Public Knowledge Graphs

IE650 Knowledge Graphs

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Previously on “Knowledge Graphs”

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

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

Introduction

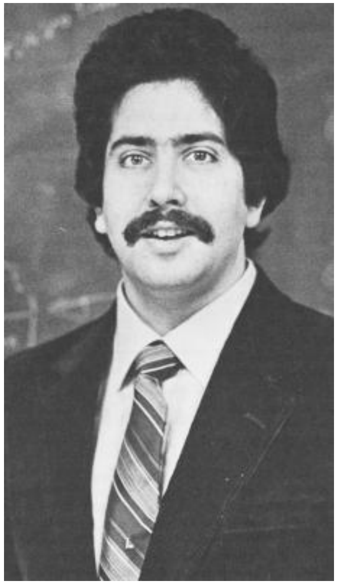
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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	public
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	private
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	

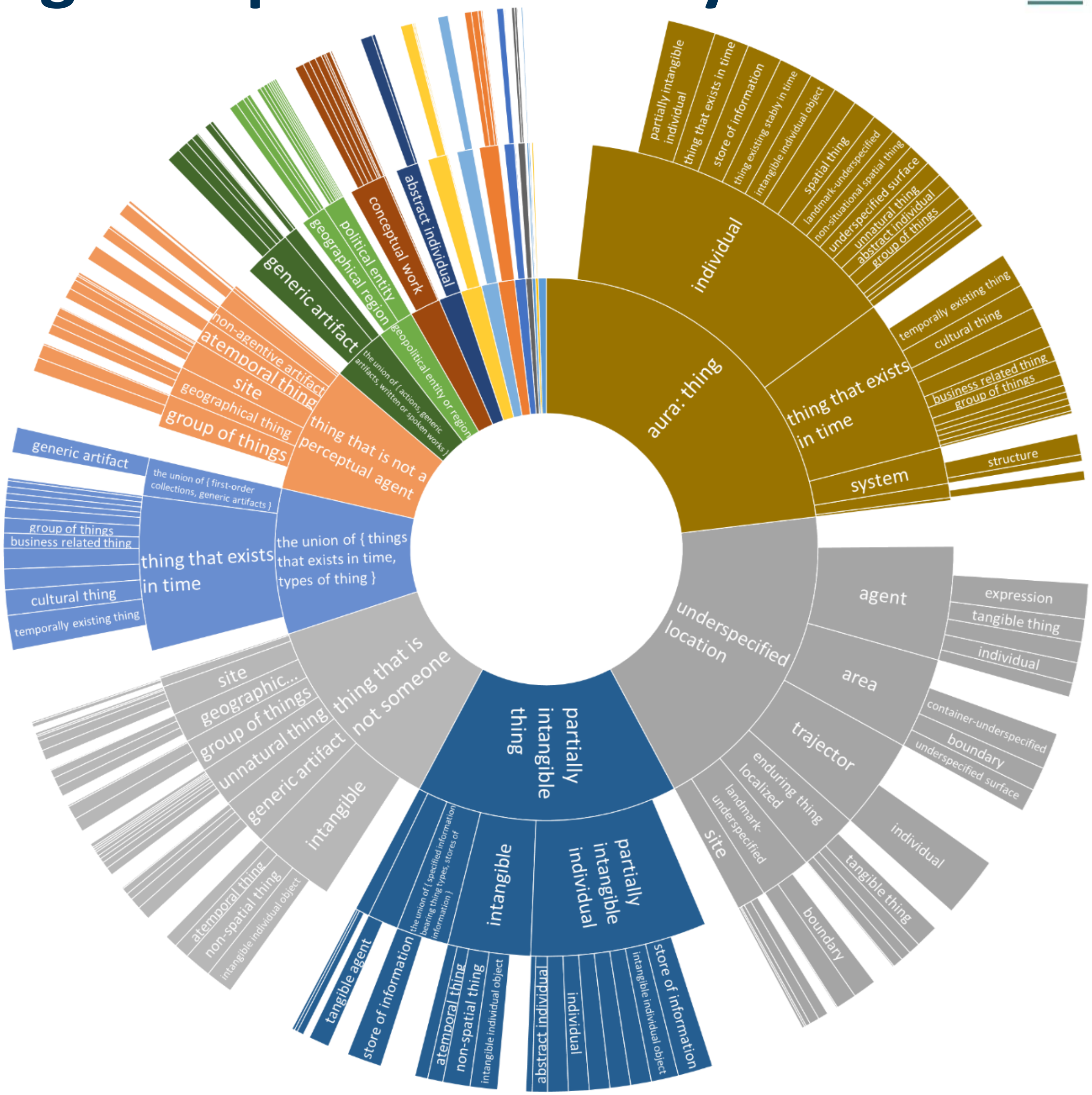
Paulheim: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web 8:3 (2017), pp. 489-508

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

Knowledge Graph Creation: Freebase

UNIVERSITY OF MANNHEIM
Data and Web Science Group

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

Freebase™

- 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

Knowledge Graph Creation

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

Max. user involvement

Max. degree of formality

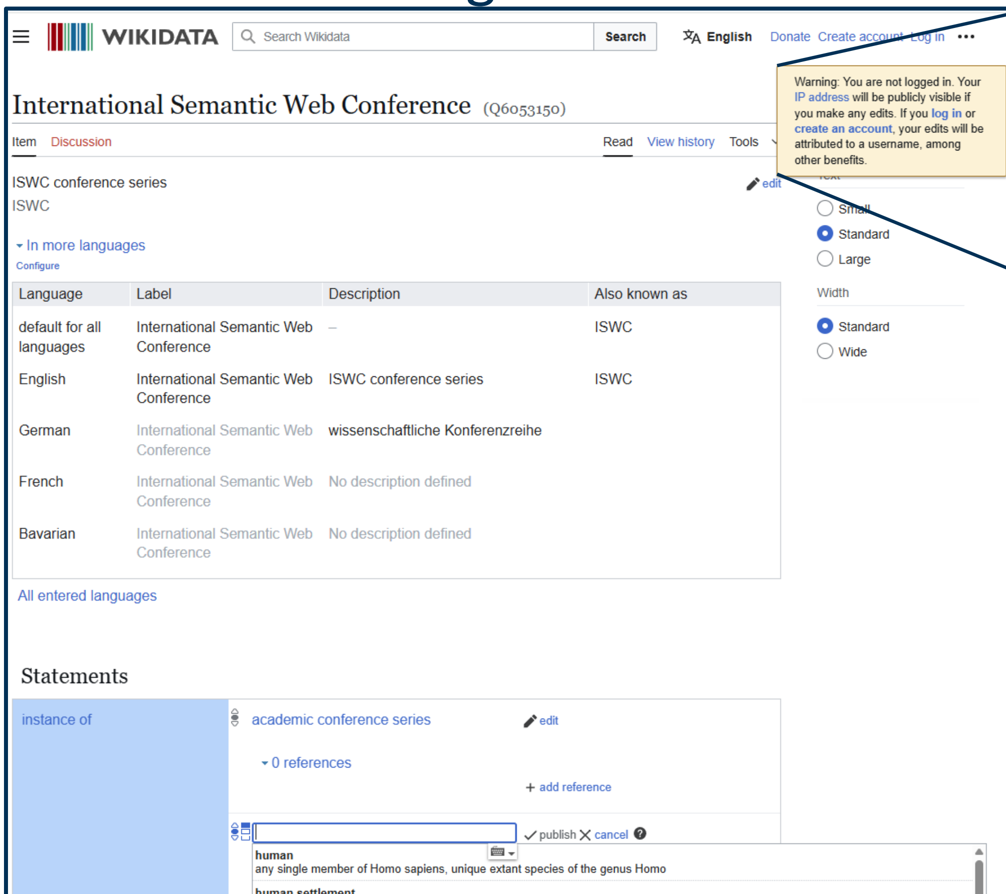
Knowledge Graph Creation: Wikidata

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

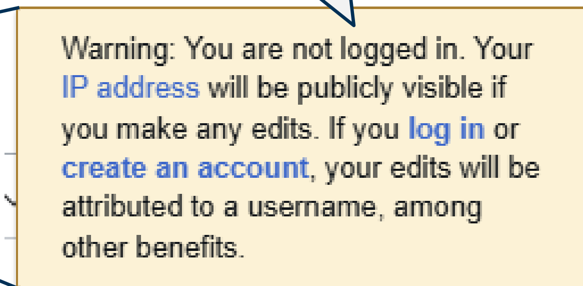
- Present
  - One of the largest public knowledge graphs (see later)
  - Includes rich provenance

Knowledge Graph Creation: Wikidata

- Collaborative editing

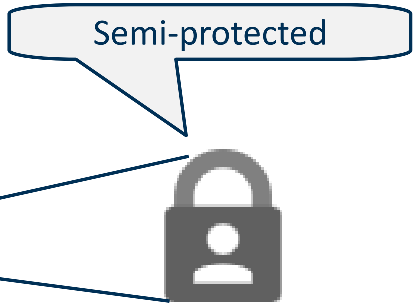


Even without login



Knowledge Graph Creation: Wikidata

- Some pages are (semi-) protected



WIKIDATA English Donate Create account Log in

University of Mannheim (Q317070)

Item Discussion Read View history Tools

university in Germany

Uni Mannheim | UMA | Hochschule Mannheim

In more languages

Configure

Language	Label	
default for all languages	Universität Mannheim	
English	University of Mannheim	
German	Universität Mannheim	
French	université de Mannheim	
Bavarian	Universität Mannheim	No description defined

All entered languages

Statements

instance of	public university
	0 references

Protection types

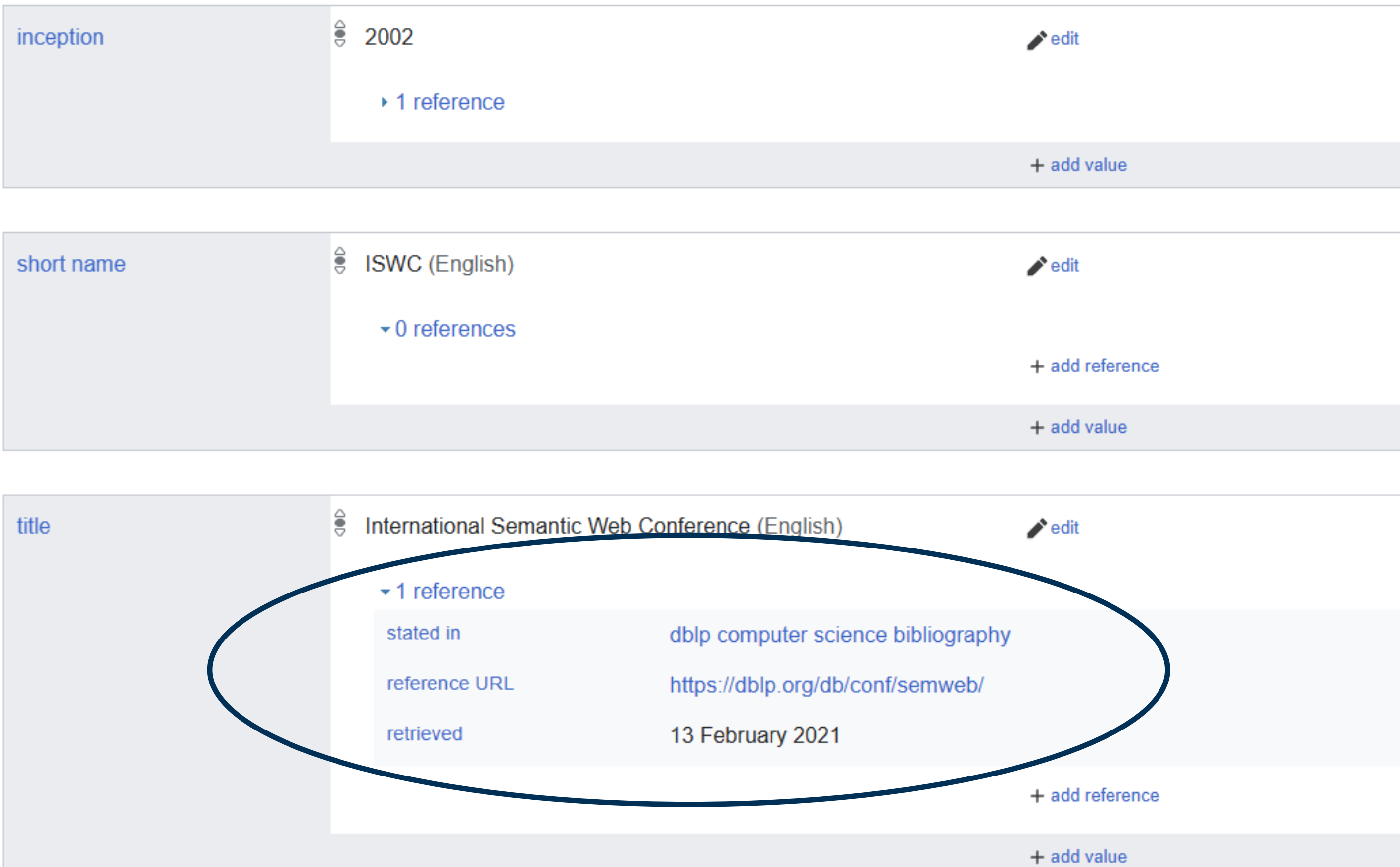
There are three kinds of protection that can be applied to existing pages.

- **Semi-protection** (symbolized by 🔒) allows only autoconfirmed users/confirmed users to edit the page.
- **Full protection** (symbolized by 🔒 or 🔒) only allows administrators to edit the page.
- **Move protection** (symbolized by 🔒) does not affect who can edit pages, but does prevent anyone except for administrators from moving affected pages to new locations. Most high traffic pages are move protected. Move protection cannot be applied to items or properties.
- **Create-protection** (symbolized by 🔒, for both full and semi-create protection) can be applied to deleted or nonexistent pages. Like move protection, however, it cannot be applied to deleted items or properties.
  - **Full create protection** prevents the page from being created except by administrators.
  - **Semi-create protection** also allows autoconfirmed and confirmed users to create.

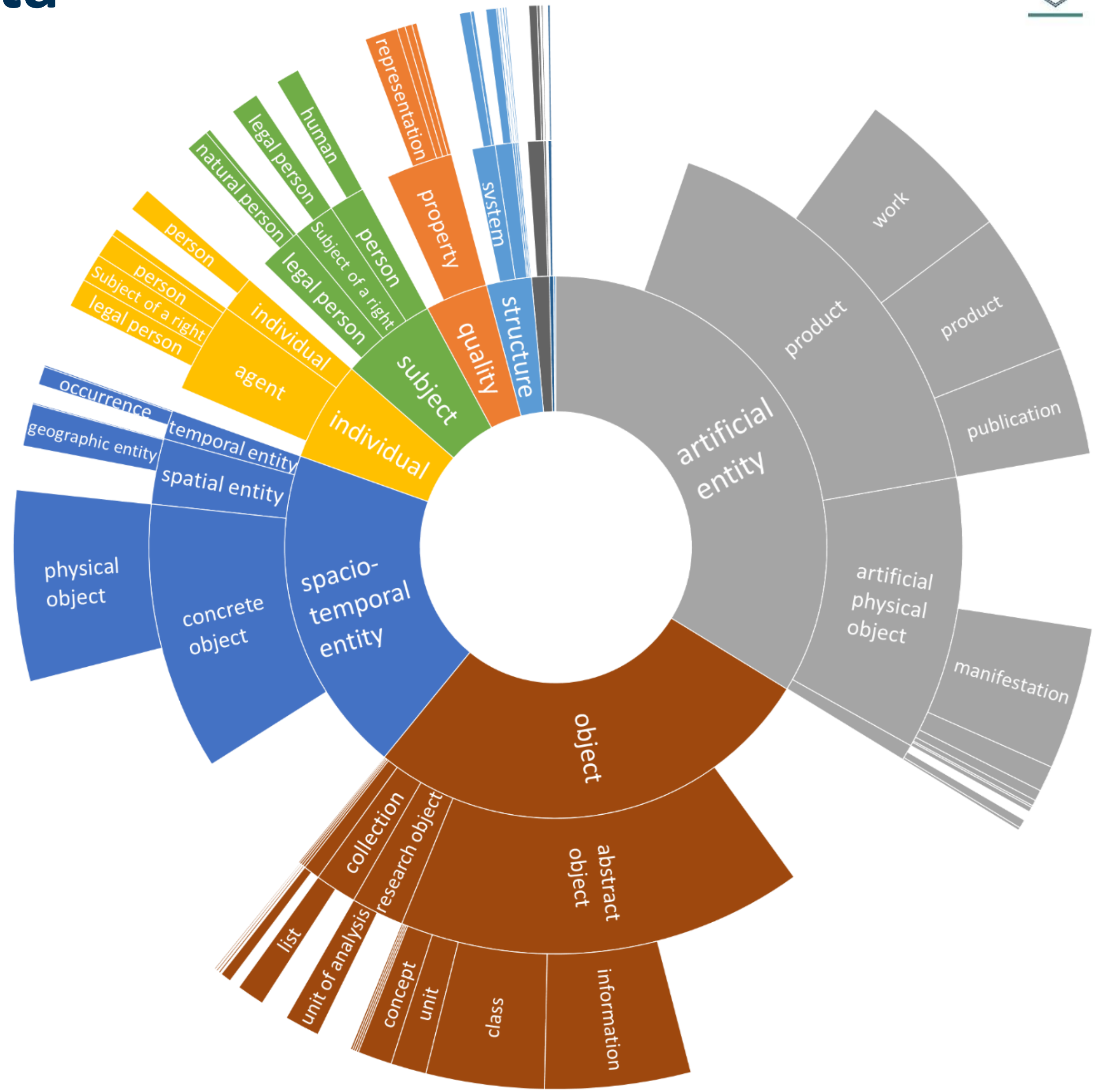
https://www.wikidata.org/wiki/Wikidata:Protection_policy

Knowledge Graph Creation: Wikidata

- Provenance



Wikidata



artificial entity
product
work
product
publication
artificial physical object
manifestation
object
abstract object
research object
collection
list
unit of analysis
concept
unit
class
information
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
natural person
Subject of a right
person
legal person
human
quality
property
representation
structure
system

Knowledge Graph Creation

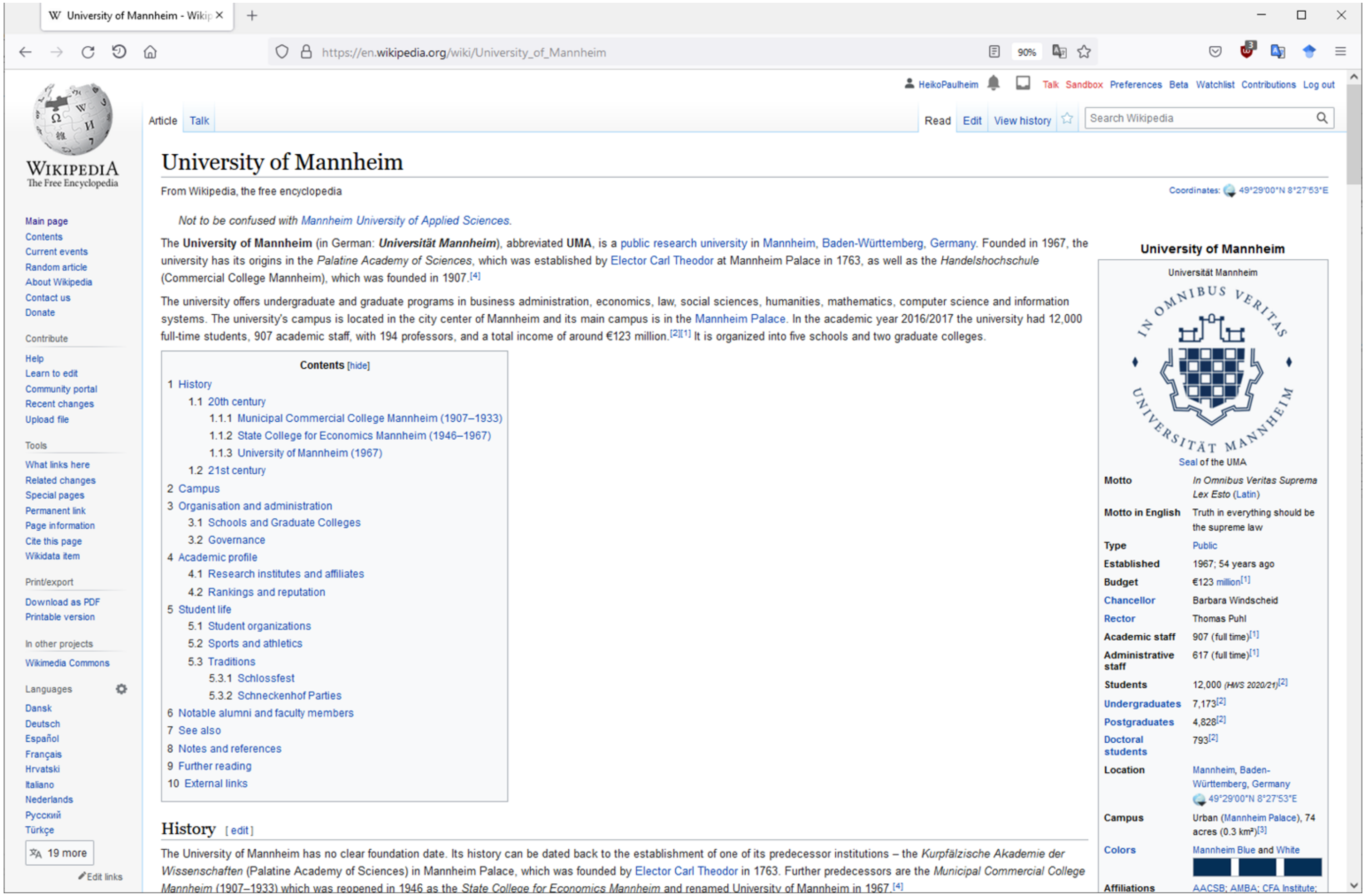
- Lesson learned no. 3:
  - There is not one truth (but allowing for plurality adds complexity)

Knowledge Graph Creation: DBpedia & YAGO

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

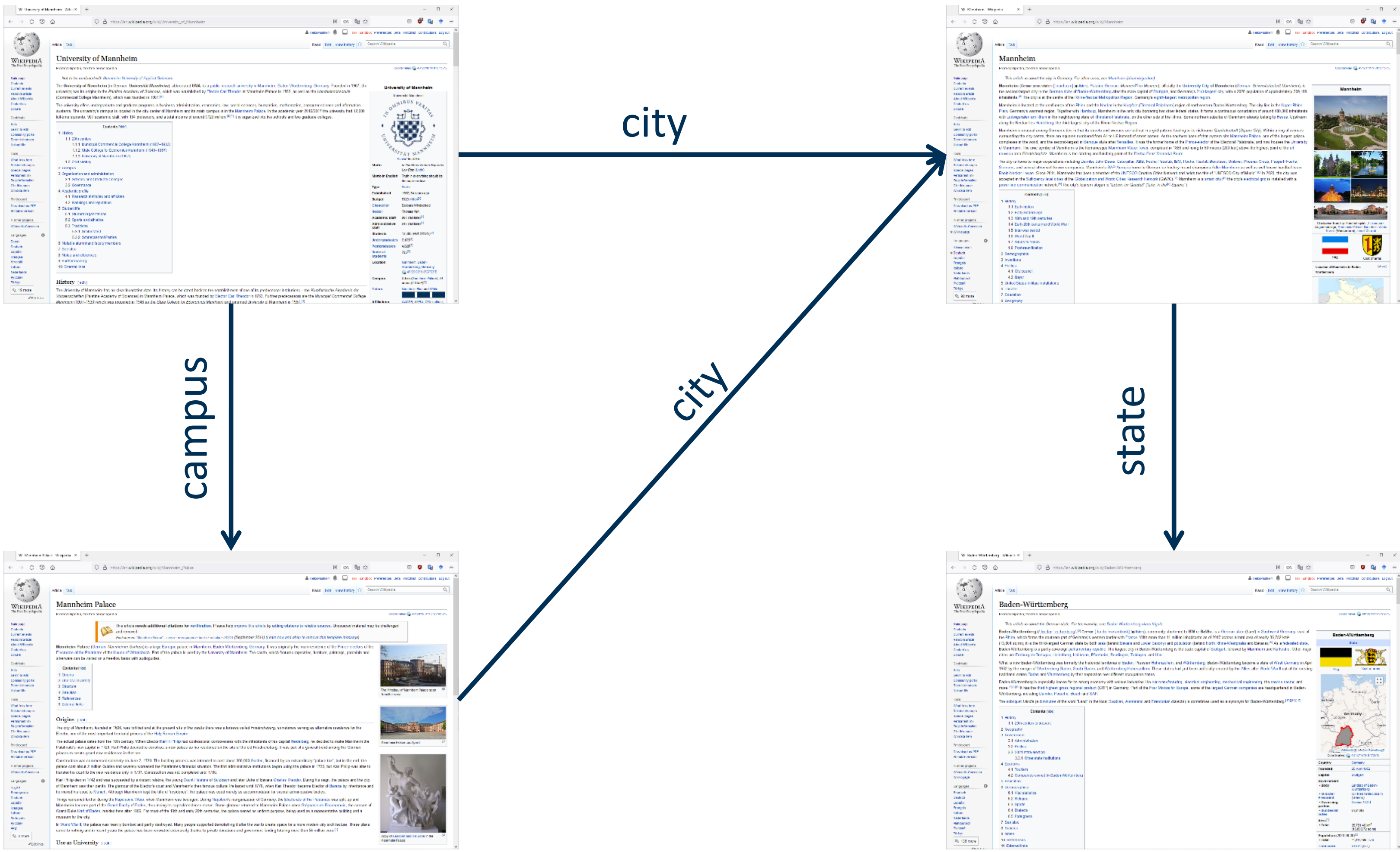
- Present
  - Two of the most used knowledge graphs
  - Next to Wikidata

Wikipedia as a Knowledge Graph



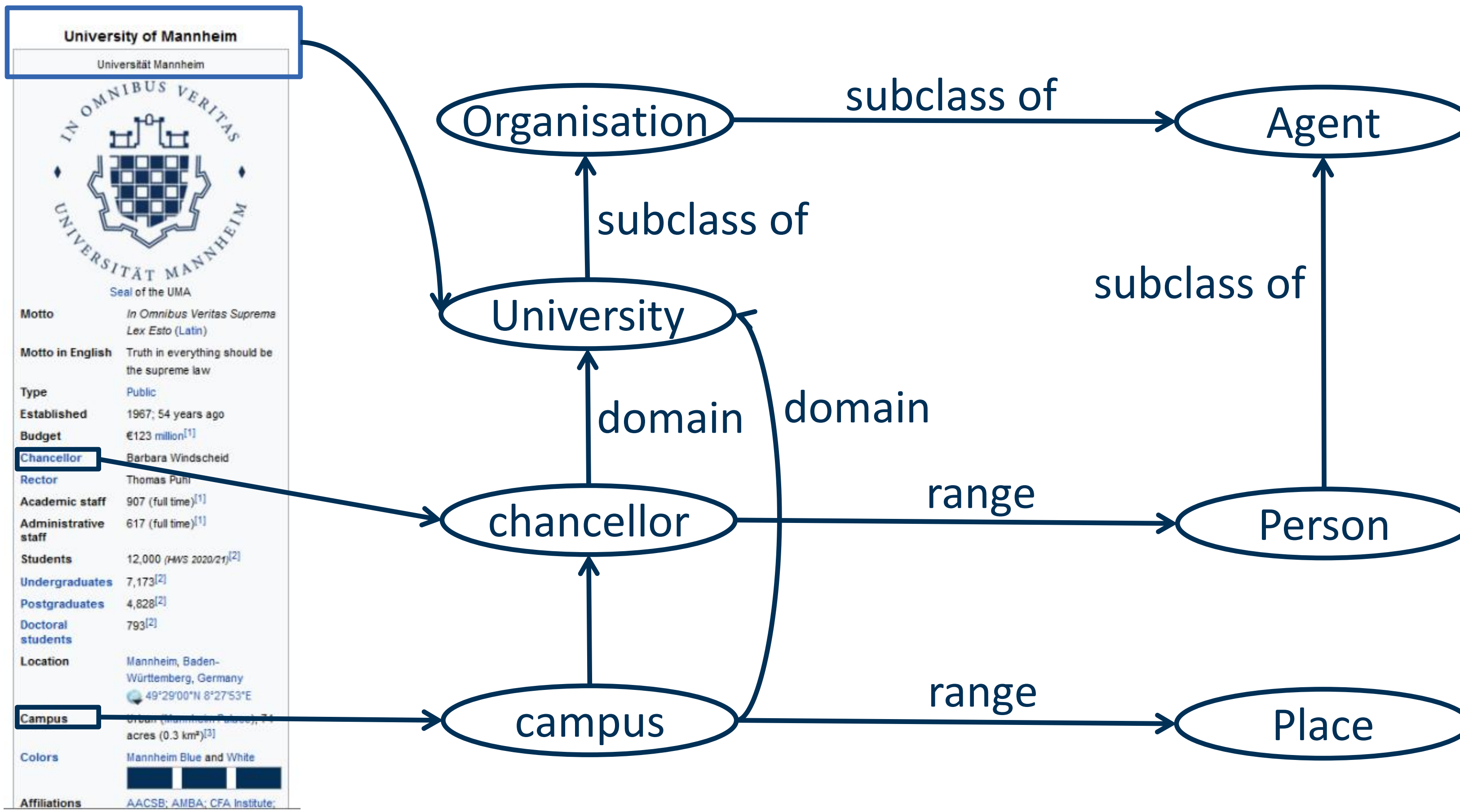
Wikipedia as a Knowledge Graph

UNIVERSITY OF MANNHEIM
Data and Web Science Group



Wikipedia as a Knowledge Graph

- Mapping to a central schema/ontology



DBpedia

University of Mannheim

Universität Mannheim

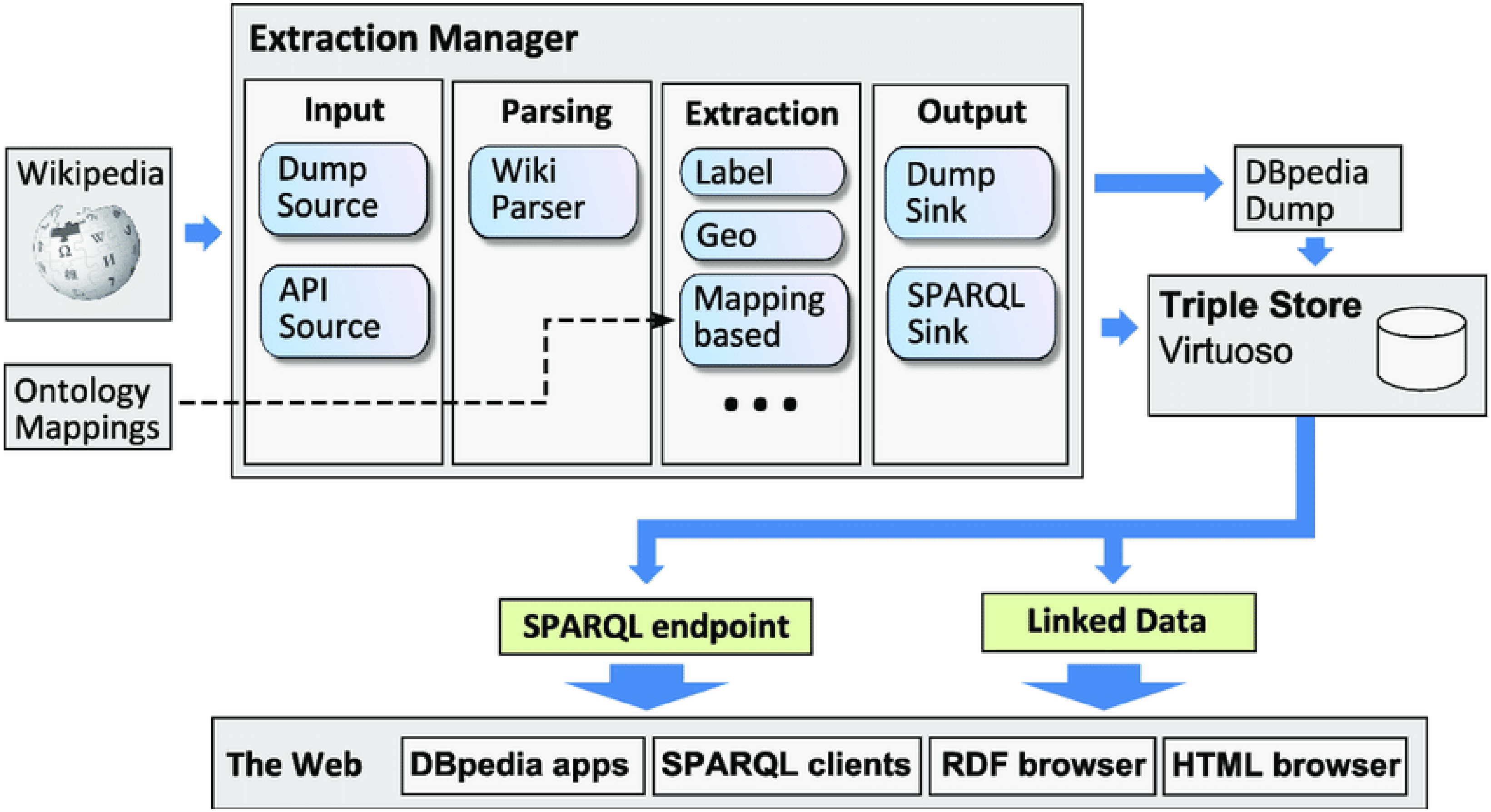
Seal of the UMA

Motto	*In Omnibus Veritas Suprema Lex Esto* (Latin)
Motto in English	Truth in everything should be the supreme law
Established	1763: Theodoro Palatinae 1907: Handelshochschule 1967: Universität Mannheim
Type	Public
Endowment	€115 million
Chancellor	Susann-Annette Storm
Rector	Ernst-Ludwig von Thadden
Academic staff	800 (full time)
Administrative staff	550 (full time)
Students	12,151 *(HWS 2013/14)*[1]
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<re
|doctoral       =249<ref n
|profess        =
|city           =[[Mannhe
|state          =[[Baden-W
|country        =[[German
|coor           = {{Coord
```


```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF>
  <rdf:Description rdf:about="http://dbpedia.org/resource/University_of_Mannheim">
    <dbp:mottoeng xml:lang="en">Truth in everything should be the supreme law</dbp:mottoeng>
    <dbp:chancellor xml:lang="en">Susann-Annette Storm</dbp:chancellor>
    <dbp:budget rdf:datatype="http://dbpedia.org/datatype/euro">1.23E8</dbp:budget>
    <dbp:campus xml:lang="en">Urban , 74 acres</dbp:campus>
    <dbp:caption xml:lang="en">Seal of the UMA</dbp:caption>
    <dbp:city rdf:resource="http://dbpedia.org/resource/Mannheim" />
    <dbp:colors xml:lang="en">Mannheim Blue and White</dbp:colors>
    <dbp:country rdf:resource="http://dbpedia.org/resource/Germany" />
    <dbp:imageName xml:lang="en">Uni_Mannheim_Seal.png</dbp:imageName>
    <dbp:logo xml:lang="en">University of Mannheim.svg</dbp:logo>
    <dbp:logoSize rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">280</dbp:logoSize>
    <dbp:nativeName xml:lang="en">Universität Mannheim</dbp:nativeName>
    <dbp:rector xml:lang="en">Thomas Puhl</dbp:rector>
    <dbp:state rdf:resource="http://dbpedia.org/resource/Baden-Württemberg" />
    <dbp:type rdf:resource="http://dbpedia.org/resource/Public_University" />
    <dbp:website rdf:resource="https://www.uni-mannheim.de/en/%7C2=www.uni-mannheim.de" />
    <georss:point>49.4832 8.4647</georss:point>
    ...
  </rdf:Description>
</rdf:RDF>
```

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

Test this mapping (or in namespace File or Creator) with some example Wikipedia pages. Chec

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 (

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.

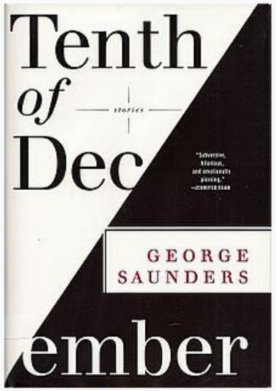
Read more about editing the ontology schema.

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 audiovisuelle, généralement pour le 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

Tenth of December



US release cover

Author	George Saunders
Language	English
Genre	Short story collection
Publisher	Random House
Publication date	January 8, 2013
Publication place	United States
Media type	Print (hardcover)
Pages	208
ISBN	0812993802
OCLC	785558855

```
{{Infobox book
| name = Tenth of December
| author = [[George Saunders]]
| language = English
| country = United States
| genre = [[Short story collection]]
| publisher = [[Random House]]
| isbn = 0812993802
| image = Tenth of December.jpg
| caption = US release cover
| release_date = January 8, 2013
| media_type = Print (hardcover)
| pages = 208
| oclc = 785558855
```

Mapping en:Infobox book

This is the mapping for the Wikipedia template Infobox book. Find usages of this Wikipedia templ
Test this mapping (or in namespace File or Creator) with some example Wikipedia pages. Ch
Read more about mapping Wikipedia templates.

Template Mapping (help)	
map to class	Book

Mappings

Property Mapping (help)	
template property	name
ontology property	foaf:name

Property Mapping (help)	
template property	genre
ontology property	literaryGenre

Property Mapping (help)	
template property	pages
ontology property	numberOfPages

DBpedia Browse using Formats

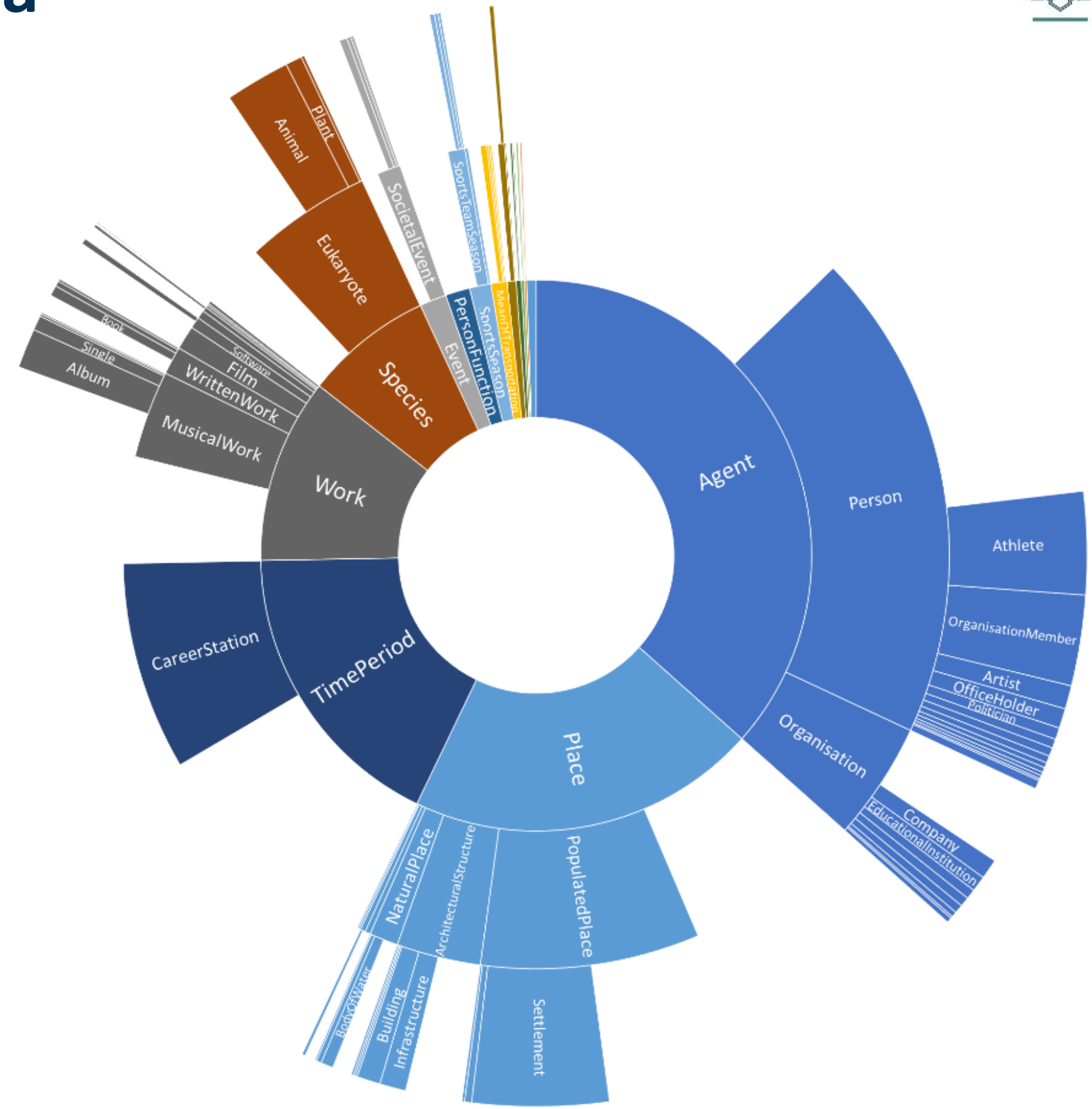
About: Tenth of December: Stories

An Entity of Type: Buch, from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

Tenth of December is a collection of short stories by American author George Sau
magazines between 1995 and 2012. The book was published on January 8, 2013, b
was a 2011 Bram Stoker Award finalist. Tenth of December was selected as one of
New York Times Book Review. The collection also won the 2013 Story Prize for sho
Folio Prize.

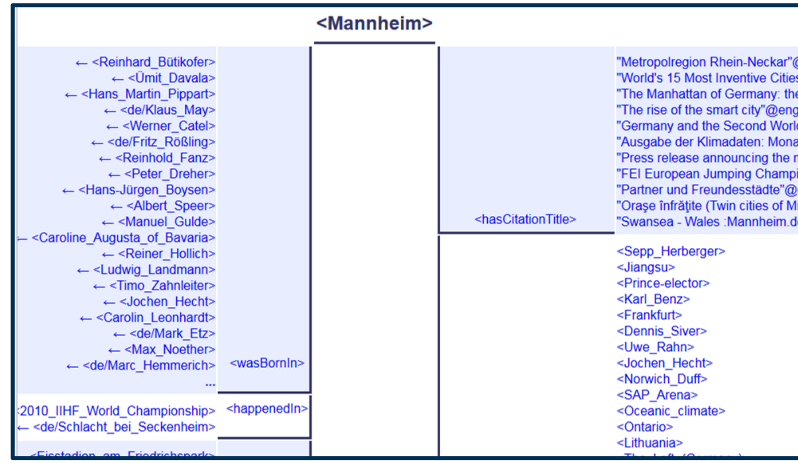
Property	Value
dbo:abstract	• Tenth of December is a collection of short stories by America 1995 and 2012. The book was published on January 8, 2013, Tenth of December was selected as one of the 10 Best Books the 2013 Story Prize for short-story collections and the inaug
dbo:author	• dbr:George_Saunders
dbo:isbn	• 0812993802
dbo:literaryGenre	• dbr:Short_story_collection
dbo:numberOfPages	• 208 (xsd:positiveInteger)
dbo:oclc	• 785558855
dbo:publisher	• dbr:Random_House
dbo:releaseDate	• 2013-01-08 (xsd:date)

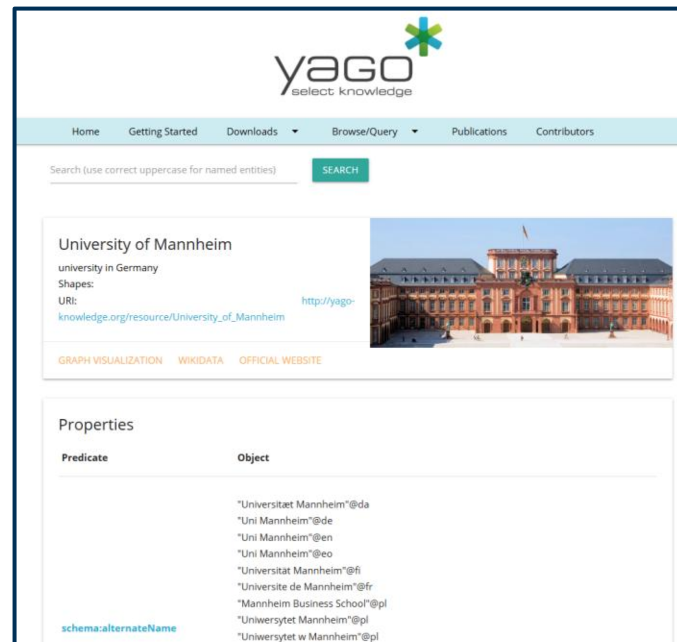
DBpedia


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

YAGO

- YAGO 1
  - Wikipedia categories for types
    - Plus WordNet as upper structure
- YAGO 2
  - Anchored in time and space
- YAGO 3
  - Furth improvements + multiple languages
- YAGO 4
  - Based on Wikidata
    - Refines taxonomy and properties by using schema.org





https://yago-knowledge.org/

YAGO 3

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

Knowledge Graph Creation

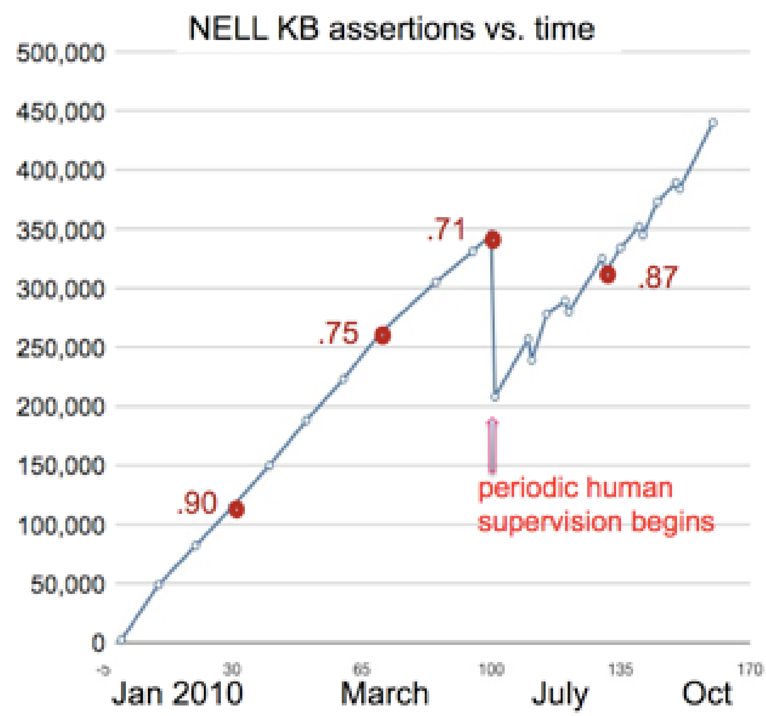
- Lesson learned no. 4:
  - Heuristics help increasing coverage (at the cost of accuracy)

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 2018
  - Continuously ran for ~8 years
  - New release every few days



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

Knowledge Graph Creation: NELL

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Extraction of a Knowledge Graph from a Text Corpus

Nine Inch Nails singer Trent Reznor, bo… 19…

...as stated by Filter singer Richard Pa…

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

patterns

"X singer Y"
→ band_member(X,Y)

facts

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

Recently-Learned Facts twitter Refresh

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

Max. 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!

GPTKB

UNIVERSITY OF MANNHEIM
Data and Web Science Group

GPTKB Browse Query Compare Download Publications Contributors Search...

Welcome to GPTKB

A large general-domain knowledge base entirely from a large language model

100 million triples	6.1 million entities	10x less cost

Overview

This interface presents **GPTKB v1.5** (Hu et al. arXiv, 2025b), a large general-domain knowledge base (KB) entirely from a large language model (LLM). It demonstrates the feasibility of large-scale KB construction from LLMs, while highlighting specific challenges arising around entity recognition, entity and property canonicalization, and taxonomy construction (Hu et al. ACL, 2025a).

Based on GPT-4.1, **GPTKB v1.5** contains 100 million triples for more than 6.1 million entities, at a cost 10x less than previous KBC projects. We also provide GPTKB v1.1 for download.

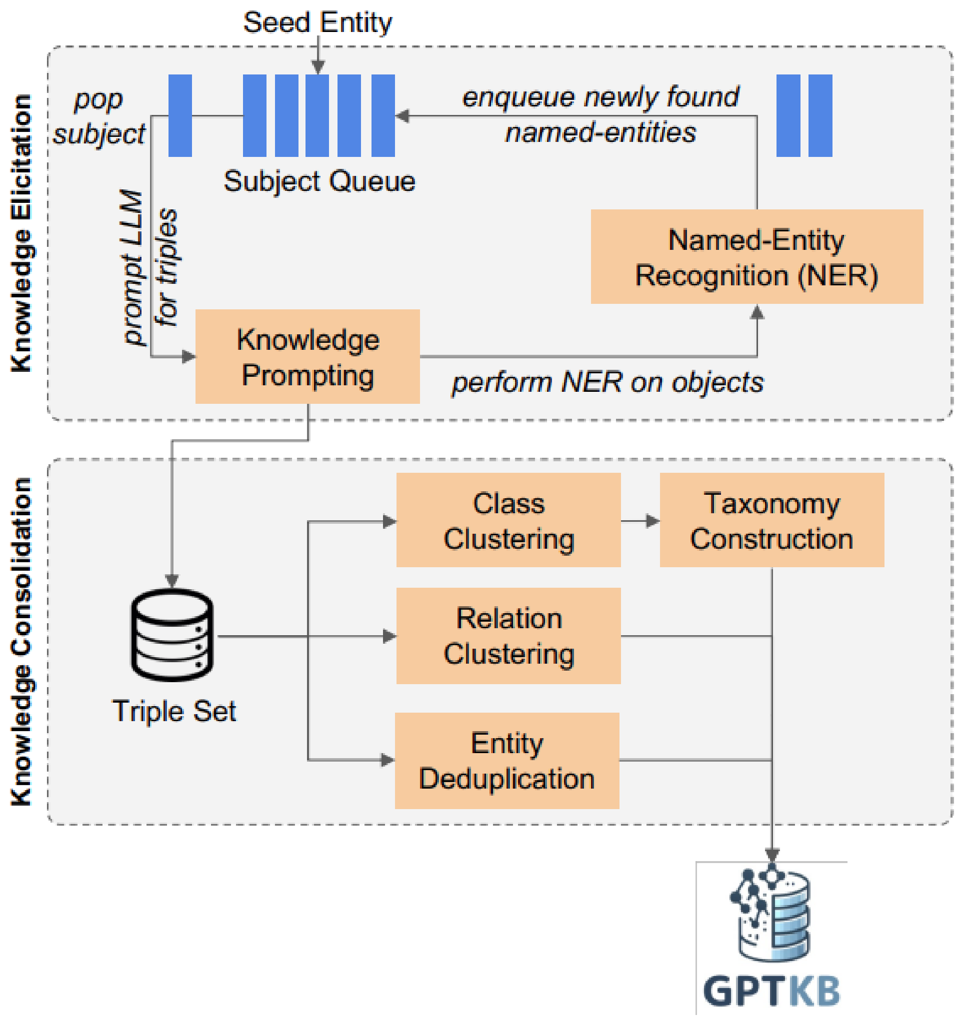
GPTKB is a landmark for two fields:

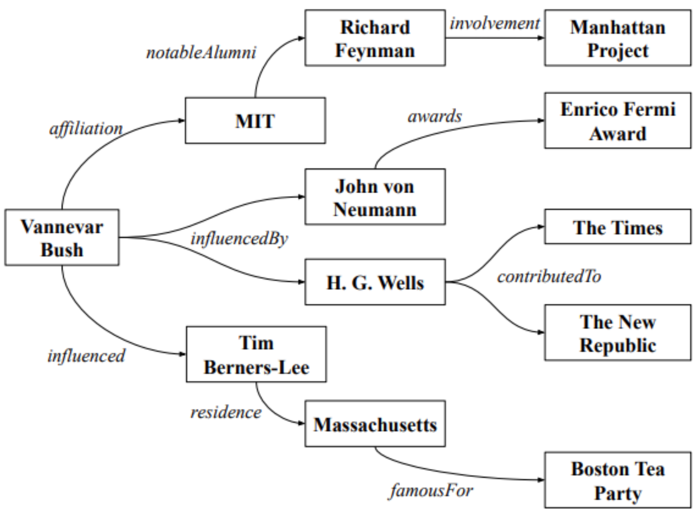
- For NLP, for the first time, it provides constructive insights into the knowledge (or beliefs) of LLMs.
- For the Semantic Web, it shows novel ways forward for the long-standing challenge of general-domain KB construction.

Using the search field, you can find entities by name, and browse their triples. Alternatively, you can write SPARQL queries or download the whole KB as a TTL file. Paper and code are also available.

https://gptkb.org/

GPTKB





https://arxiv.org/pdf/2411.04920

GPTKB

You are a knowledge base construction expert. Given a subject entity, return all facts that you know for the subject as a list of subject, predicate, object triples. The number of facts may be very high, between 50 to 100 or more, for very popular subjects. For less popular subjects, the number of facts can be very low, like 5 or 10.

Important:
- If you don't know the subject, return an empty list.
- If the subject is not a named entity, return an empty list.
- If the subject is a named entity, include at least one triple where predicate is "instanceOf".
- Do not get too wordy.
- Separate several objects into multiple triples with one object.

Figure 9: Prompt for knowledge elicitation.

You are an expert on named entity recognition (NER). Your task is to classify if given phrases are named entities (e.g., persons, organizations, works of art), or not (e.g., literals, dates, URLs, verbose phrases). Each phrase is given to you in a line.

Figure 10: Prompt for named-entity recognition (NER).

GPTKB

Entities	2.9M
Triples	101M
Relations	567K (788K before canonicalization)
Classes	4,715 (103K before canonicalization)
Triple objects	37M entities, 64M literals
Avg. triples/entity	35
Avg. label length	21.8 characters
Avg. outlinks	4
Subject-precision*	74% Verifiable, 9% Plausible, 17% Unverifiable
Subjects in Wikidata*	37% in WD, 63% not in WD
Triple-precision*	31% True, 61% Plausible, 1% Implausible, 7% False

* *as a weighted average across layers.*

https://arxiv.org/pdf/2411.04920

Non-Public Knowledge Graphs

UNIVERSITY OF MANNHEIM
Data and Web Science Group

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

- However, we usually know only little about them

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

Comparison of Knowledge Graphs

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Release cycles

Instant updates: DBpedia live, Freebase, Wikidata | Days: NELL (Caution!) | Months: DBpedia | Years: YAGO, Cyc

- Size and density

2017

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

2020

	DBpedia	YAGO	Wikidata	BabelNet
# Instances	5,044,223	6,349,359	52,252,549	7,735,436
# Assertions	854,294,312	479,392,870	732,420,508	178,982,397
Avg. linking degree	21.30	48.26	6.38	0.00
Median ingoing edges	0	0	0	0
Median outgoing edges	30	95	10	9
# Classes	760	819,292	2,356,259	6,044,564
# Relations	1355	77	6,236	22
Avg. depth of class tree	3.51	6.61	6.43	4.11
Avg. branching factor of class tree	4.53	8.48	36.48	71.0
Ontology complexity	SHOFD	SHOIF	SOD	SO

	Cyc	NELL	CaLiGraph	Voldemort
# Instances	122,441	5,120,688	7,315,918	55,861
# Assertions	2,229,266	60,594,443	517,099,124	693,428
Avg. linking degree	3.34	6.72	1.48	0
Median ingoing edges	0	0	0	0
Median outgoing edges	3	0	1	5
# Classes	116,821	1,187	755,963	621
# Relations	148	440	271	294
Avg. depth of class tree	5.58	3.13	4.74	3.17
Avg. branching factor of class tree	5.62	6.37	4.81	5.40
Ontology complexity	SHOIFD	SROIF	SHOD	SH

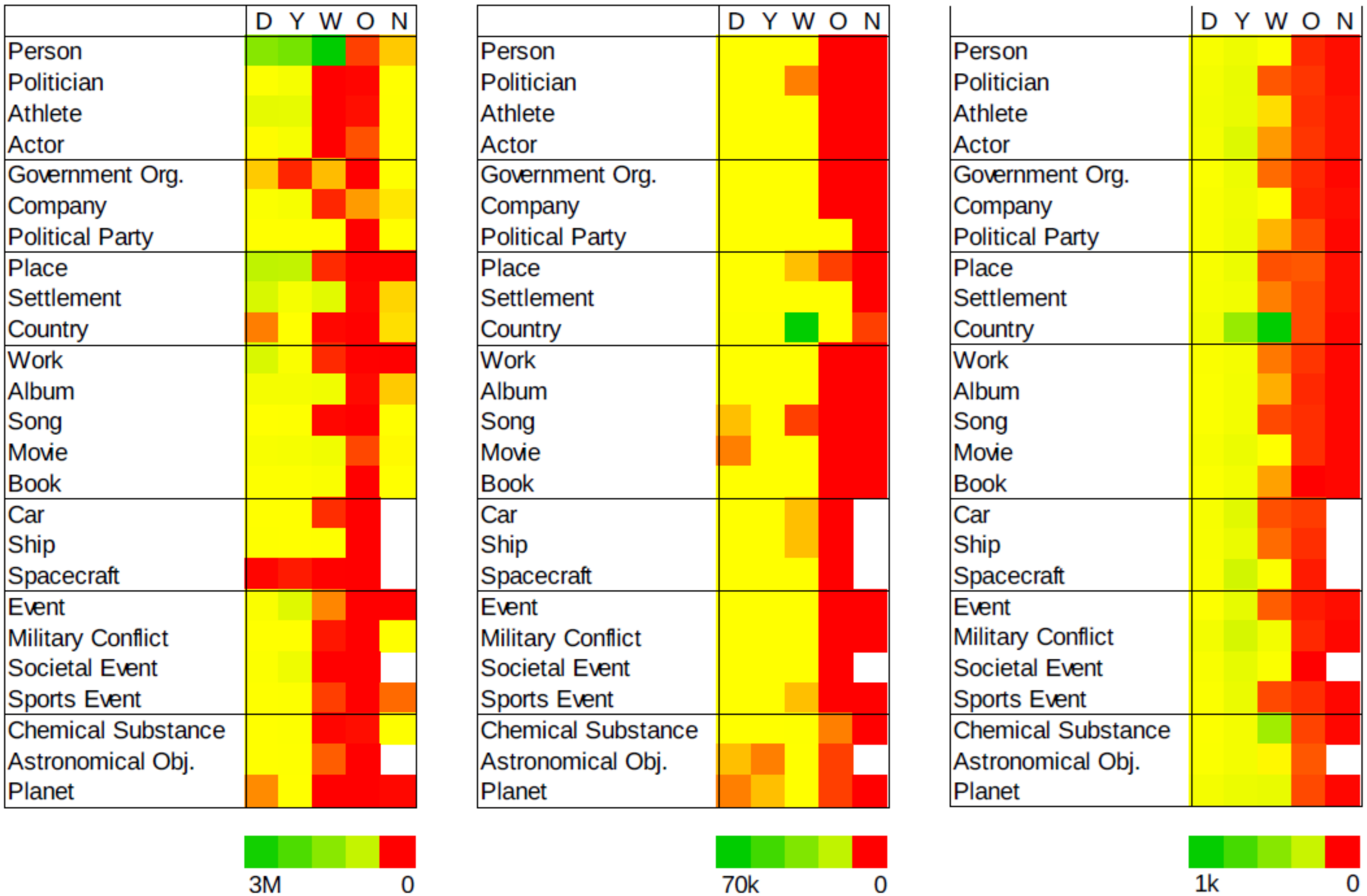
Heist, Hertling, Ringler & Paulheim: Knowledge Graphs on the Web – an Overview. 2020

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

UNIVERSITY OF MANNHEIM
Data and Web Science Group



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

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

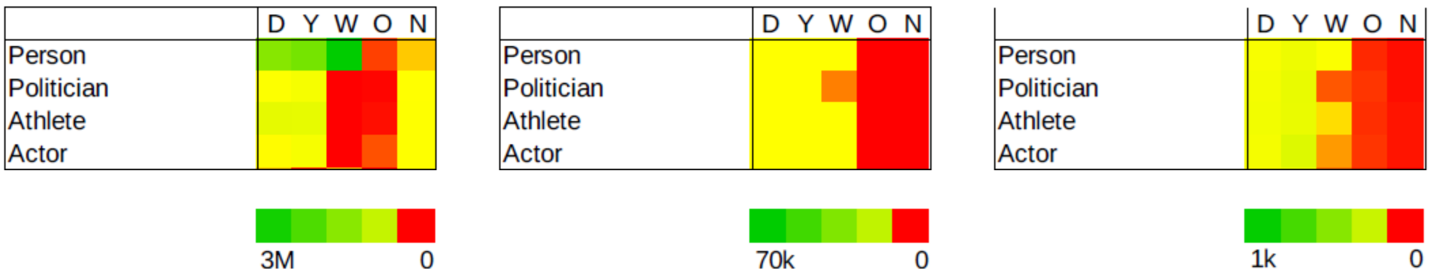
Comparison of Knowledge Graphs

- 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…



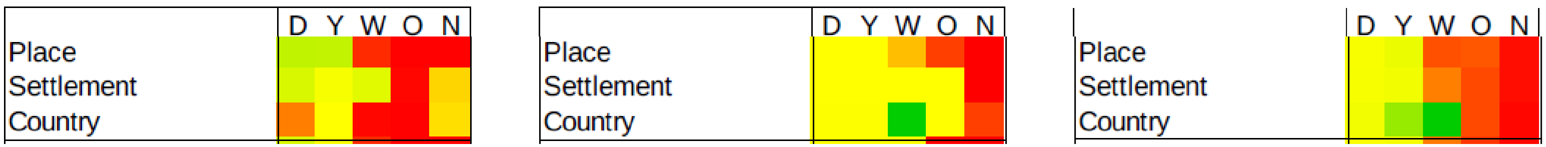
(a) Number of instances (b) Average indegree (c) Average outdegree

- 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.)

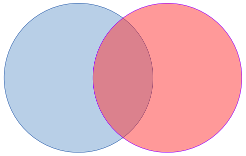


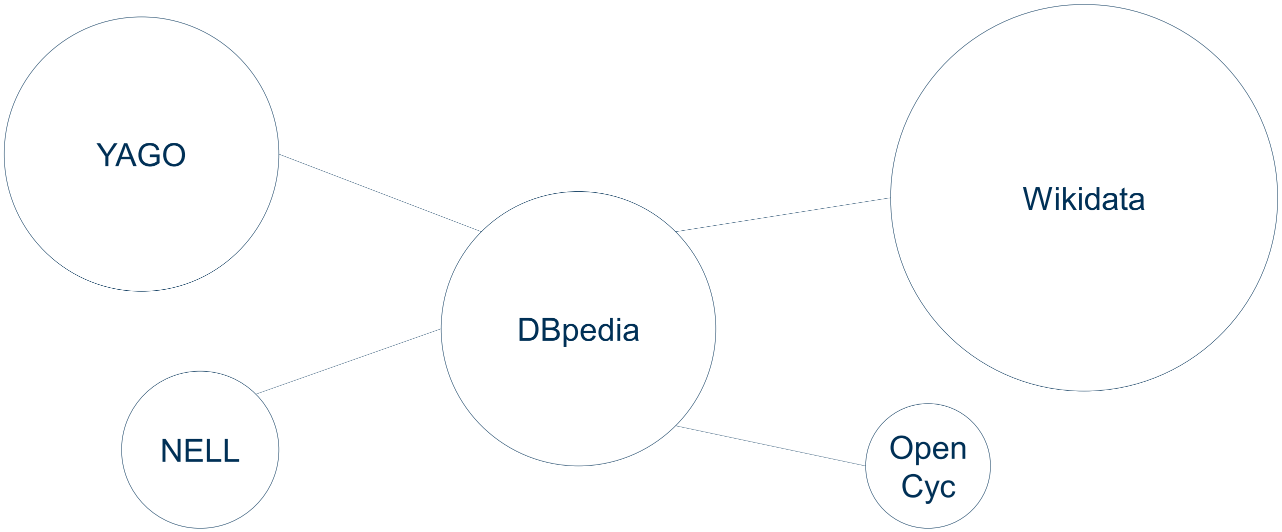
(a) Number of instances (b) Average indegree (c) Average outdegree

- 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

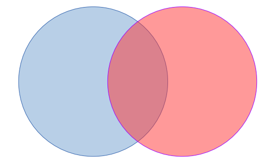
- How largely do knowledge graphs overlap?
- They are interlinked, so we can simply count links
  - For NELL, we use links to Wikipedia as a proxy





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

Overlap of Knowledge Graphs

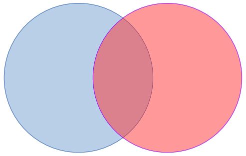


- How largely do knowledge graphs overlap?
- They are interlinked, so we can simply count links
  - For NELL, we use links to Wikipedia as a proxy

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

Overlap of Knowledge Graphs

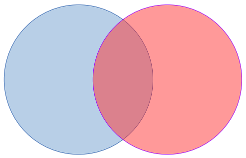
- Links between Knowledge Graphs are incomplete
  - The Open World Assumption also holds for interlinks



- 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



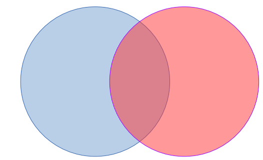
- Idea:
  - Given that the link set F is found
  - And the (unknown) actual link set would be C

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

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

Overlap of Knowledge Graphs



- Mathematical derivation:
  - Definition of recall: $R = \frac{|F_{correct}|}{|C|}$

    unknown

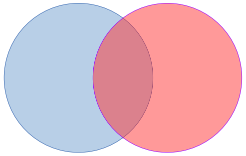
  - Definition of precision: $\mathrm{P} = \frac{|F_{correct}|}{|F|}$
- Resolve both to $|F_{correct}|$, substitute, and resolve to $|C|$

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

- The existing links are used to compute precision and recall

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

Overlap of Knowledge Graphs

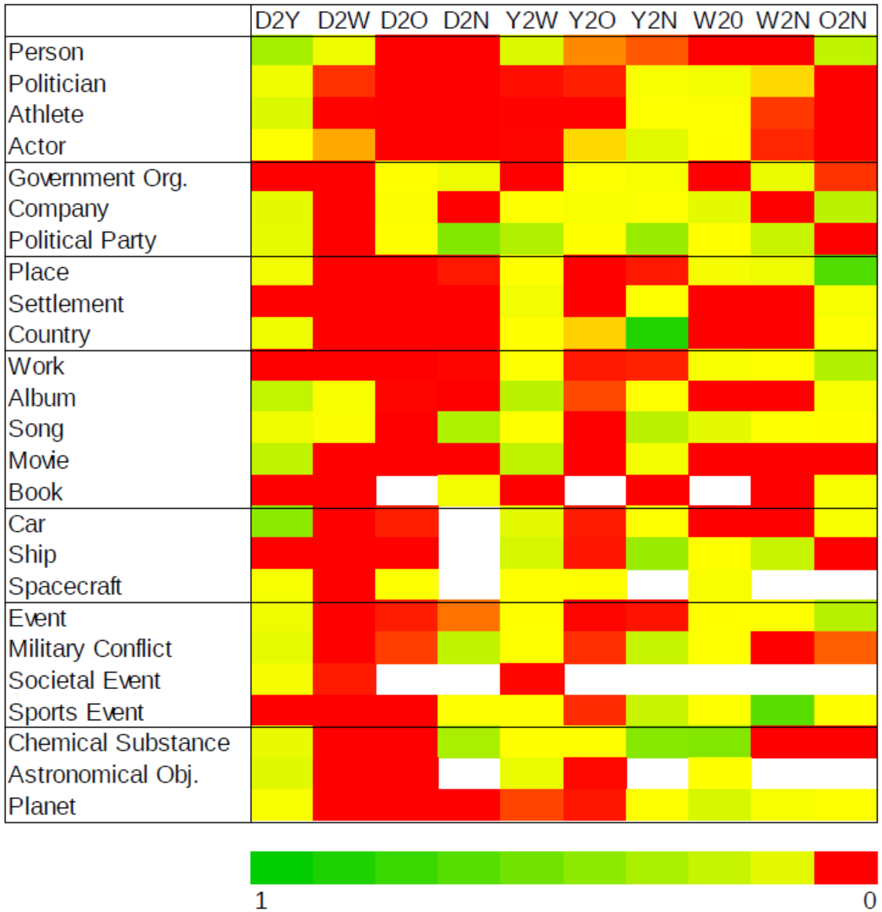


- Experiment:
  - We use the same 25 classes as before
  - Measure 1: overlap relative to smaller KG (i.e., potential gain)
  - Measure 2: overlap relative to explicit links
    (i.e., importance of improving links)

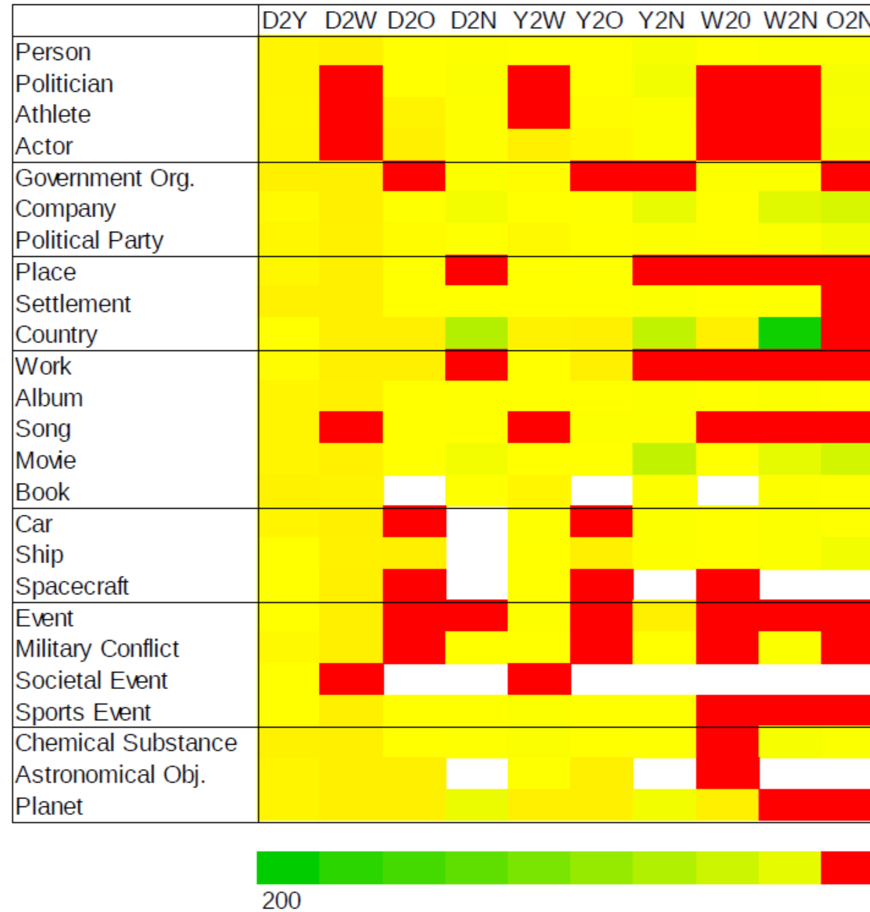
- Link generation with 16 different metrics and thresholds
  - Intra-class correlation coefficient for |C|: 0.969
  - Intra-class correlation coefficient for |F|: 0.646
- Bottom line:
  - Despite variety in link sets generated, the overlap is estimated reliably
  - The link generation mechanisms do not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

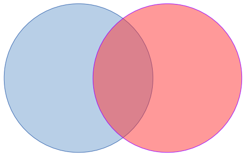


(b) Overlap relative to existing links

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

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

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	

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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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**

acquisition by Google estimated as $60-300M

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

Paulheim: How much is a triple? Estimating the Cost of Knowledge Graph Creation, 2018

Intermezzo: Knowledge Graph Creation Cost

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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

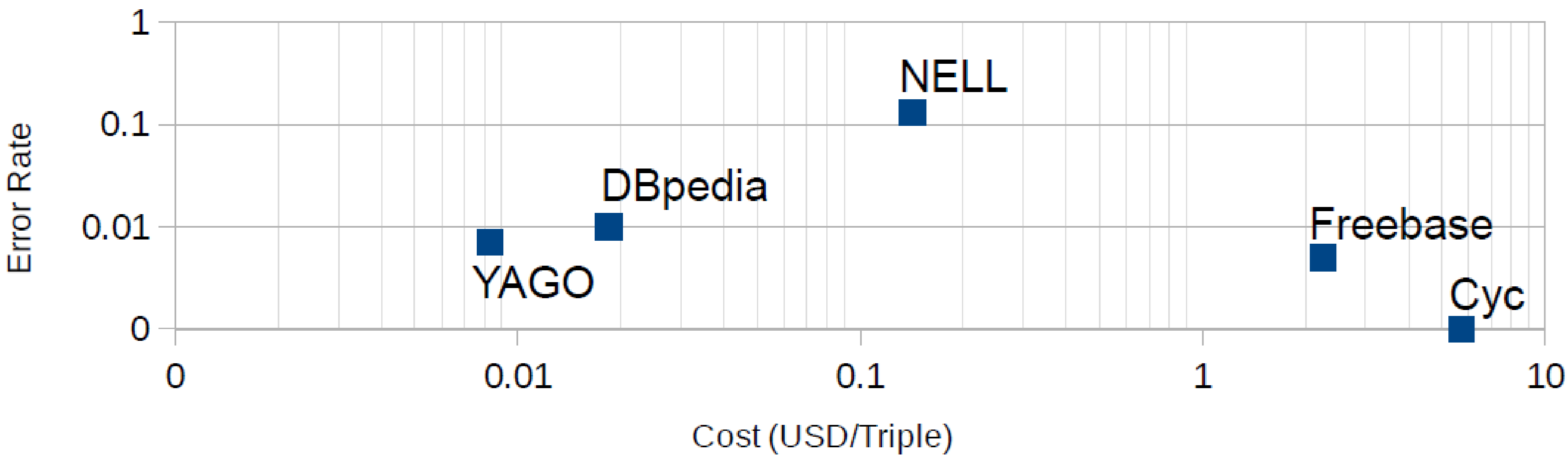
    → 0.83c per statement
  - NELL: 103k LOC

    → 14.25c per statement
- Compared to manual curation: saving factor 16-250

Paulheim: How much is a triple? Estimating the Cost of Knowledge Graph Creation, 2018

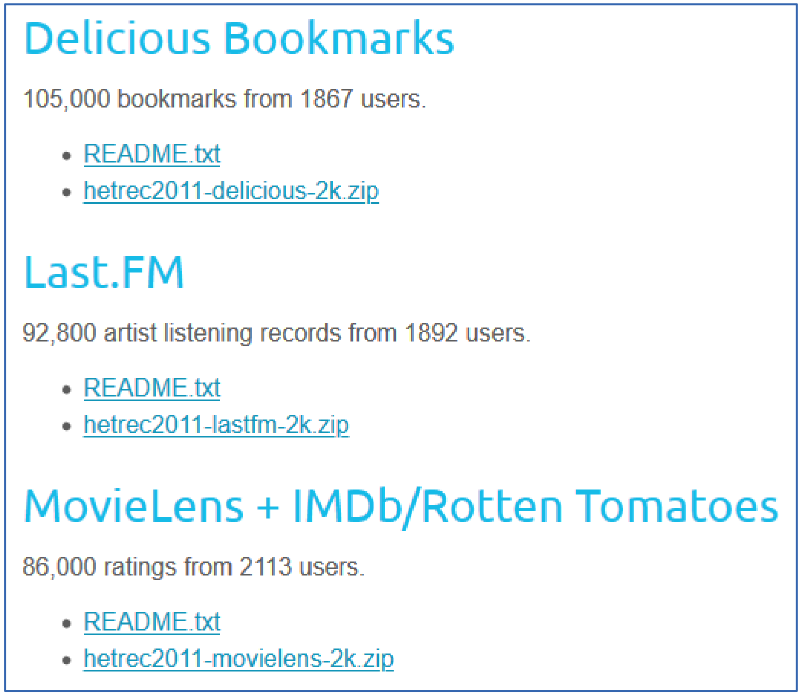
Intermezzo: Knowledge Graph Creation Cost

- Graph error rate against cost
  - we can pay for accuracy
  - NELL is a bit of an outlier



Enhancing the Coverage of Knowledge Graphs

- Study for KG-based Recommender Systems*
  - DBpedia (likewise: YAGO) has a coverage of
    - 85% for movies
    - 63% for music artists
    - 31% for books



https://grouplens.org/datasets/

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

University of Mannheim
Data and Web Science Group

Enhancing the Coverage of Knowledge Graphs

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

Delicious Bookmarks

105,000 bookmarks from 1867 users.

- README.txt
- hetrec2011-delicious-2k.zip

Last.FM

92,800 artist listening records from 1892 users.

- README.txt
- hetrec2011-lastfm-2k.zip

MovieLens + IMDb/Rotten Tomatoes

86,000 ratings from 2113 users.

- README.txt
- hetrec2011-movielens-2k.zip

https://grouplens.org/datasets/

List of intelligent dance music artists

From Wikipedia, the free encyclopedia

This section **does not cite any sources**. Please help improve this section by adding ci
Find sources: "List of intelligent dance music artists" – news · newspapers · books · scholar · JSTOR

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

Contents [hide]

[edit]

- 808 State
- µ-ziq

A-K [edit]

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

Entity Extraction from List Pages

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Lists form (shallow) hierarchies

Lists of writers

From Wikipedia, the free encyclopedia

The following are **lists of writers**:

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

Top of page

Contents [hide]

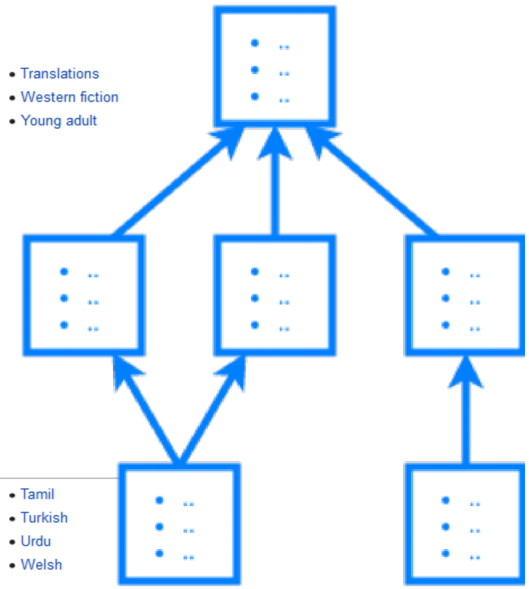
Lists by language (non-English) [edit]

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

Top of page

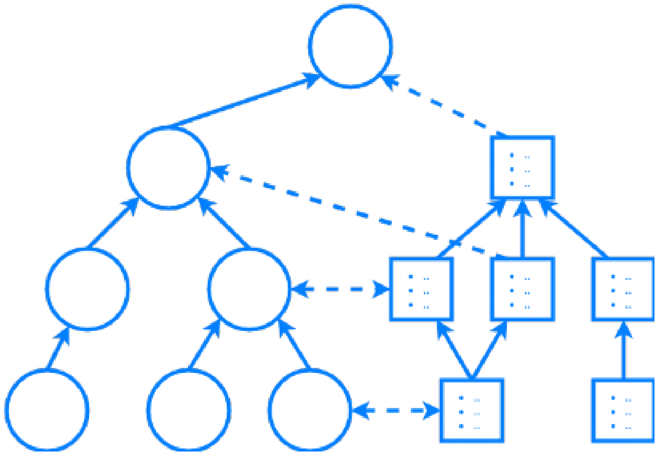
Lists by ethnicity or nationality [edit]

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



Entity Extraction from List Pages

- Idea: Align with category graph
- Equivalence:
  - "List of Japanese Writers"
    ↔ Category:Japanese Writers
- Subsumption:
  - "List of Japanese
    Speculative Fiction Writers"
    → Category:Japanese Writers



Classifying Red Links

- Not all entities on a list page belong to the same category
- Idea:
  - Learn classifier to tell subject entities from non-subject entities
- Distant learning approach
  - Positive examples:
    - Entities that are in the corresponding category
  - Negative examples
    - Entities that are in a category which is disjoint
    - e.g., Book <> Writer

- Patricia Aakhus (1952–2012), *The Voyage of Mael Duin's Curragh*
- Atia Abawi
- Edward Abbey (1927–1989), *The Monkey Wrench Gang*
- Lynn Abbey (born 1948), *Daughter of the Bright Moon*
- Belle Kendrick Abbott (1842–1893), *Leah Mordecai*
- Eleanor Hallowell Abbott (1872–1958), poet, novelist and short story writer
- Hailey Abbott, *Summer Boys*
- Megan Abbott (born 1971), *Die A Little*
- Shana Abé, *A Rose in Winter*
- Louise Abeita (1926–2014), Native American Isleta Pueblo writer, *I am a Pueblo Indian Girl*
- Robert H. Abel (1941–2017)
- Aberjhani
- Walter Abish (born 1931), *How German Is It*
- Abiola Abrams (born 1976), TV host, art filmmaker and author, *Dare*
- Diana Abu-Jaber (born 1960), *Arabian Jazz*
- Susan Abulhawa, *Mornings in Jenin*
- Kathy Acker (1947–1997), *Blood and Guts in High School*
- Cherry Adair, *Black Magic*
- Alice Adams (1926–1999), *Beautiful Girl*
- Victoria Aveyard (born 1990), *Red Queen* series

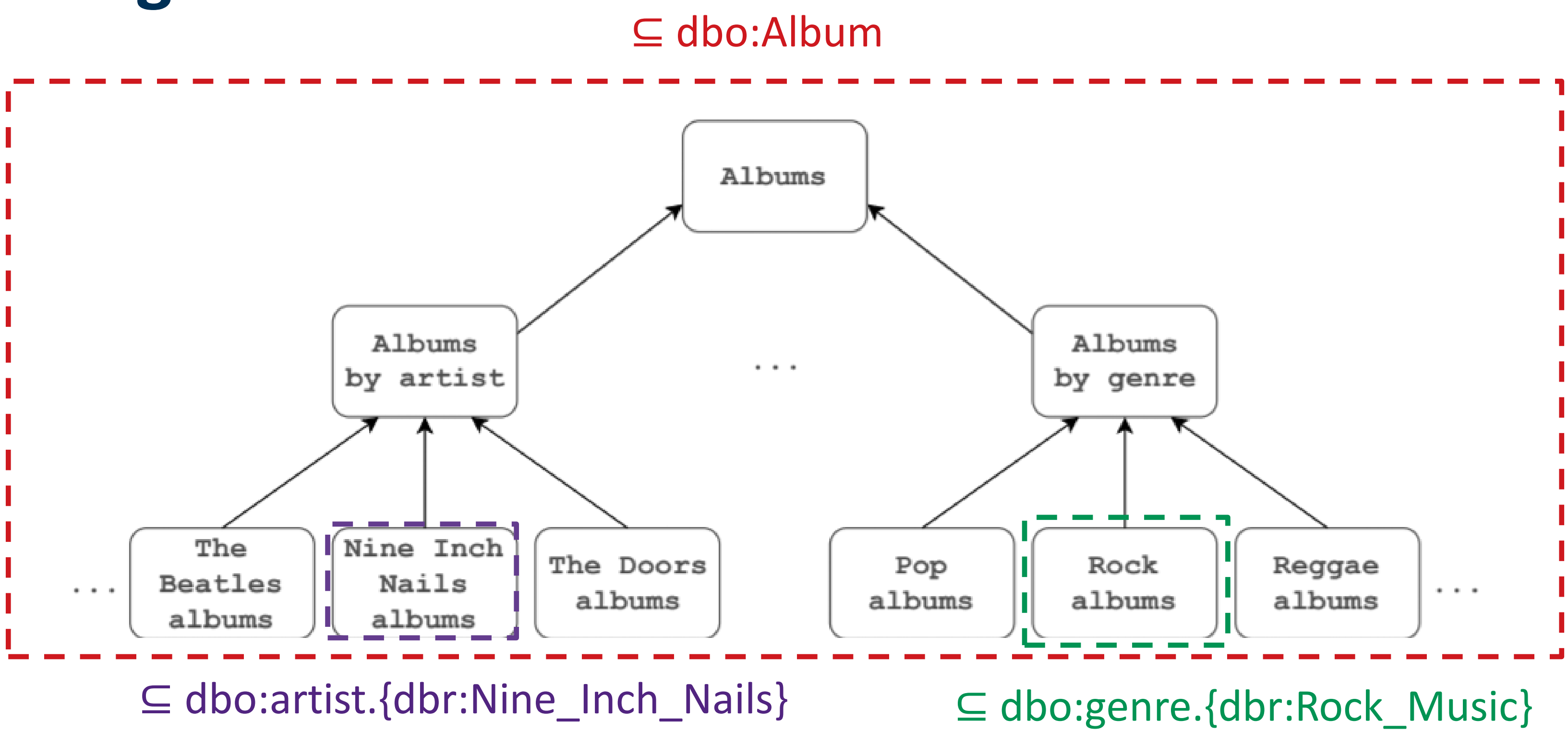
Increasing Level of Detail

- YAGO uses categories for types
  - e.g., Category:American Industrial Groups
  - but does not analyze them further

- `:NineInchNails a :AmericanIndustrialGroup.`
  - “Things, not Strings”?

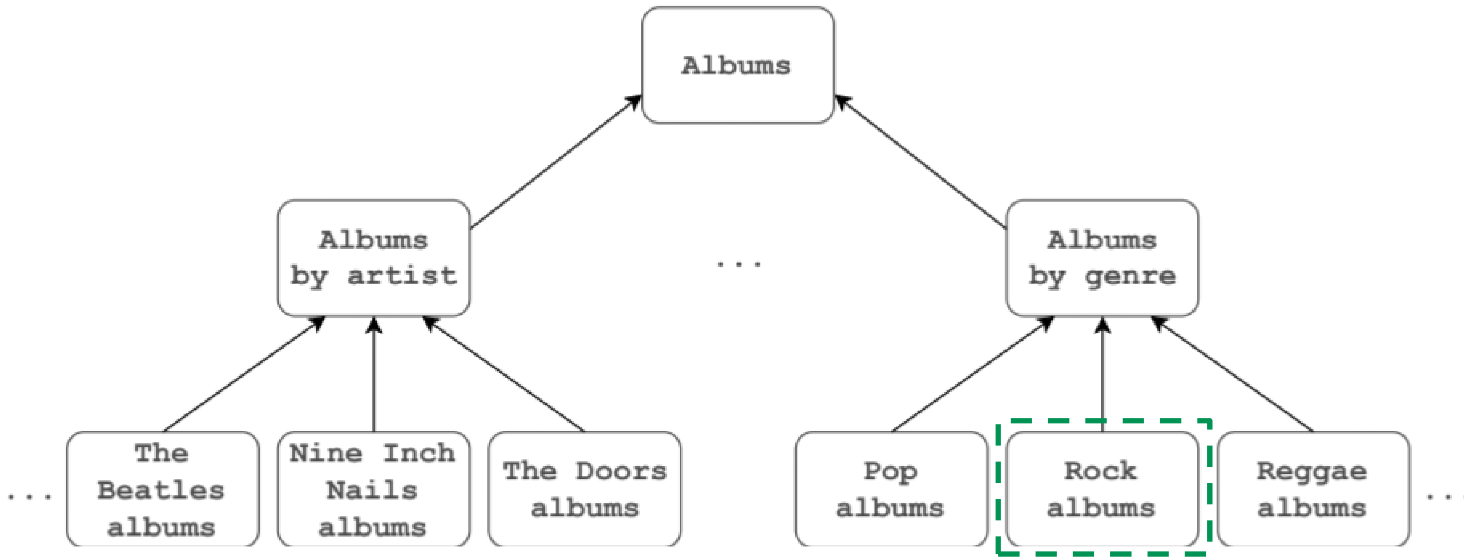
- ```
  :NineInchNails a :MusicalGroup ;
                   hometown :United_States ;
                   genre :Industrial .
  ```
  ```

Cat2Ax: Axiomatizing Wikipedia Categories



Heist & Paulheim (2019): Uncovering the Semantics of Wikipedia Categories

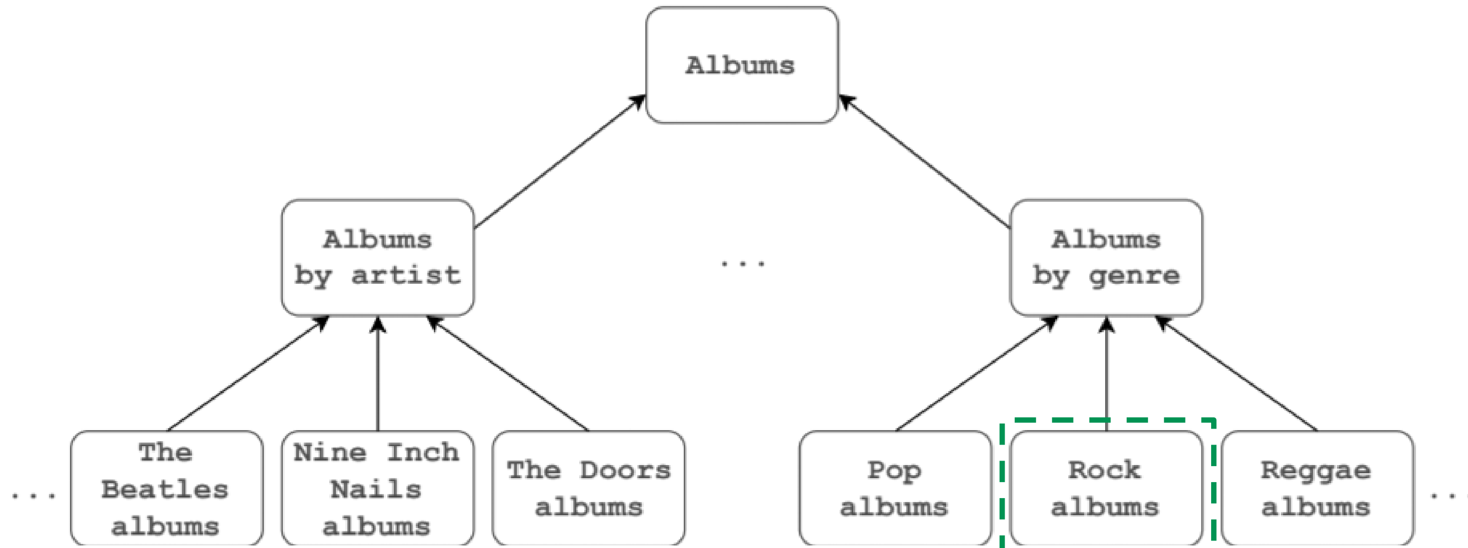
Cat2Ax: Axiomatizing Wikipedia Categories



⊆ dbo:genre.{dbr:Rock_Music}
⊆ dbo:genre.{dbr:Rock_(Rapper)}

Heist & Paulheim (2019): Uncovering the Semantics of Wikipedia Categories

Cat2Ax: Axiomatizing Wikipedia Categories



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

Heist & Paulheim (2019): Uncovering the Semantics of Wikipedia Categories

CaLiGraph Example

rdfs:label	• Tiamat
owl:sameAs	• dbr:Tiamat_(band)
clgo:activeYearsStartYear	• 1987
clgo:genre	• Symphonic metal
clgo:hometown	• Sweden

Category: Musical Groups established in 1987

List of symphonic metal bands

Category: Swedish death metal bands
List of Swedes in Music

Pushing Entity Coverage Further

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Beyond red links (2020)

Cinematic films

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

- Beyond explicit lists (2021)

Members [edit]

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

Former members [edit]

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

Discography [edit]

Albums [edit]

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

Entity Extraction from List Pages

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

Cinematic films

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

Beyond List Pages

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Many pages contain list-like constructs
- Usually
  - Small
  - Same type
  - Same relation to page entity
  - More grey links

Axl Rose

From Wikipedia, the free encyclopedia

...

Discography [edit]

with Guns N' Roses [edit]

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

with Hollywood Rose [edit]

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

with Rapidfire [edit]

- *Ready to Rumble* EP (2014)

Guest appearances [edit]

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

Filmography [edit]

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

Beyond List Pages

UNIVERSITY OF MANNHEIM
Data and Web Science Group

(artist, Axl_Rose)

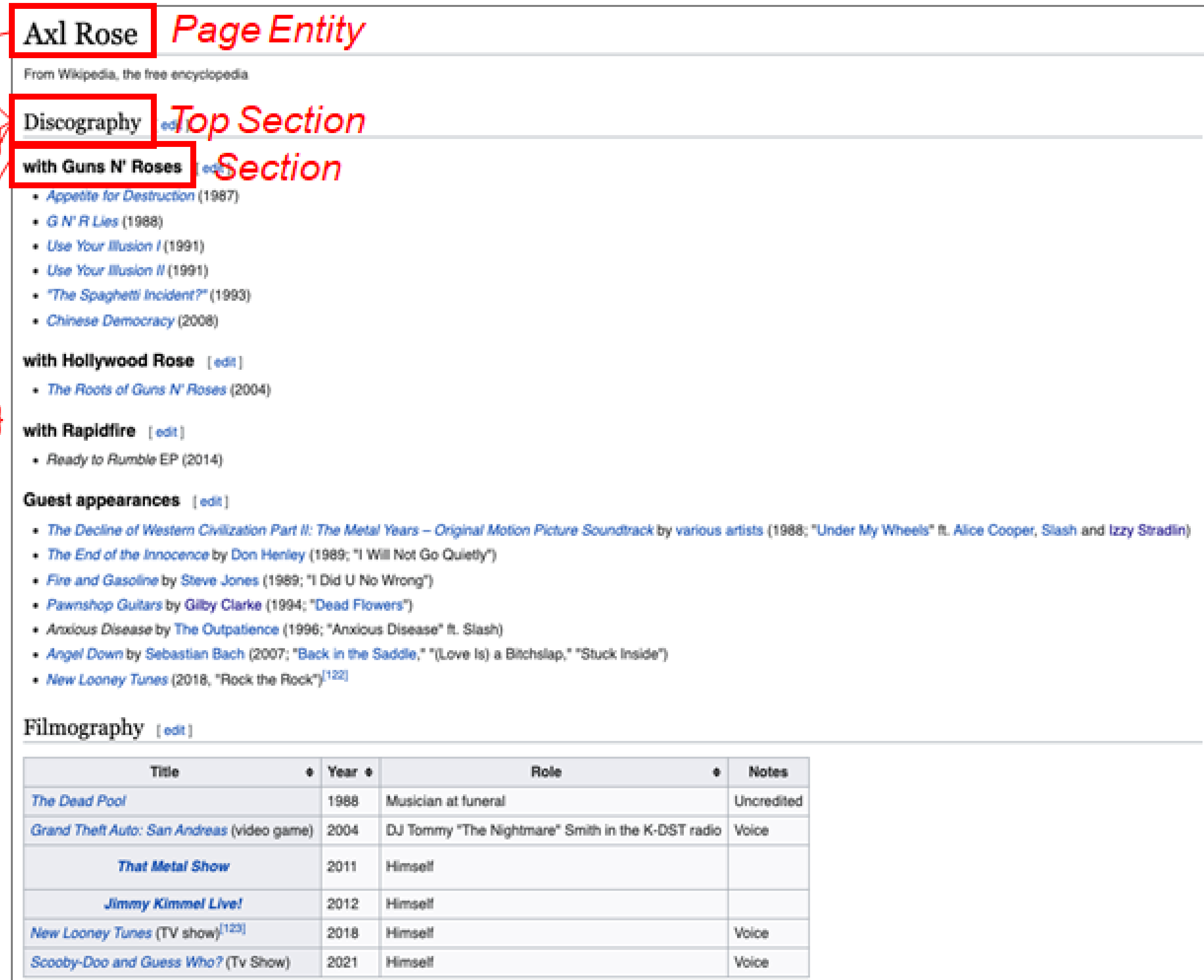
$\exists topSection.\{„Discography“\}$
$\sqsubseteq \exists artist.\{$>PageEntity<$\}$

(type, MusicalWork)

$\exists topSection.\{„Discography“\}$
$\sqsubseteq MusicalWork$

(musicalBand, Guns_N'_Roses)

$\exists topSection.\{„Discography“\}$
$\sqcap\ sectionEntityType.\{Band\}$
$\sqsubseteq$
$\exists musicalBand.\{$>SectionEntity<$\}$



Axl Rose — Page Entity

From Wikipedia, the free encyclopedia

Discography [edit] — Top Section

with Guns N' Roses [edit] — Section

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

with Hollywood Rose [edit]

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

with Rapidfire [edit]

- *Ready to Rumble* EP (2014)

Guest appearances [edit]

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

Filmography [edit]

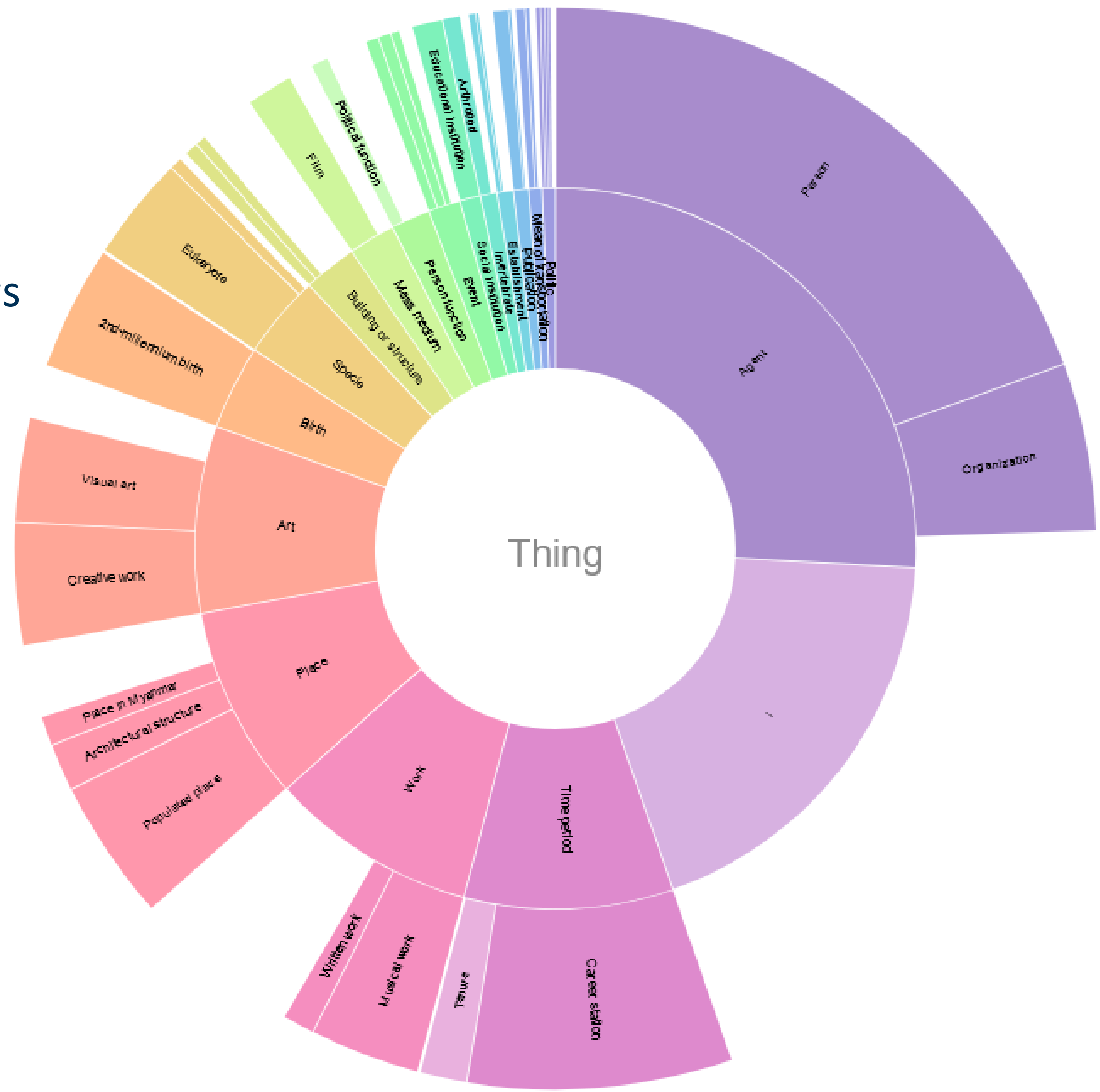
Title	Year	Role	Notes
The Dead Pool	1988	Musician at funeral	Uncredited
Grand Theft Auto: San Andreas (video game)	2004	DJ Tommy "The Nightmare" Smith in the K-DST radio	Voice
That Metal Show	2011	Himself	
Jimmy Kimmel Live!	2012	Himself	
New Looney Tunes (TV show)[123]	2018	Himself	Voice
Scooby-Doo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

Beyond List Pages

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

CaLiGraph at a Glance

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



Entity Disambiguation

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

Members [edit]

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

Former members [edit]

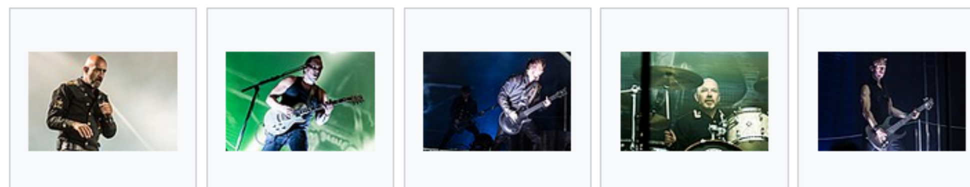
- Bradley Bills – live drums (2013–2014)
- Rüdiger Esch – bass guitar (1989–1997, 2005–2011)
- Christoph "Nook" Michelfeit – drums, electronic percussion
- Bernward Malaka – bass guitar (1980–1982)
- Hendrik Thiesbrummel – live drums (2016–2018)
- Frank Köllges – drums
- Eva Gossling – saxophone (1981)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

?

Members [edit]

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

Eisbrecher, line-up at Rockharz Open Air 2018



Alexander "Alexx" Wesselsky | Jochen "Noel Pix" Seibert | Jürgen Plangger | Achim Färber | Rupert Keplinger

Former live members [edit]

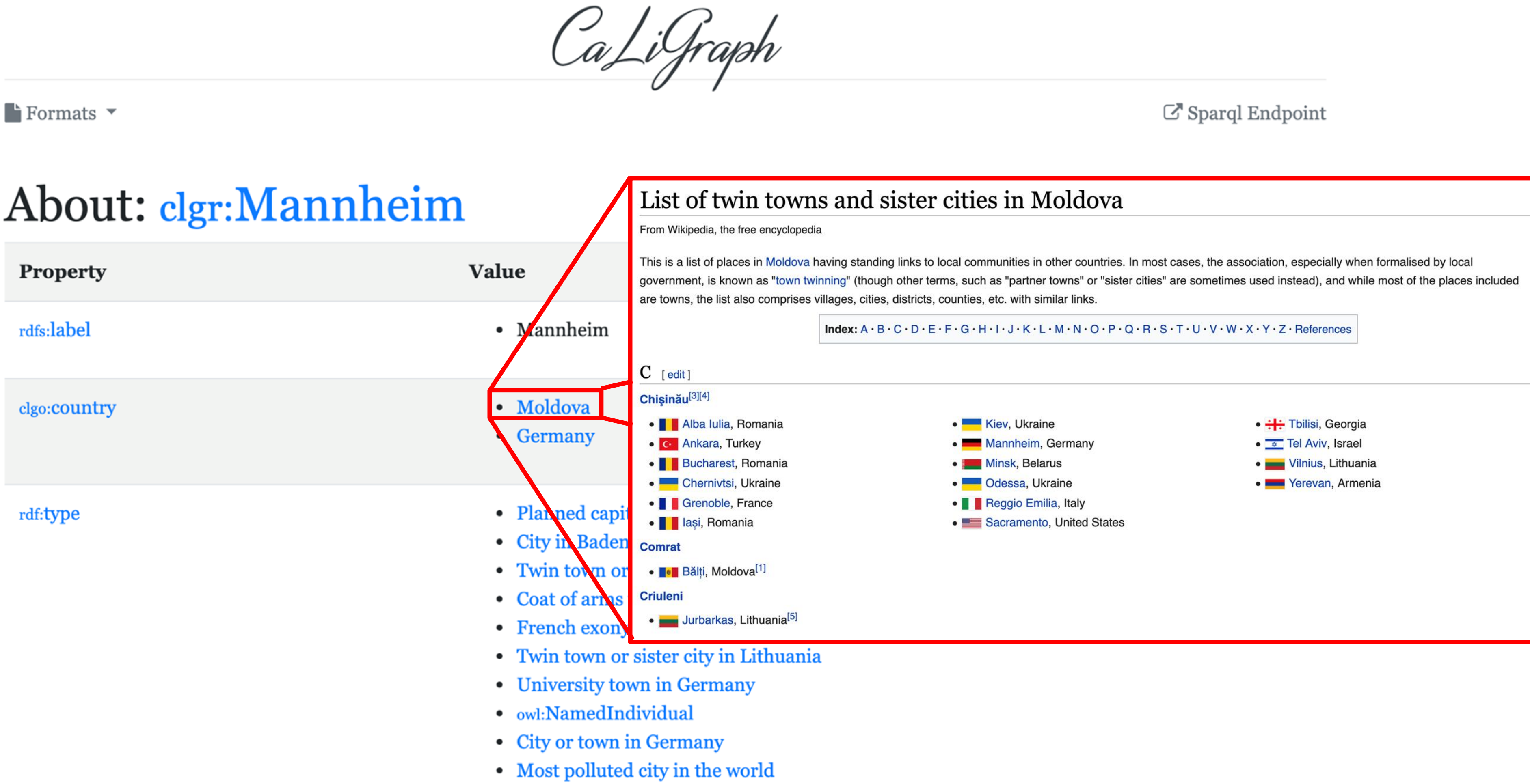
- Felix Primc – rhythm guitar (2003–2007)
- Micheal Behnke – bass (2003–2007)
- Martin Motnik – bass (2007–2008)
- Olli Pohl – bass (2008–2010, 2015)
- Dominik Palmer – bass (2010–2013)
- Rene Greil – drums (2003–2011)

Touring members [edit]

- Sebastien Angrand – drums (2010)

CaLiGraph Glitches

UNIVERSITY OF MANNHEIM
Data and Web Science Group



CaLiGraph

Formats

Sparql Endpoint

About: clgr:Mannheim

Property	Value
rdfs:label	• Mannheim
clgo:country	• Moldova • Germany
rdf:type	• Planned capit • City in Baden • Twin town or • Coat of arms • French exon • Twin town or sister city in Lithuania • University town in Germany • owl:NamedIndividual • City or town in Germany • Most polluted city in the world

List of twin towns and sister cities in Moldova

From Wikipedia, the free encyclopedia

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

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

C [edit]

Chișinău[3][4]

- Alba Iulia, Romania
- Ankara, Turkey
- Bucharest, Romania
- Chernivtsi, Ukraine
- Grenoble, France
- Iași, Romania
- Kiev, Ukraine
- Mannheim, Germany
- Minsk, Belarus
- Odessa, Ukraine
- Reggio Emilia, Italy
- Sacramento, United States
- Tbilisi, Georgia
- Tel Aviv, Israel
- Vilnius, Lithuania
- Yerevan, Armenia

Comrat

- Bălți, Moldova[1]

Criuleni

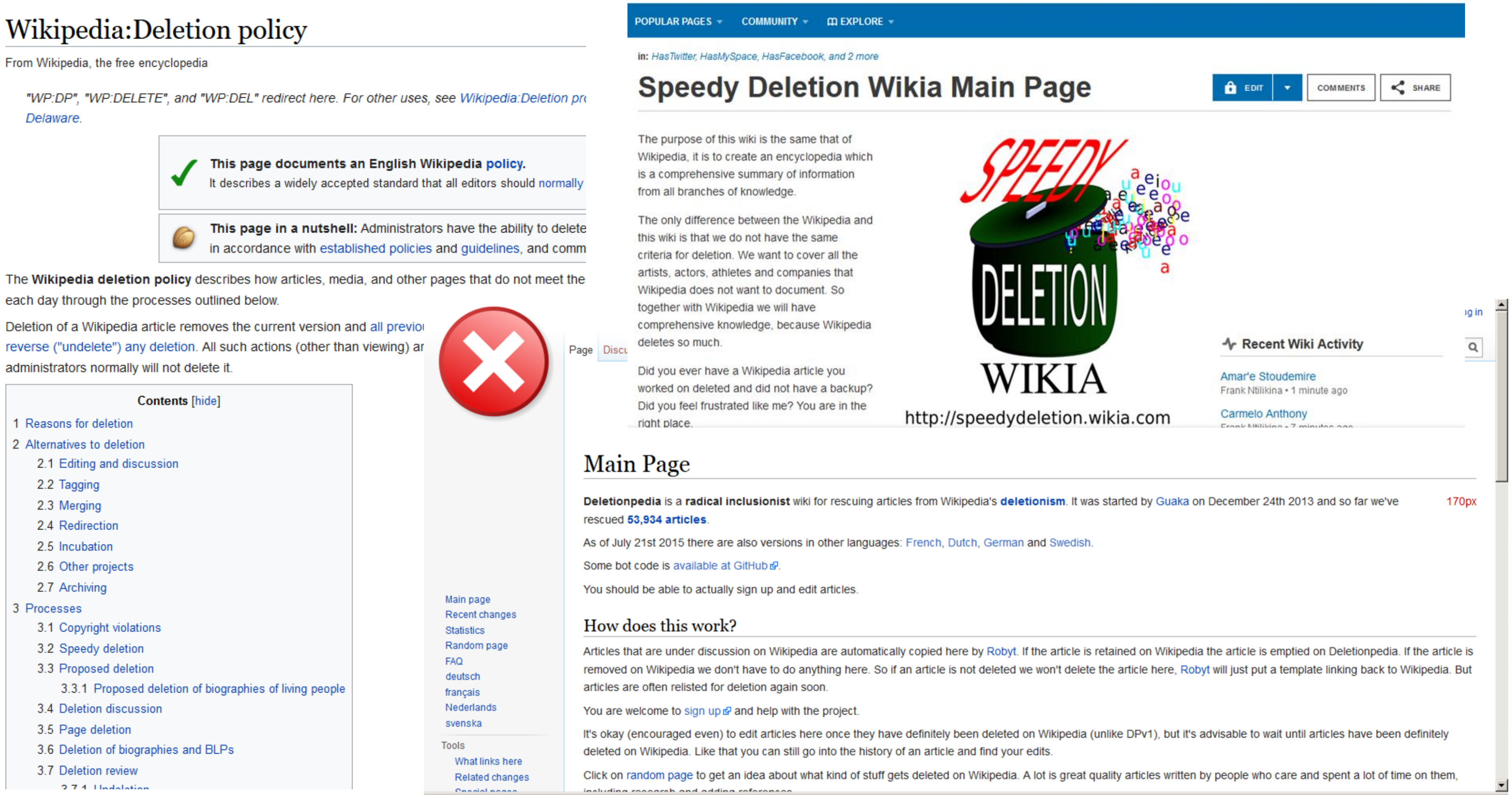
- Jurbarkas, Lithuania[5]

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



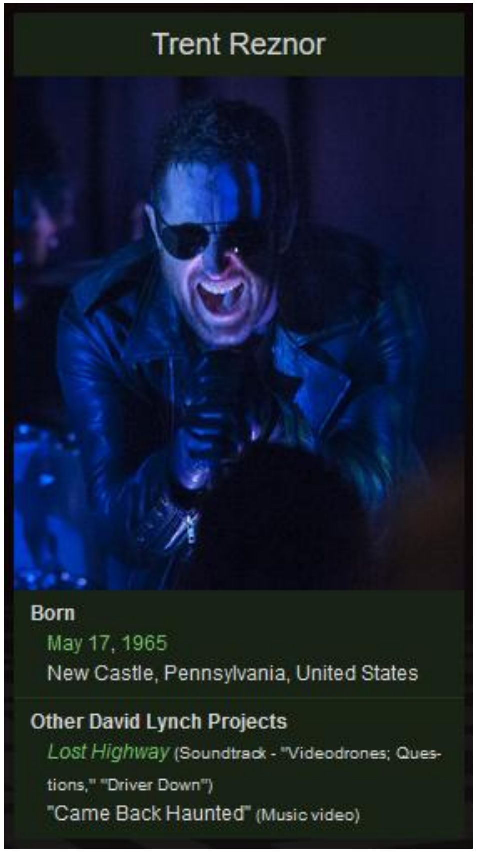
From DBpedia to DBkWik

UNIVERSITY OF MANNHEIM
Data and Web Science Group

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

From DBpedia to DBkWik

- Collecting Data from a Multitude of Wikis



From DBpedia to DBkWik

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

From DBpedia to DBkWik

- Example from Harry Potter Wiki

DBkWik | Browse using | Formats | Faceted Browser | Sparql Endpoint

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 • dbkwik:HarryPotter/resource/Category:Hogwarts_Houses
skos:altLabel	• Gryffindor House • Gryffindor Students • Griffindor • Gryffindors • Griffyndor • Griffyndor Students • Gryffindor Girl • 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

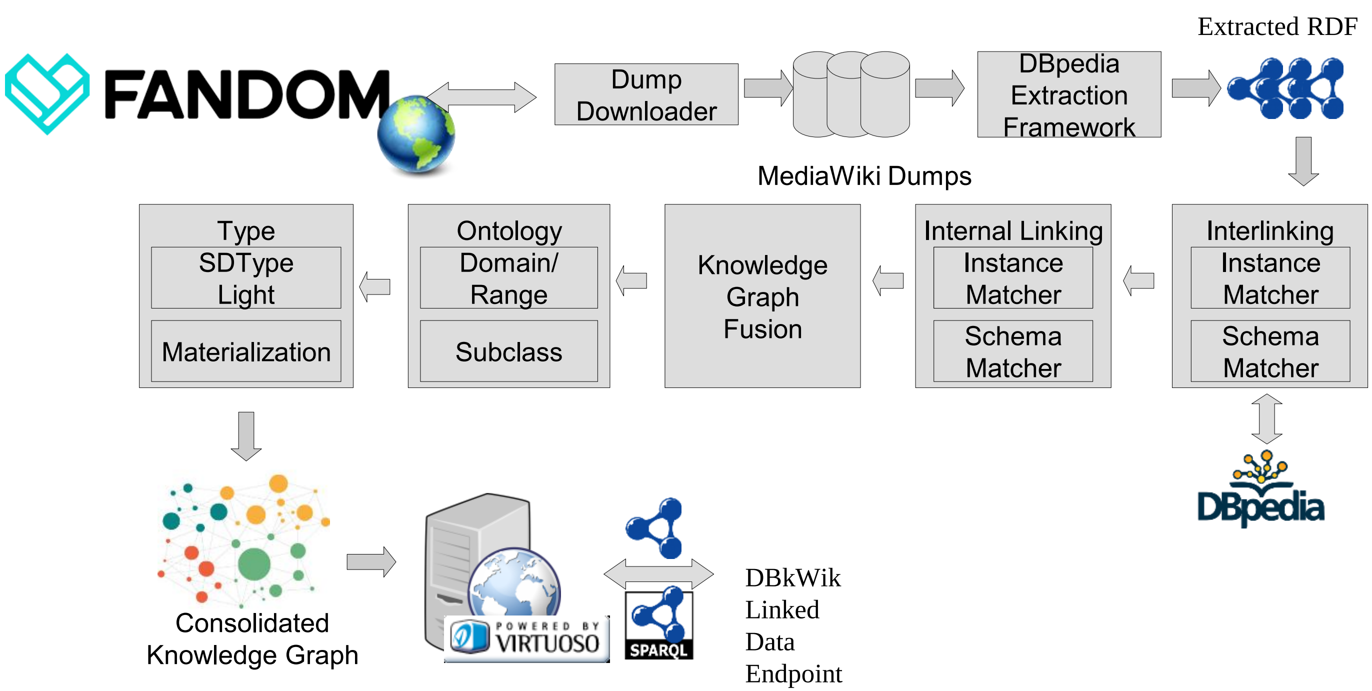
- Differences to DBpedia
  - DBpedia has manually created mappings to an ontology
  - Wikipedia has one page per subject
  - Wikipedia has global infobox conventions (more or less)

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

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

From DBpedia to DBkWik

- Heuristics
  - Ontology induction
  - Instance/Schema Matching



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

From DBpedia to DBkWik

- Downloaded ~15k Wiki dumps from Fandom
  - 52.4GB of data, roughly the size of the English Wikipedia
- Prototype: extracted data for ~250 Wikis
  - 4.3M instances, ~750k linked to DBpedia
  - 7k classes, ~1k linked to DBpedia
  - 43k properties, ~20k linked to DBpedia
  - ...including duplicates!
- Link quality
  - Good for classes, OK for properties (F1 of .957 and .852)
  - Needs improvement for instances (F1 of .641)

From DBpedia to DBkWik

- Scalability of matching:
  - Pairwise matching does not scale
  - 300k Wikis, 1 minute for each pair → 171k years
- Iteratively match and merge
  - 300k Wikis, 1 minute for each match&merge run → 200 days
- Tree-shaped execution plan
  - Parallelizable
  - Hierarchical clustering by topic
  - Whole run under a week

WebIsALOD

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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

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

WebIsALOD

- Example:

About: fiction writer

Premodifier: fiction
Head noun: writer

Same concepts
http://dbpedia.org/resource/Fiction

Broader concepts

label	provenance	confidence
writer	isap:391280092	0.799331
great idea	isap:493244047	0.672180
several magazine	isap:104101164	0.655684
category	isap:104762336	0.477191
artist	isap:387107910	0.471280
blog	isap:492616562	0.458511
writers	isap:439522913	0.427701
story	isap:122402598	0.306667
group	isap:115379219	0.299656
poet	isap:492284397	0.287519

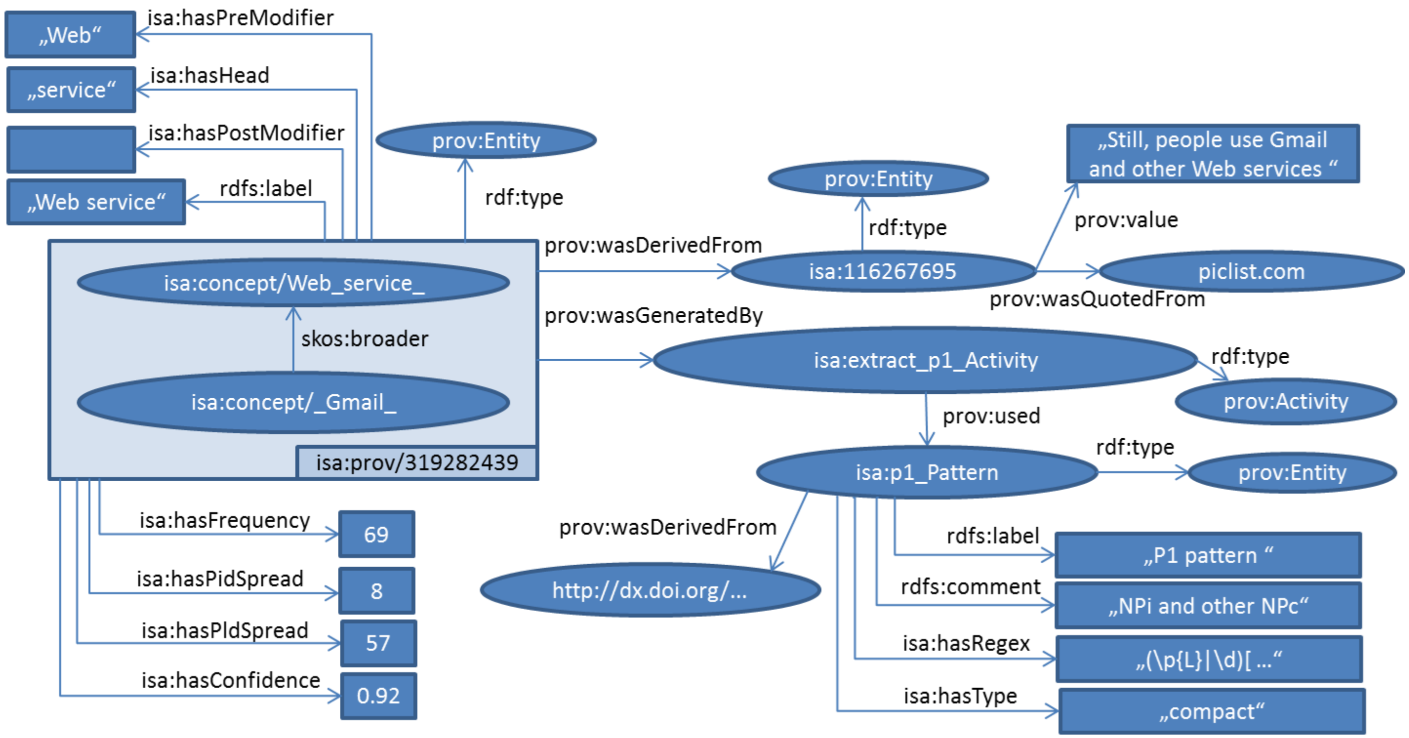
Narrower concepts

label	provenance	confidence
george orwell	isap:386468501	0.662121
science fiction	isap:275868279	0.635886
franz kafka	isap:159147340	0.602015
steve almond	isap:392552636	0.581515
dan brown	isap:157209267	0.574584
james joyce	isap:159394667	0.561794
stephen king	isap:306753456	0.557354
flannery oconnor	isap:266705231	0.555096
alice munro	isap:162537618	0.552608
ayn rand	isap:301402665	0.526857

http://webisa.webdatacommons.org/

WebIsALOD

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

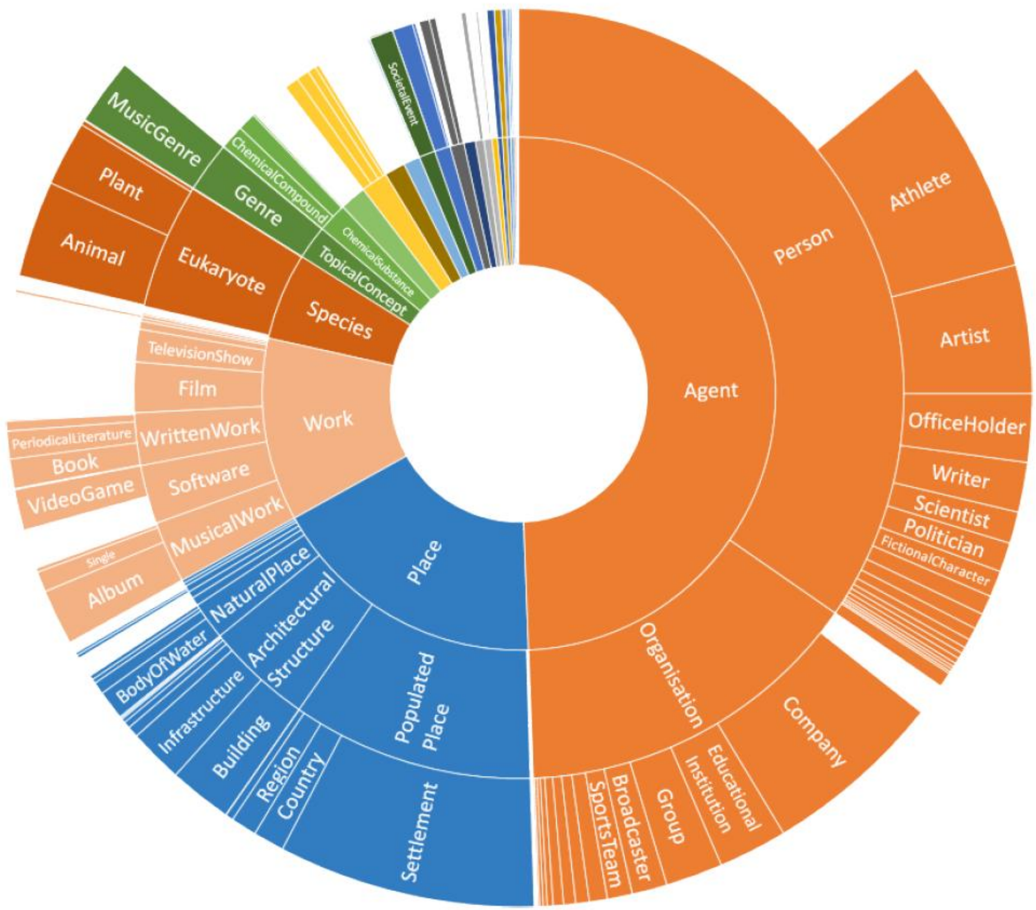


Hertling & Paulheim: WebIsALOD: Providing Hypernymy Relations extracted from the Web as Linked Open Data. ISWC 2017

WebIsALOD

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Estimated contents breakdown



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

WebIsALOD

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Main challenge
  - Original dataset is quite noisy (<10% correct statements)
  - Recap: coverage vs. accuracy
  - Simple thresholding removes too much knowledge
- Approach
  - Train RandomForest model for predicting correct vs. wrong statements
  - Using all the provenance information we have
  - Use model to compute confidence scores
- Current ongoing research
  - Using transformers and a larger training set

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

WebIsALOD

University of Mannheim
Data and Web Science Group

- 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?