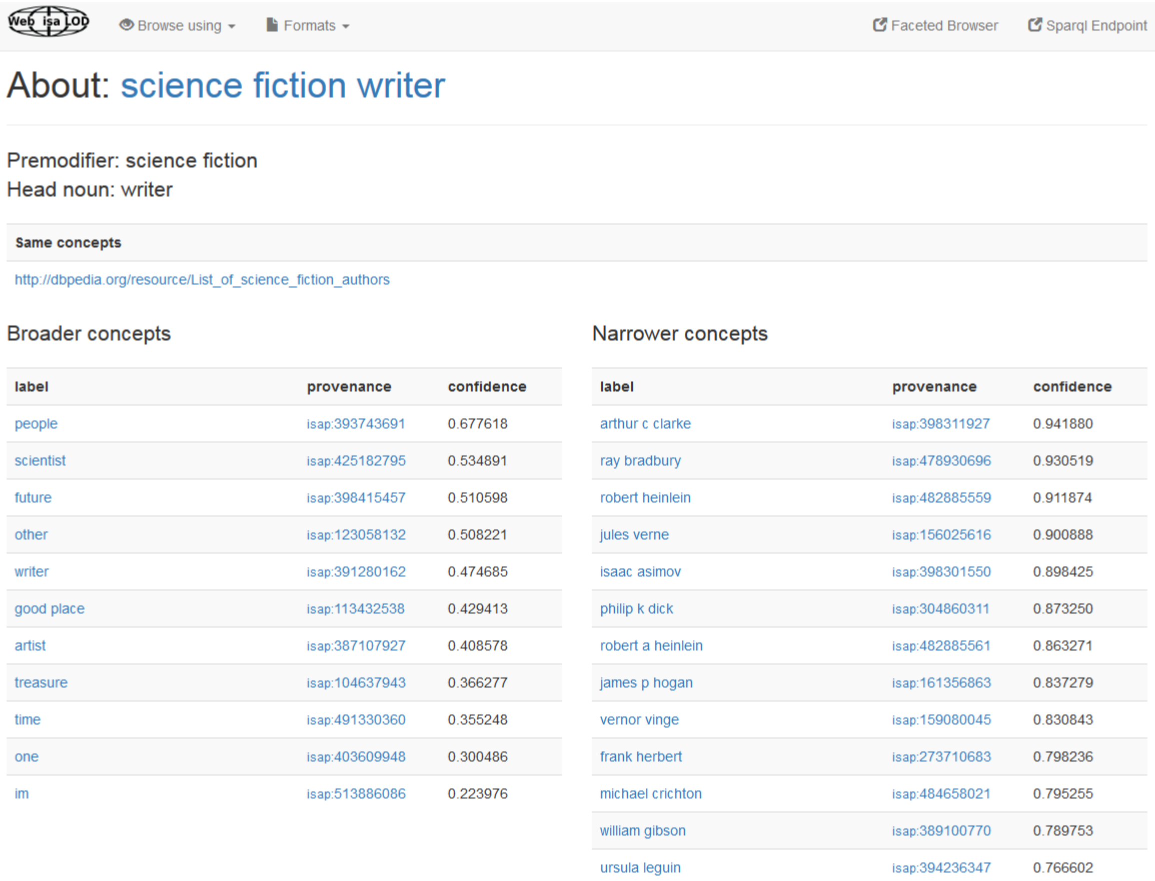
- Extraction of type information using Hearst-like patterns, e.g.,
  - T, such as X
  - X, Y, and other T
- Text corpus: common crawl
  - ~2 TB crawled web pages
  - Fast implementation: regex over text
  - “Expensive” operations only applied once regex has fired
- Resulting database
  - 400M hypernymy relations

Common Crawl

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

# WebIsALOD

- Example:



About: science fiction writer

Premodifier: science fiction
Head noun: writer

**Same concepts**

http://dbpedia.org/resource/List_of_science_fiction_authors

Broader concepts

| label | provenance | confidence |
|---|---|---|
| people | isap:393743691 | 0.677618 |
| scientist | isap:425182795 | 0.534891 |
| future | isap:398415457 | 0.510598 |
| other | isap:123058132 | 0.508221 |
| writer | isap:391280162 | 0.474685 |
| good place | isap:113432538 | 0.429413 |
| artist | isap:387107927 | 0.408578 |
| treasure | isap:104637943 | 0.366277 |
| time | isap:491330360 | 0.355248 |
| one | isap:403609948 | 0.300486 |
| im | isap:513886086 | 0.223976 |

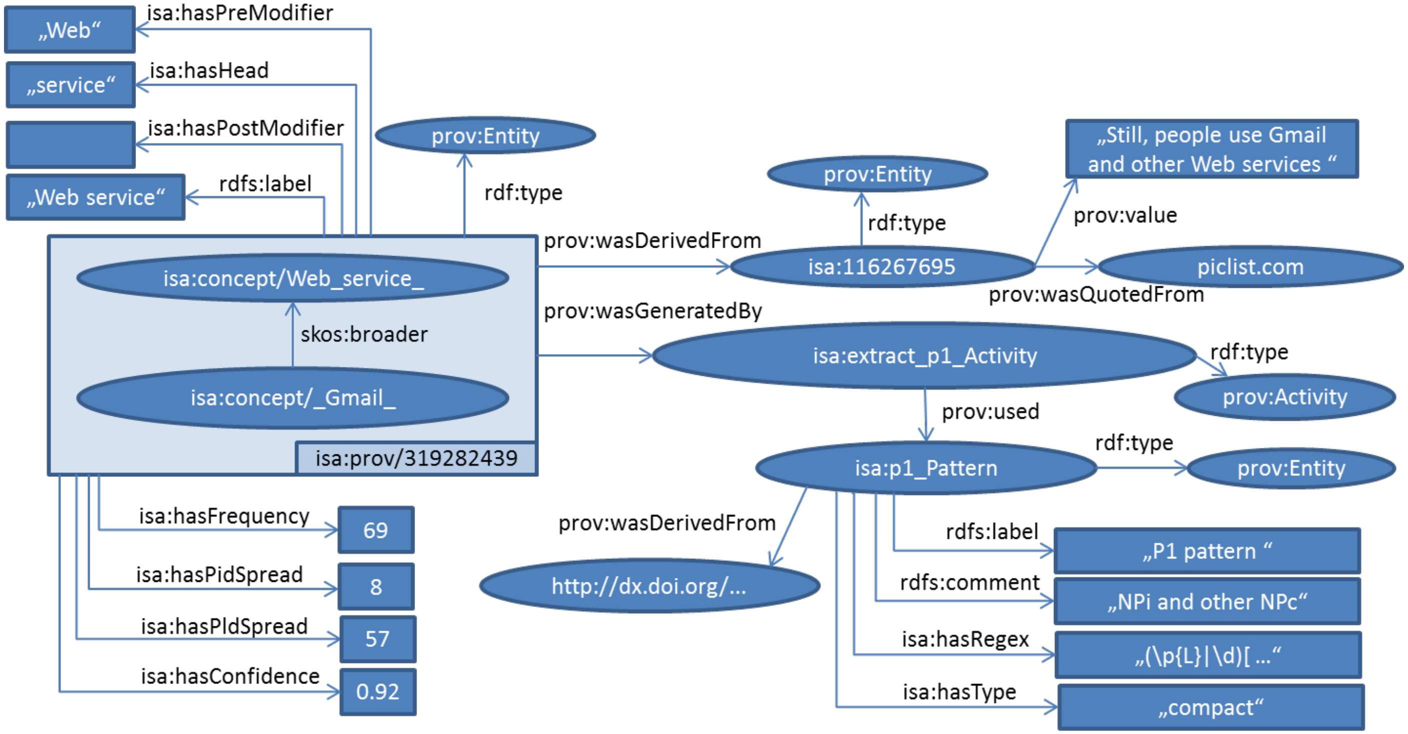
Narrower concepts

| label | provenance | confidence |
|---|---|---|
| arthur c clarke | isap:398311927 | 0.941880 |
| ray bradbury | isap:478930696 | 0.930519 |
| robert heinlein | isap:482885559 | 0.911874 |
| jules verne | isap:156025616 | 0.900888 |
| isaac asimov | isap:398301550 | 0.898425 |
| philip k dick | isap:304860311 | 0.873250 |
| robert a heinlein | isap:482885561 | 0.863271 |
| james p hogan | isap:161356863 | 0.837279 |
| vernor vinge | isap:159080045 | 0.830843 |
| frank herbert | isap:273710683 | 0.798236 |
| michael crichton | isap:484658021 | 0.795255 |
| william gibson | isap:389100770 | 0.789753 |
| ursula leguin | isap:394236347 | 0.766602 |

http://webisa.webdatacommons.org/

# WebIsALOD

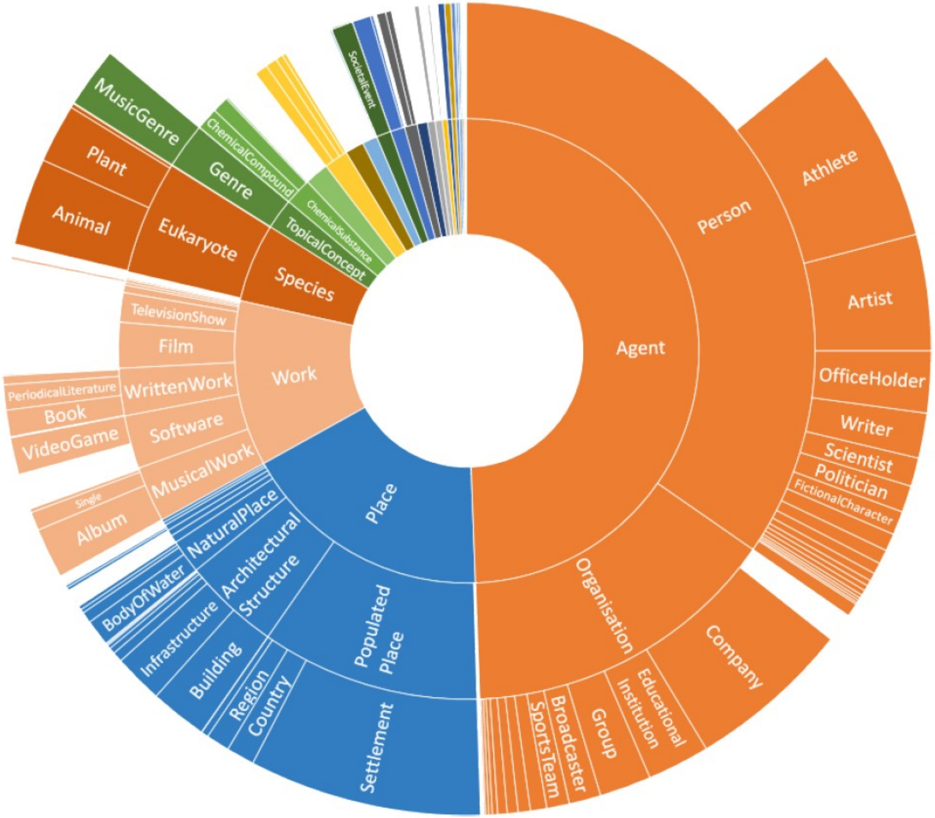
- Initial effort: transformation to a LOD dataset
  - including rich provenance information



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

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?