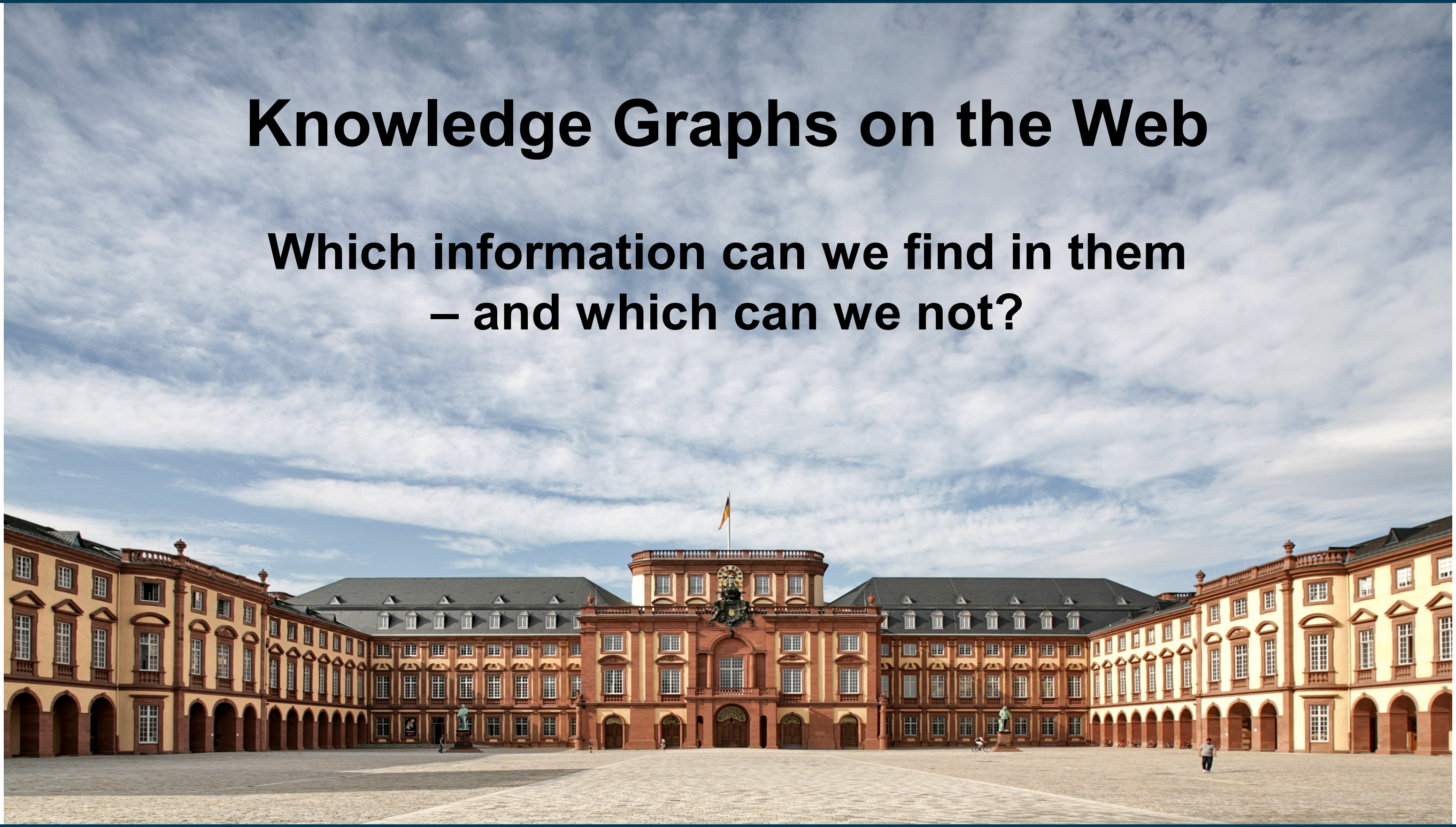
UNIVERSITÄT MANNHEIM



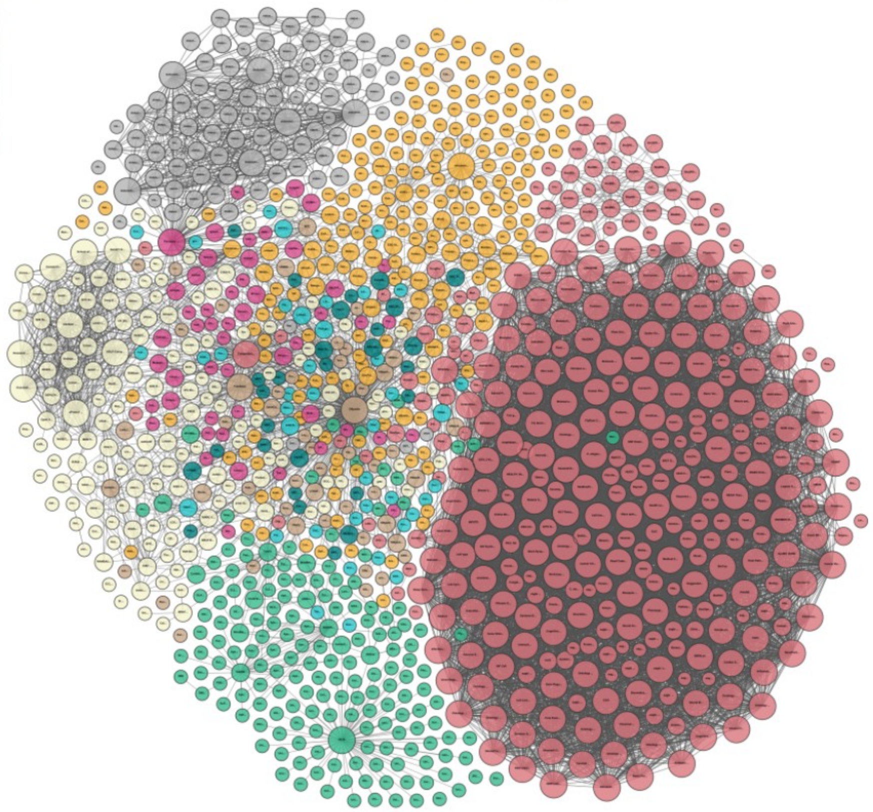
Knowledge Graphs on the Web

Which information can we find in them – and which can we not?

Heiko Paulheim

Introduction

- You’ve seen this, haven’t you?



Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

Introduction

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

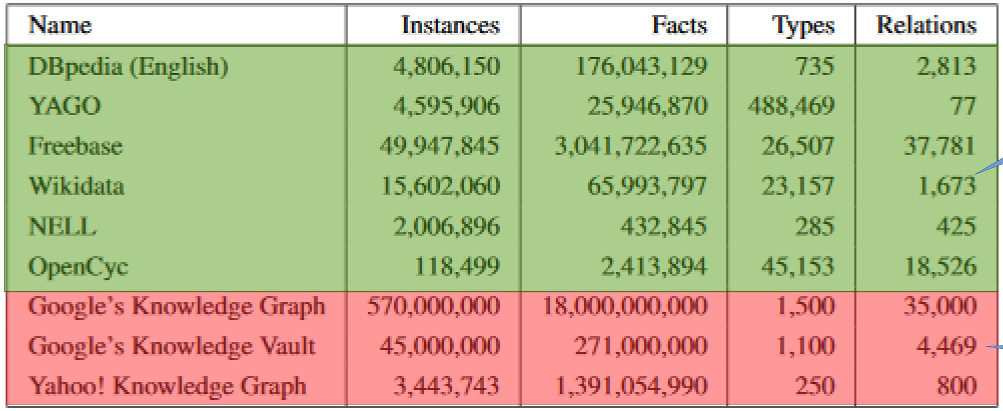
Introduction

- Knowledge Graph definitions
- Many people talk about KGs, few give definitions

- 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

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

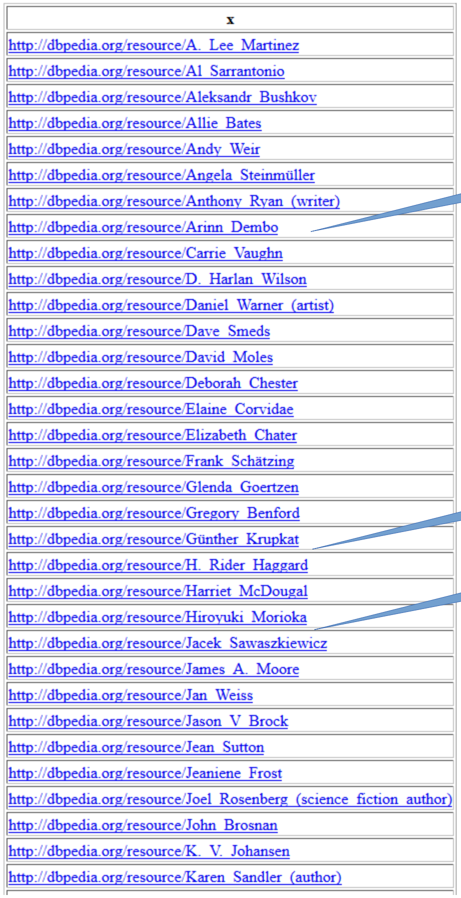
Finding Information in Knowledge Graphs

- Find list of science fiction writers in DBpedia

```
select ?x where
       {?x a dbo:Writer .
        ?x dbo:genre dbr:Science_Fiction}
order by ?x
```

Finding Information in Knowledge Graphs

- Results from DBpedia



x
http://dbpedia.org/resource/A._Lee_Martinez
http://dbpedia.org/resource/Al_Sarrantonio
http://dbpedia.org/resource/Aleksandr_Bushkov
http://dbpedia.org/resource/Allie_Bates
http://dbpedia.org/resource/Andy_Weir
http://dbpedia.org/resource/Angela_Steinmüller
http://dbpedia.org/resource/Anthony_Ryan_(writer)
http://dbpedia.org/resource/Arinn_Dembo
http://dbpedia.org/resource/Carrie_Vaughn
http://dbpedia.org/resource/D._Harlan_Wilson
http://dbpedia.org/resource/Daniel_Warner_(artist)
http://dbpedia.org/resource/Dave_Smeds
http://dbpedia.org/resource/David_Moles
http://dbpedia.org/resource/Deborah_Chester
http://dbpedia.org/resource/Elaine_Corvidae
http://dbpedia.org/resource/Elizabeth_Chater
http://dbpedia.org/resource/Frank_Schätzing
http://dbpedia.org/resource/Glenda_Goertzen
http://dbpedia.org/resource/Gregory_Benford
http://dbpedia.org/resource/Günther_Krupkat
http://dbpedia.org/resource/H._Rider_Haggard
http://dbpedia.org/resource/Harriet_McDougal
http://dbpedia.org/resource/Hiroyuki_Morioka
http://dbpedia.org/resource/Jacek_Sawaszkiewicz
http://dbpedia.org/resource/James_A._Moore
http://dbpedia.org/resource/Jan_Weiss
http://dbpedia.org/resource/Jason_V_Brock
http://dbpedia.org/resource/Jean_Sutton
http://dbpedia.org/resource/Jeaniene_Frost
http://dbpedia.org/resource/Joel_Rosenberg_(science_fiction_author)
http://dbpedia.org/resource/John_Brosnan
http://dbpedia.org/resource/K._V._Johansen
http://dbpedia.org/resource/Karen_Sandler_(author)

Arthur C. Clarke?

H.G. Wells?

Isaac Asimov?

Finding Information in Knowledge Graphs

- Questions in this talk
  - What can we find in different Knowledge Graphs?
  - Why do we sometimes not find what we expect to find?
  - What can be done about this?

- ...and:
  - What new Knowledge Graphs are currently developed?

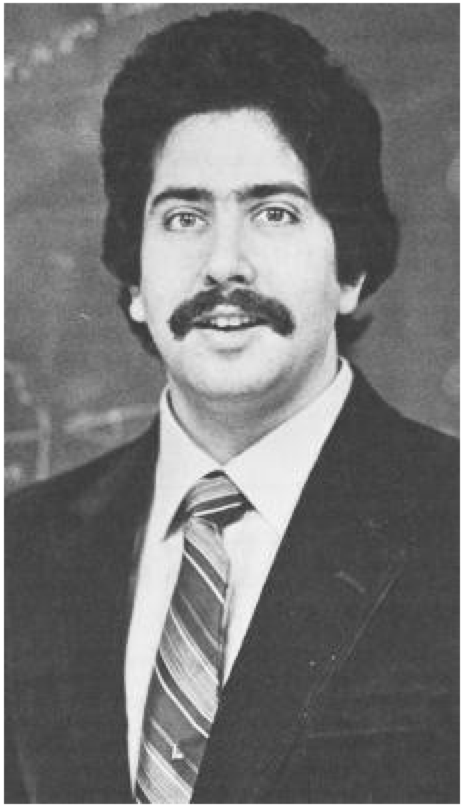
Outline

- How are Knowledge Graphs created?
- What is inside public Knowledge Graphs?
  - Knowledge Graph profiling
- Addressing typical problems
  - Errors
  - Incompleteness
- New Kids on the Block
  - WebIsALOD
  - DBkWik
- Take Aways

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
  - >900 person years
  - Far from completion
  - Used to exist until 2017



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)

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

- 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

yago
select knowledge
DBpedia

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

- Today
  - Still running
  - New release every few days

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

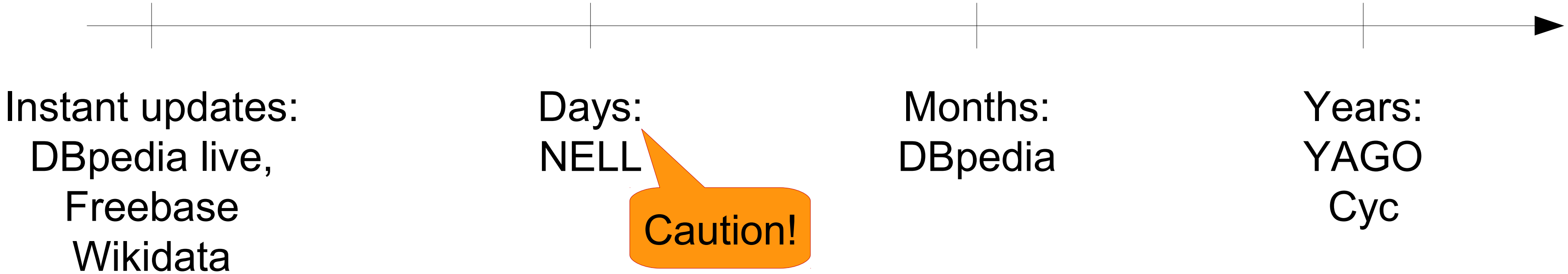
Non-Public Knowledge Graphs

- 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

Comparison of Knowledge Graphs

- Release cycles



Instant updates:
DBpedia live,
Freebase
Wikidata

Days:
NELL

Caution!

Months:
DBpedia

Years:
YAGO
Cyc

- Size and density

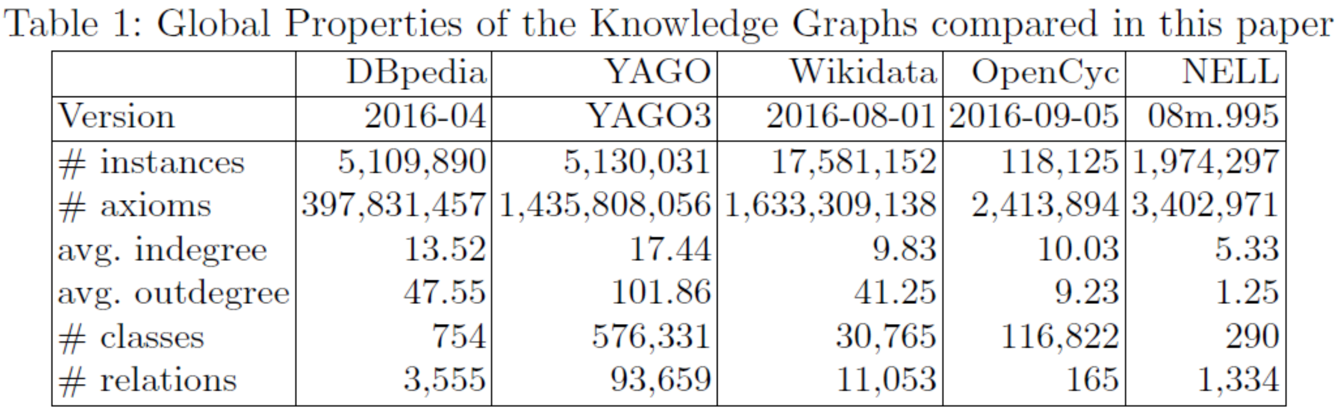


Table 1: Global Properties of the Knowledge Graphs compared in this paper

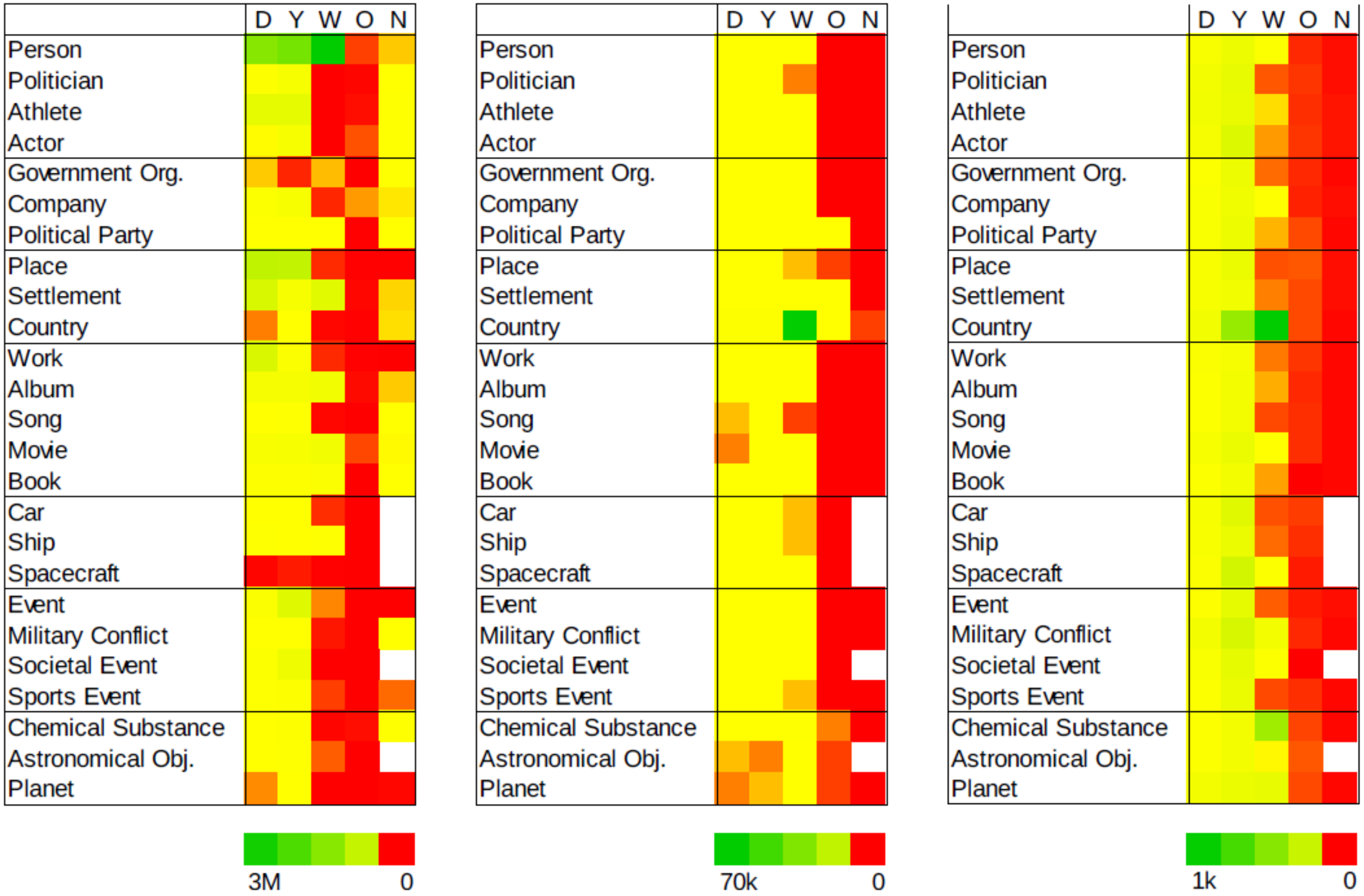
	DBpedia	YAGO	Wikidata	OpenCyc	NELL
Version	2016-04	YAGO3	2016-08-01	2016-09-05	08m.995
# instances	5,109,890	5,130,031	17,581,152	118,125	1,974,297
# axioms	397,831,457	1,435,808,056	1,633,309,138	2,413,894	3,402,971
avg. indegree	13.52	17.44	9.83	10.03	5.33
avg. outdegree	47.55	101.86	41.25	9.23	1.25
# 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

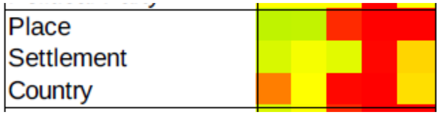
Comparison of Knowledge Graphs

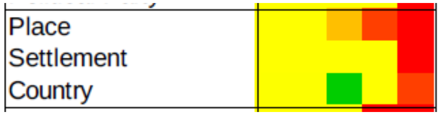
- Summary findings:
  - Persons: more in Wikidata
    (twice as many persons as DBpedia and YAGO)
  - Countries: more details in Wikidata
  - Places: most in DBpedia
  - Organizations: most in YAGO
  - Events: most in YAGO
  - Artistic works:
    - Wikidata contains more movies and albums
    - YAGO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

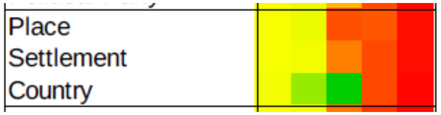
Ringler & Paulheim: *One Knowledge Graph to Rule them All?* KI 2017

Caveats

- Reading the diagrams right…

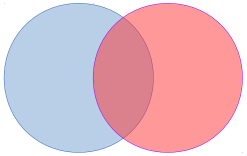




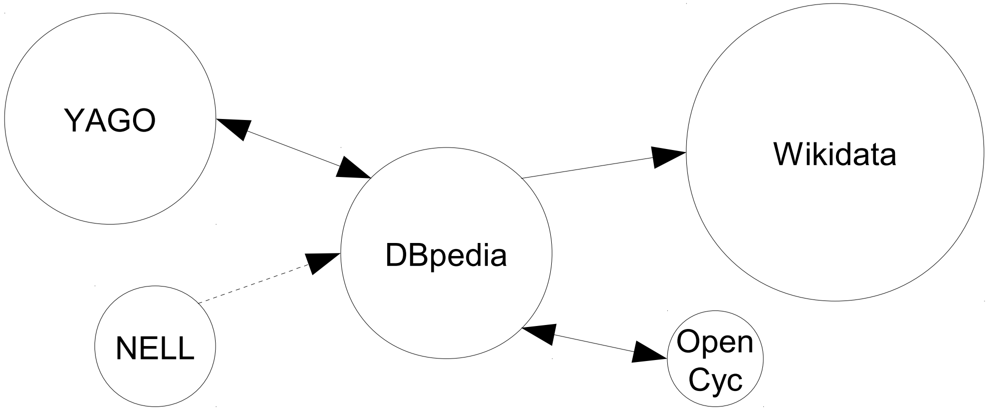


- So, Wikidata contains more data on countries, but less countries?
- First: Wikidata only counts current, actual countries
  – DBpedia and YAGO also count historical countries
- “KG1 contains less of X than KG2” can mean
  – it actually contains less instances of X
  – it contains equally many or more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

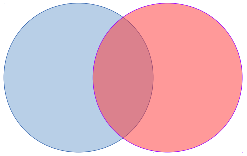


- How largely do knowledge graphs overlap?
- They are interlinked, so we can simply count links
  - For NELL, we use links to Wikipedia as a proxy



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

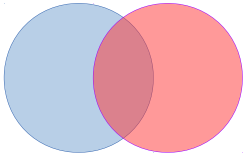
Overlap of Knowledge Graphs



- 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

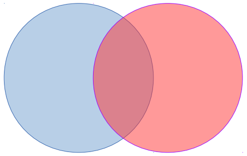


- Links between Knowledge Graphs are incomplete
  - The Open World Assumption also holds for interlinks

- 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



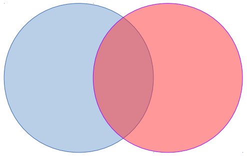
- Idea:
  - Given that the link set F is found
  - And the (unknown) actual link set would be C

- 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



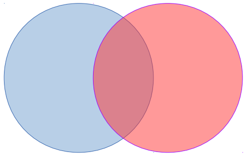
- Mathematical derivation:
  - Definition of recall: $R = \frac{|F_{correct}|}{|C|}$
  - Definition of precision: $P = \frac{|F_{correct}|}{|F|}$

- 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

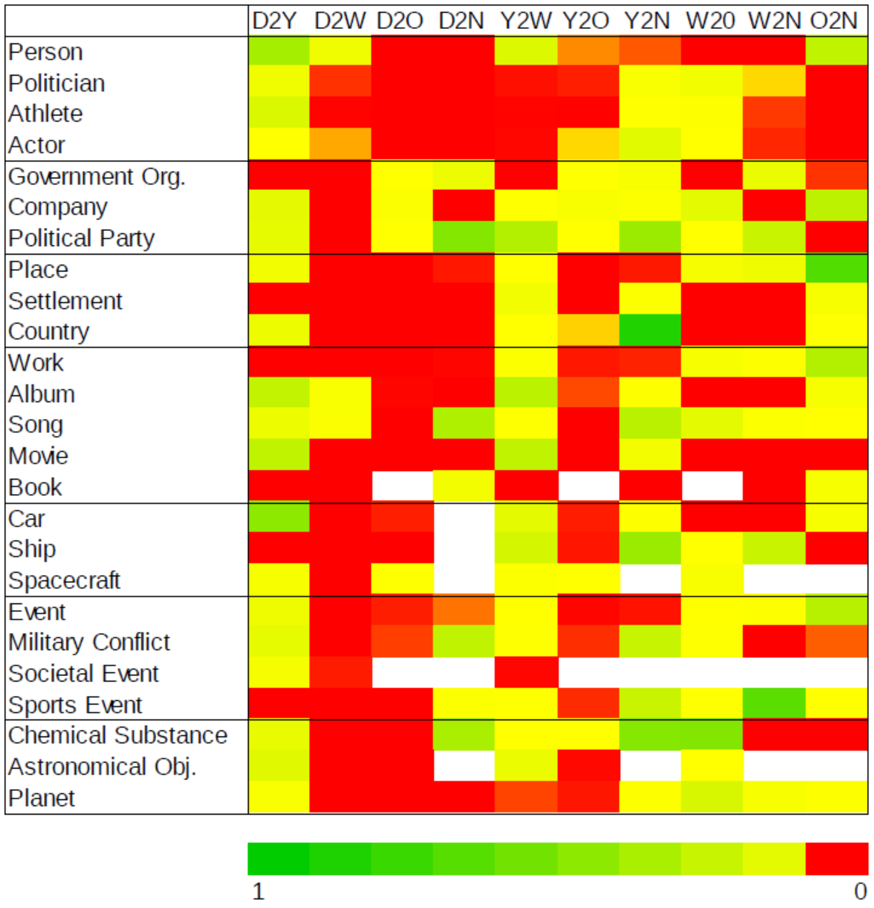


- Experiment:
  - We use the same 25 classes as before
  - Measure 1: overlap relative to smaller KG (i.e., potential gain)
  - Measure 2: overlap relative to explicit links
    (i.e., importance of improving links)

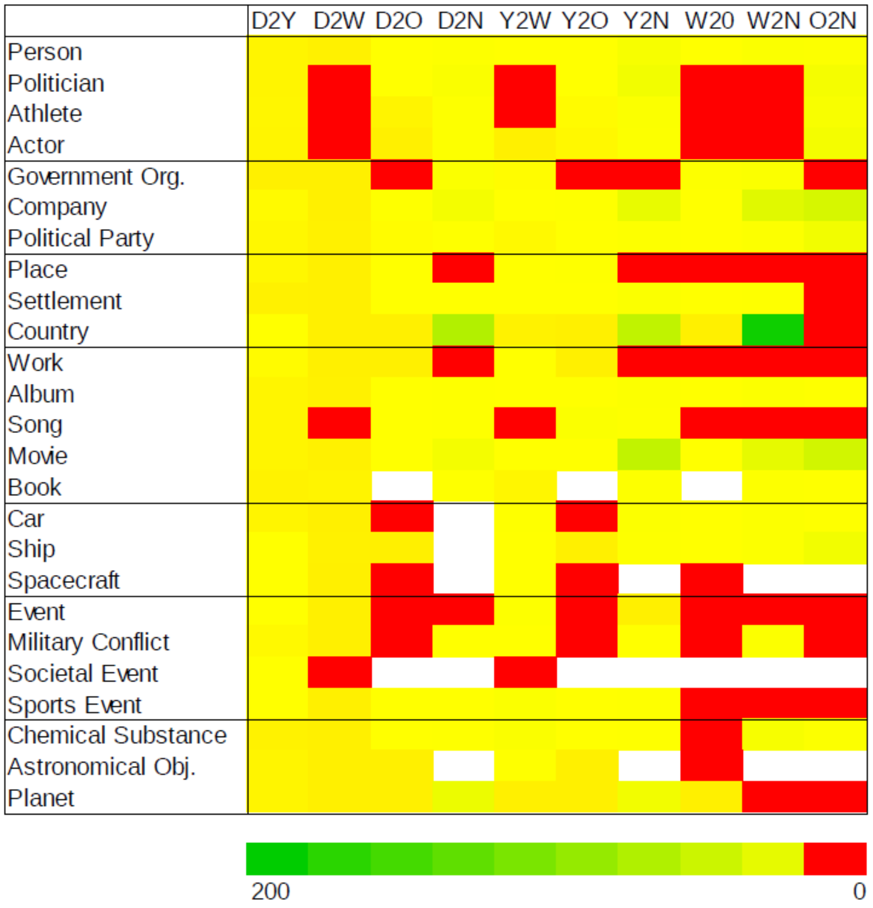
- Link generation with 16 different metrics and thresholds
  - Intra-class correlation coefficient for |C|: 0.969
  - Intra-class correlation coefficient for |F|: 0.646
- Bottom line:
  - Despite variety in link sets generated, the overlap is estimated reliably
  - The link generation mechanisms do not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

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

Overlap of Knowledge Graphs



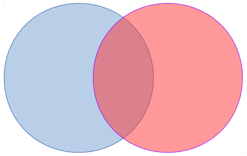
(a) Overlap as potential gain



(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

Common Errors in Knowledge Graphs

- Using DBpedia as an Example
  - ...but most of those hold for other KGs as well
  - ...each KG has its own advantages and shortcomings
- Recap: using mappings & heuristics for extraction from Wikipedia
- Something to keep in mind:
  - Wikipedia is made for humans
  - Not necessarily: for facilitating easy Knowledge Graph creation

Common Errors in Knowledge Graphs

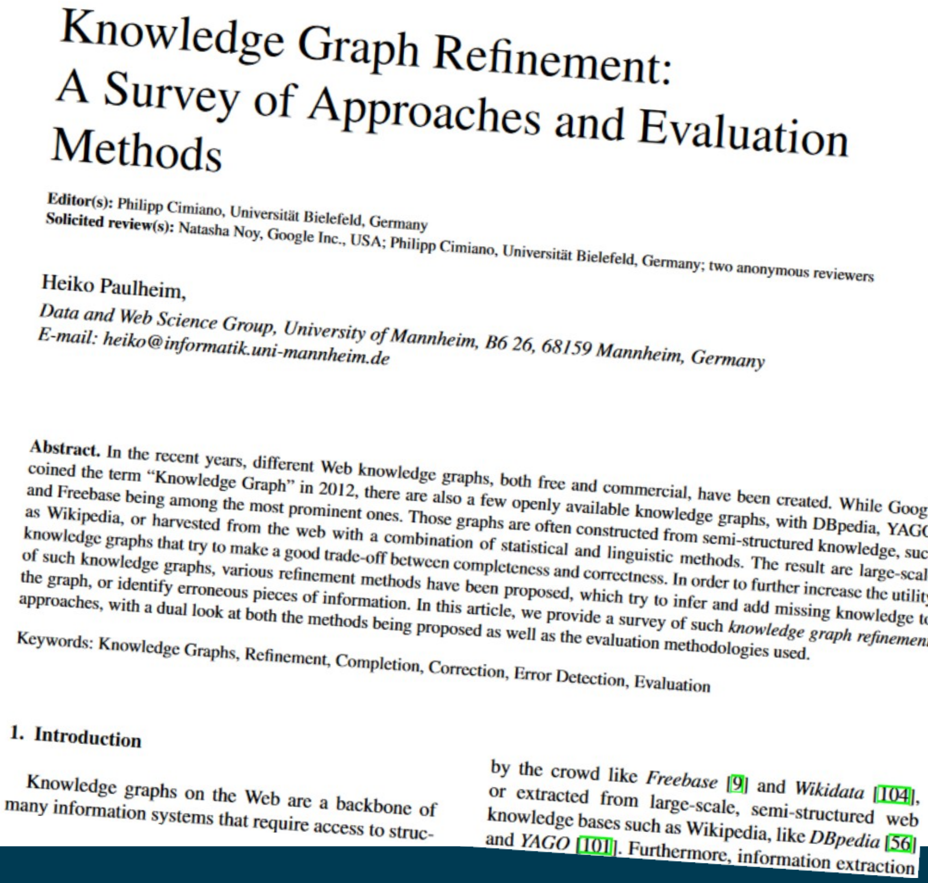
- What reasons can cause incomplete results?

```
select ?x where
       {?x a dbo:Writer .
        ?x dbo:genre dbr:Science_Fiction}
order by ?x
```

- Two possible problems:
  - The resource at hand is not of type `dbo:Writer`
  - The genre relation to `dbr:Science_Fiction` is missing

Common Errors in Knowledge Graphs

- Various works on *Knowledge Graph Refinement*
  - Knowledge Graph completion
  - Error detection
- See, e.g., 2017 survey in Semantic Web Journal



Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods

Editor(s): Philipp Cimiano, Universität Bielefeld, Germany
Solicited review(s): Natasha Noy, Google Inc., USA; Philipp Cimiano, Universität Bielefeld, Germany; two anonymous reviewers

Heiko Paulheim,
Data and Web Science Group, University of Mannheim, B6 26, 68159 Mannheim, Germany
E-mail: heiko@informatik.uni-mannheim.de

Abstract. In the recent years, different Web knowledge graphs, both free and commercial, have been created. While Google coined the term "Knowledge Graph" in 2012, there are also a few openly available knowledge graphs, with DBpedia, YAGO, and Freebase being among the most prominent ones. Those graphs are often constructed from semi-structured knowledge, such as Wikipedia, or harvested from the web with a combination of statistical and linguistic methods. The result are large-scale knowledge graphs that try to make a good trade-off between completeness and correctness. In order to further increase the utility of such knowledge graphs, various refinement methods have been proposed, which try to infer and add missing knowledge to the graph, or identify erroneous pieces of information. In this article, we provide a survey of such *knowledge graph refinement* approaches, with a dual look at both the methods being proposed as well as the evaluation methodologies used.

Keywords: Knowledge Graphs, Refinement, Completion, Correction, Error Detection, Evaluation

1. Introduction

Knowledge graphs on the Web are a backbone of many information systems that require access to struc-

by the crowd like *Freebase* [9] and *Wikidata* [104], or extracted from large-scale, semi-structured web knowledge bases such as Wikipedia, like *DBpedia* [56] and *YAGO* [101]. Furthermore, information extraction

Paulheim: Knowledge Graph Refinement – A Survey of Approaches and Evaluation Methods. SWJ 8(3), 2017

Common Errors in Knowledge Graphs

- Missing types
  - Estimate (2013) for DBpedia: at least 2.6M type statements are missing
  - Using YAGO as “ground truth”

- “Well, we’re semantics folks, we have ontologies!”
  -
```
CONSTRUCT {?x a ?t}
WHERE { {?x ?r ?y . ?r rdfs:domain ?t}
UNION   {?y ?r ?x . ?r rdfs:range  ?t} }
```

Bizer & Paulheim: Type Inference on Noisy RDF Data. In: ISWC 2013

Common Errors in Knowledge Graphs

- Experiment of RDFS reasoning for typing `Germany`
- Results:
  - `Place, PopulatedPlace, Award, MilitaryConflict, City, Country, EthnicGroup, Genre, Stadium, Settlement, Language, MontainRange, PersonFunction, Race, RouteOfTransportation, Building, Mountain, Airport, WineRegion`
- Bottom line: RDFS reasoning *accumulates* errors
  - Germany is the object of 44,433 statements
  - 15 single wrong statements can cause those 15 errors
  - i.e., an error rate of only 0.03% (that is unlikely to achieve)

Bizer & Paulheim: Type Inference on Noisy RDF Data. In: ISWC 2013

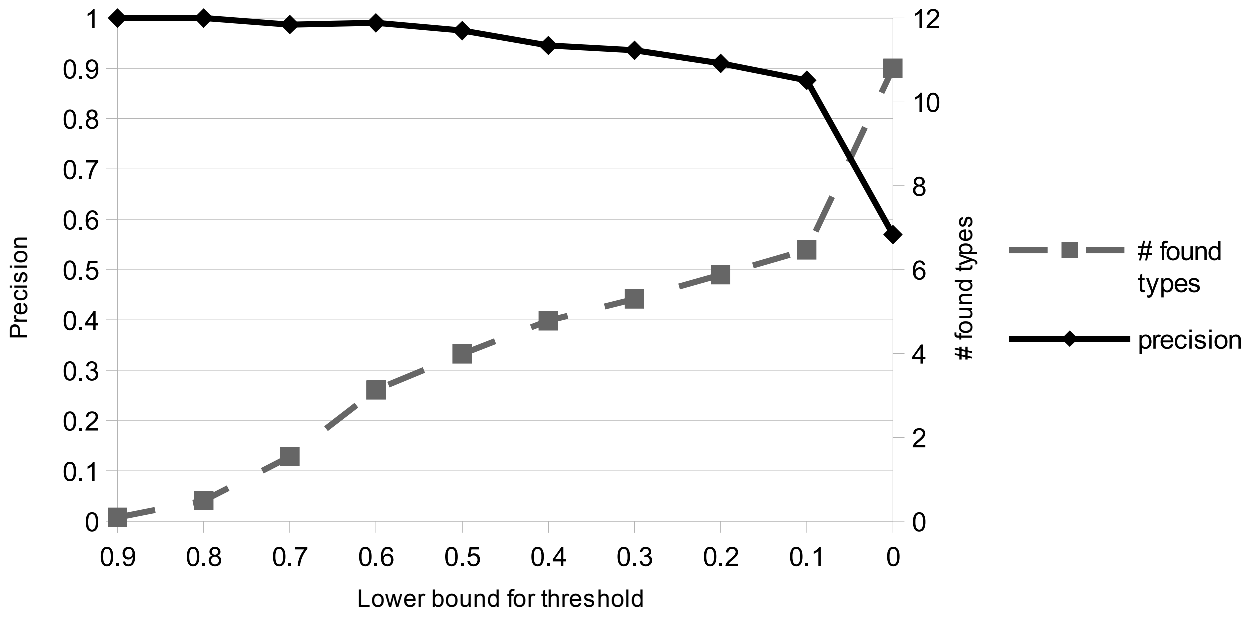
Common Errors in Knowledge Graphs

- Required: a noise-tolerant approach
- SDType (meanwhile included in DBpedia)
  - Use statistical distributions of properties and object types
    - P(C|p) → probability of object being of type C when observing property p in a statement
  - Averaging scores for all statements of a resource
  - Weighting properties by discriminative power
- Since DBpedia 3.9: typing ~1M untyped resources at precision >0.95
- Refinement:
  - Filtering resources of non-instance pages and list pages

Bizer & Paulheim: Type Inference on Noisy RDF Data. In: ISWC 2013

Common Errors in Knowledge Graphs

- Recap
  - Trade-off coverage vs. accuracy



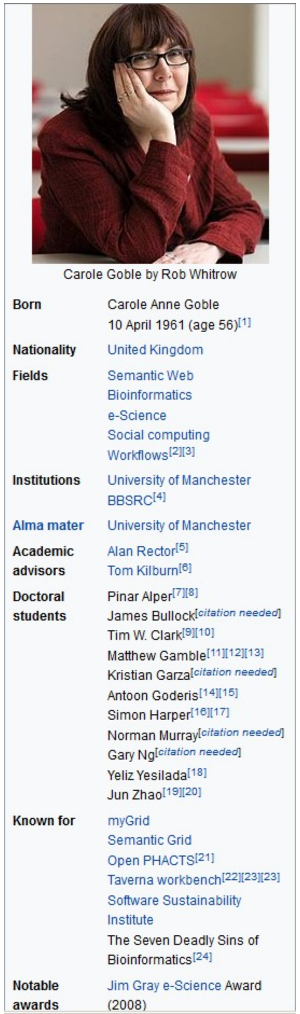
Common Errors in Knowledge Graphs

- The same idea applied to identification of noisy statements
  - i.e., a statement is implausible if the distribution of its object's types deviates from the overall distribution for the predicate
- Removing ~20,000 erroneous statements from DBpedia
- Error analysis
  - Errors in Wikipedia account for ~30%
  - Other typical problems: see following slides

Bizer & Paulheim: Improving the quality of linked data using statistical distributions. In: IJSWIS 10(2), 2014

Common Errors in Knowledge Graphs

- Typical errors
  - links in longer texts are not interpreted correctly
  - dbr:Carole_Goble dbo:award dbr:Jim_Gray



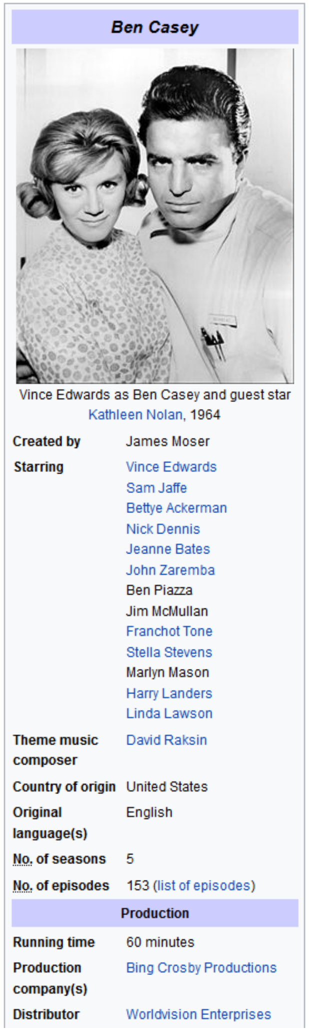
Carole Goble by Rob Whitrow

Born	Carole Anne Goble 10 April 1961 (age 56)[1]
Nationality	United Kingdom
Fields	Semantic Web Bioinformatics e-Science Social computing Workflows[2][3]
Institutions	University of Manchester BBSRC[4]
Alma mater	University of Manchester
Academic advisors	Alan Rector[5] Tom Kilburn[6]
Doctoral students	Pinar Alper[7][8] James Bullock[citation needed] Tim W. Clark[9][10] Matthew Gamble[11][12][13] Kristian Garza[citation needed] Antoon Goderis[14][15] Simon Harper[16][17] Norman Murray[citation needed] Gary Ng[citation needed] Yeliz Yesilada[18] Jun Zhao[19][20]
Known for	myGrid Semantic Grid Open PHACTS[21] Taverna workbench[22][23][23] Software Sustainability Institute The Seven Deadly Sins of Bioinformatics[24]
Notable awards	Jim Gray e-Science Award (2008)

Paulheim & Gangemi: *Serving DBpedia with DOLCE – More than Just Adding a Cherry on Top.* ISWC 2015

Common Errors in Knowledge Graphs

- Typical errors
  - Misinterpretation of redirects
  - dbr:Ben_Casey dbo:company dbr:Bing_Cosby



Ben Casey

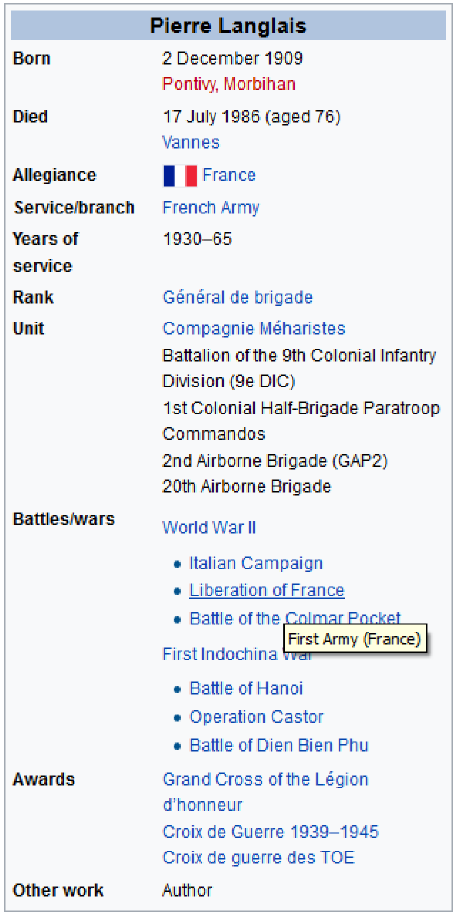
Vince Edwards as Ben Casey and guest star Kathleen Nolan, 1964

Created by	James Moser
Starring	Vince Edwards Sam Jaffe Bettye Ackerman Nick Dennis Jeanne Bates John Zaremba Ben Piazza Jim McMullan Franchot Tone Stella Stevens Marlyn Mason Harry Landers Linda Lawson
Theme music composer	David Raksin
Country of origin	United States
Original language(s)	English
No. of seasons	5
No. of episodes	153 (list of episodes)
Production	
Running time	60 minutes
Production company(s)	Bing Crosby Productions
Distributor	Worldvision Enterprises

Paulheim & Gangemi: *Serving DBpedia with DOLCE – More than Just Adding a Cherry on Top.* ISWC 2015

Common Errors in Knowledge Graphs

- Typical errors
  - Metonymy
  - dbr:Human_Nature_(band) dbo:genre dbr:Motown,
  - Links with anchors pointing to subsections in a page
  - First_Army_(France)#1944-1945



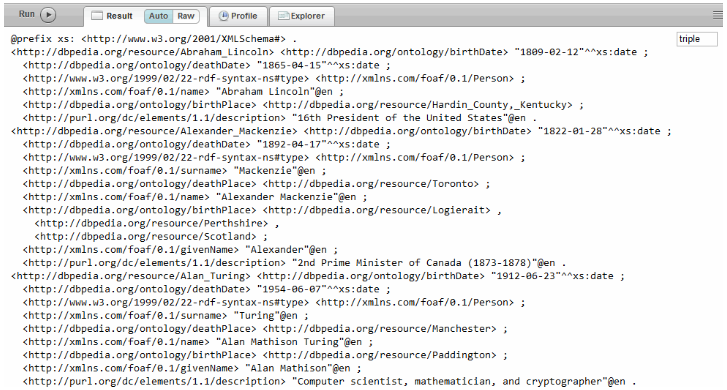
Pierre Langlais	
Born	2 December 1909 Pontivy, Morbihan
Died	17 July 1986 (aged 76) Vannes
Allegiance	France
Service/branch	French Army
Years of service	1930–65
Rank	Général de brigade
Unit	Compagnie Méharistes Battalion of the 9th Colonial Infantry Division (9e DIC) 1st Colonial Half-Brigade Paratroop Commandos 2nd Airborne Brigade (GAP2) 20th Airborne Brigade
Battles/wars	World War II • Italian Campaign • Liberation of France • Battle of the Colmar Pocket First Indochina War • Battle of Hanoi • Operation Castor • Battle of Dien Bien Phu
Awards	Grand Cross of the Légion d'honneur Croix de Guerre 1939–1945 Croix de guerre des TOE
Other work	Author

First Army (France)

Paulheim & Gangemi: *Serving DBpedia with DOLCE – More than Just Adding a Cherry on Top.* ISWC 2015

Common Errors in Knowledge Graphs

- Identifying individual errors is possible with many techniques
  - e.g., statistics, reasoning, exploiting upper ontologies, …
- ...but what do we do with those efforts?
  - they typically end up in drawers and abandoned GitHub repositories

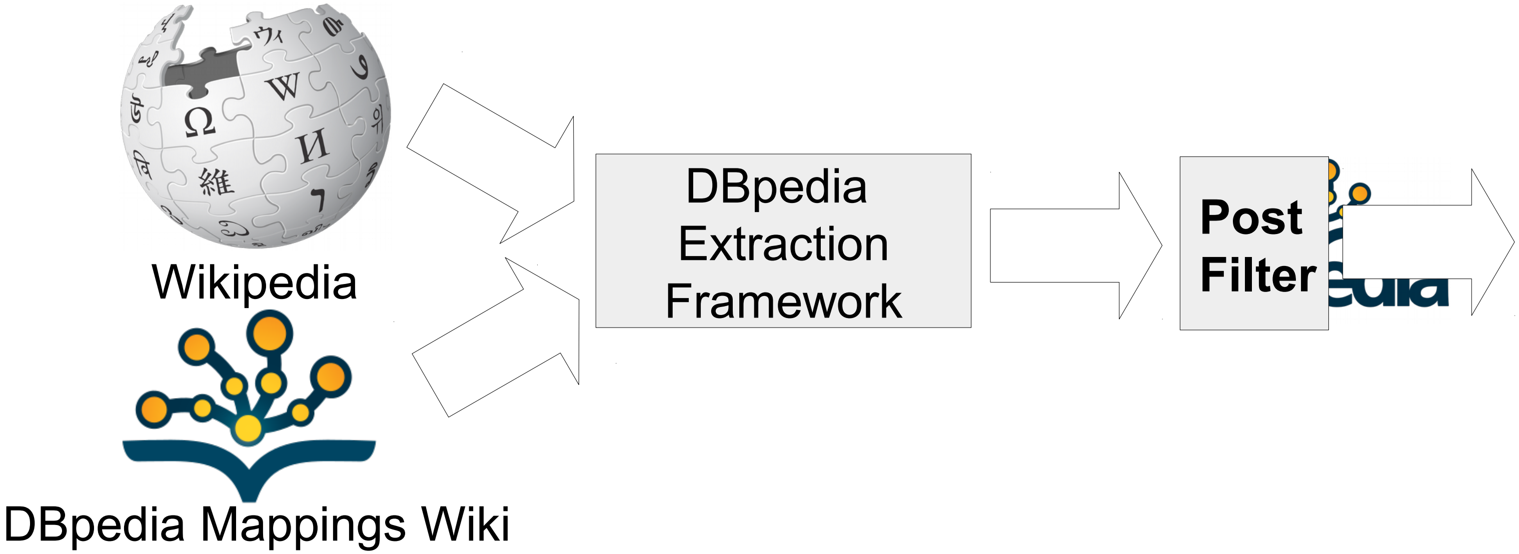


```
@prefix xs: <http://www.w3.org/2001/XMLSchema#> .
<http://dbpedia.org/resource/Abraham_Lincoln> <http://dbpedia.org/ontology/birthDate> "1809-02-12"^^xs:date ;
  <http://dbpedia.org/ontology/deathDate> "1865-04-15"^^xs:date ;
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> ;
  <http://xmlns.com/foaf/0.1/name> "Abraham Lincoln"@en ;
  <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Hardin_County,_Kentucky> ;
  <http://purl.org/dc/elements/1.1/description> "16th President of the United States"@en .
<http://dbpedia.org/resource/Alexander_Mackenzie> <http://dbpedia.org/ontology/birthDate> "1822-01-28"^^xs:date ;
  <http://dbpedia.org/ontology/deathDate> "1892-04-17"^^xs:date ;
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> ;
  <http://xmlns.com/foaf/0.1/surname> "Mackenzie"@en ;
  <http://dbpedia.org/ontology/deathPlace> <http://dbpedia.org/resource/Toronto> ;
  <http://xmlns.com/foaf/0.1/name> "Alexander Mackenzie"@en ;
  <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Logierait> ,
    <http://dbpedia.org/resource/Perthshire> ,
    <http://dbpedia.org/resource/Scotland> ;
  <http://xmlns.com/foaf/0.1/givenName> "Alexander"@en ;
  <http://purl.org/dc/elements/1.1/description> "2nd Prime Minister of Canada (1873-1878)"@en .
<http://dbpedia.org/resource/Alan_Turing> <http://dbpedia.org/ontology/birthDate> "1912-06-23"^^xs:date ;
  <http://dbpedia.org/ontology/deathDate> "1954-06-07"^^xs:date ;
  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> ;
  <http://xmlns.com/foaf/0.1/surname> "Turing"@en ;
  <http://dbpedia.org/ontology/deathPlace> <http://dbpedia.org/resource/Manchester> ;
  <http://xmlns.com/foaf/0.1/name> "Alan Mathison Turing"@en ;
  <http://dbpedia.org/ontology/birthPlace> <http://dbpedia.org/resource/Paddington> ;
  <http://xmlns.com/foaf/0.1/givenName> "Alan Mathison"@en ;
  <http://purl.org/dc/elements/1.1/description> "Computer scientist, mathematician, and cryptographer"@en .
```

Paulheim & Gangemi: *Serving DBpedia with DOLCE – More than Just Adding a Cherry on Top.* ISWC 2015
Paulheim: *Data-driven Joint Debugging of the DBpedia Mappings and Ontology.* ESWC 2017

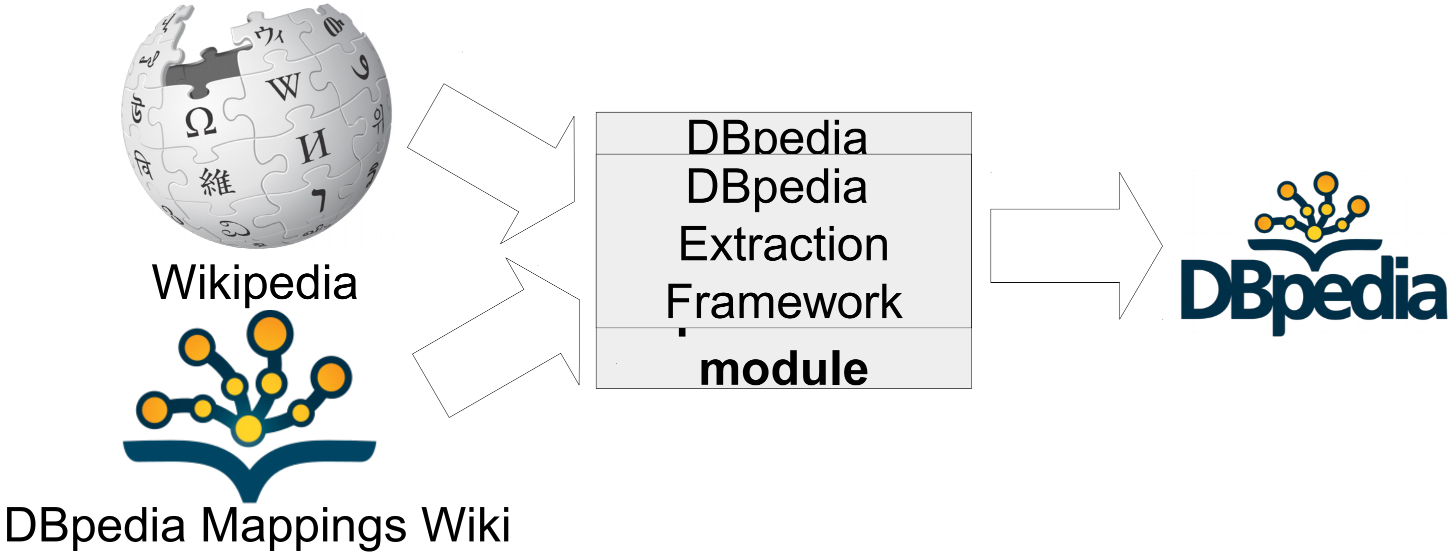
Motivation

- Possible option 1: Remove erroneous triples from DBpedia
- Challenges
  - May remove correct axioms, may need thresholding
  - Needs to be repeated for each release
  - **Needs to be materialized on all of DBpedia**



Motivation

- Possible option 2: Integrate into DBpedia Extraction Framework
- Challenges
  - Development workload
  - Some approaches are not fully automated (technically or conceptually)
  - **Scalability**



Wikipedia

DBpedia Mappings Wiki

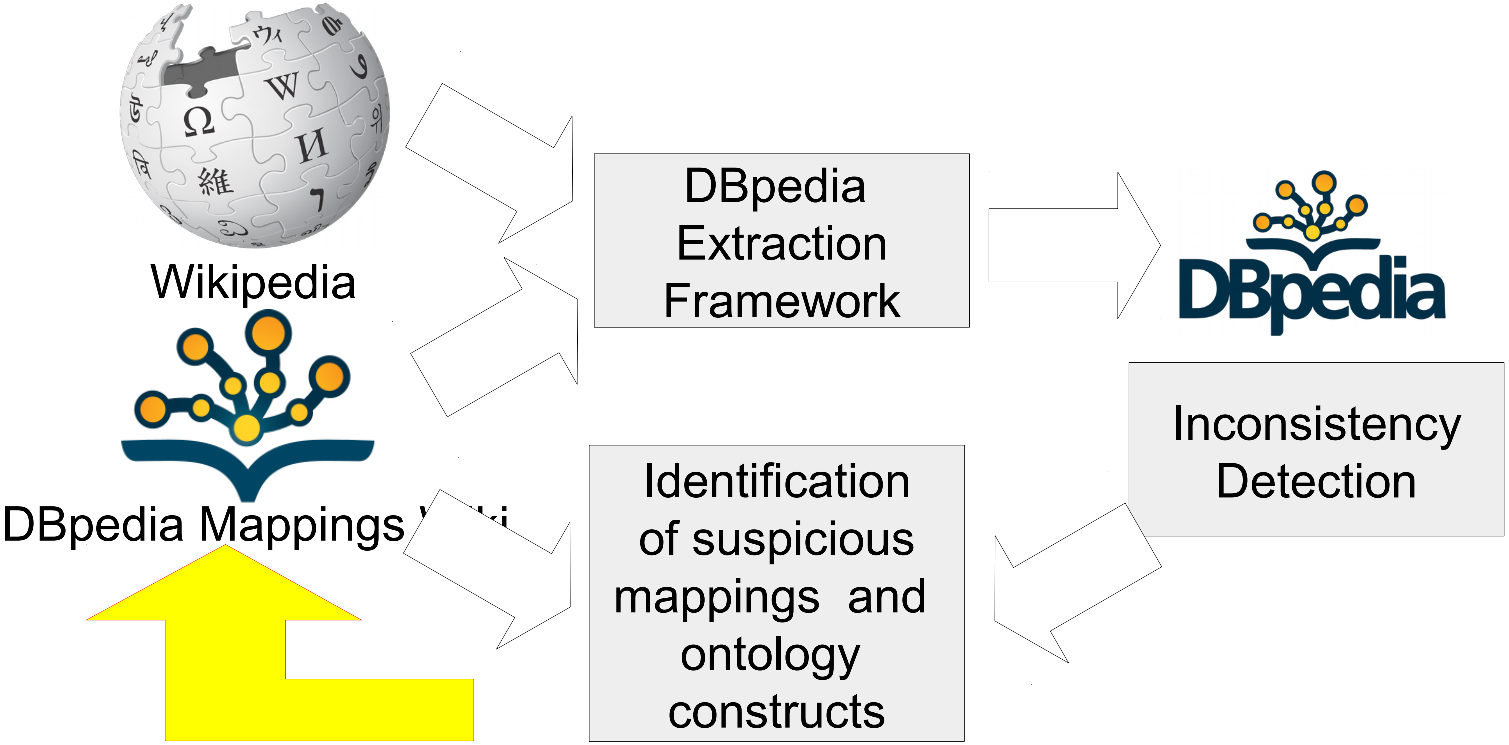
DBpedia

DBpedia Extraction Framework

module

Common Errors in Knowledge Graphs

- Goal: a third option
  - Find the root of the error and fix it!



Wikipedia

DBpedia Mappings Wiki

DBpedia Extraction Framework

DBpedia

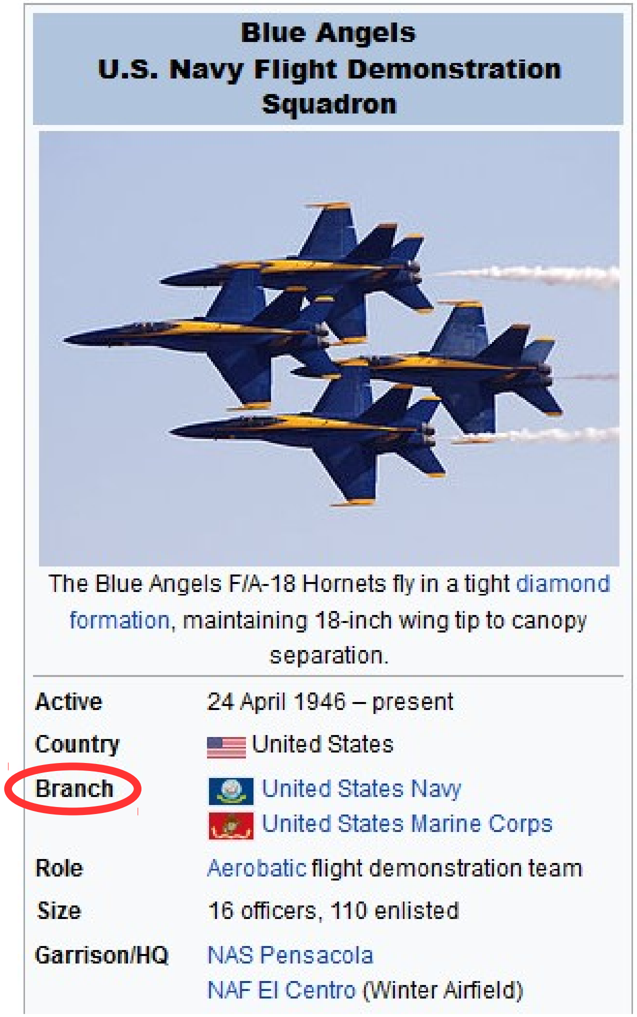
Inconsistency Detection

Identification of suspicious mappings and ontology constructs

Paulheim: *Data-driven Joint Debugging of the DBpedia Mappings and Ontology.* ESWC 2017

Common Errors in Knowledge Graphs

- Case 1: Wrong mapping
- Example:
  - *branch* in infobox military unit is mapped to *dbo:militaryBranch*
    - but *dbo:militaryBranch* has *dbo:Person* as its domain
  - correction: *dbo:commandStructure*
  - Affects 12,172 statements (31% of all *dbo:militaryBranch*)



Blue Angels
U.S. Navy Flight Demonstration Squadron

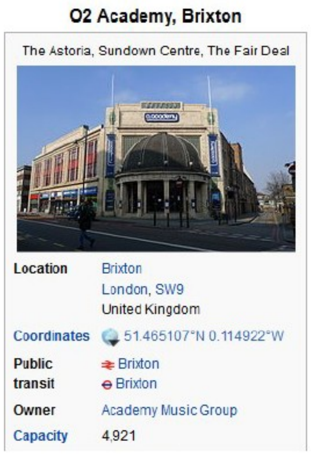
The Blue Angels F/A-18 Hornets fly in a tight diamond formation, maintaining 18-inch wing tip to canopy separation.

Active	24 April 1946 – present
Country	United States
Branch	United States Navy United States Marine Corps
Role	Aerobatic flight demonstration team
Size	16 officers, 110 enlisted
Garrison/HQ	NAS Pensacola NAF El Centro (Winter Airfield)

Common Errors in Knowledge Graphs

- Case 2: Mappings that should be removed
- Example:
  - *dbo:picture*
  - Most of the are inconsistent (64.5% places, 23.0% persons)
  - Reason: statements are extracted from picture *caption*

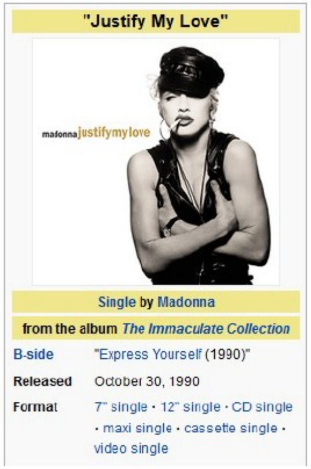
```
dbr:Brixton_Academy
    dbo:picture
    dbr:Brixton .

dbr:Justify_My_Love
    dbo:picture
    dbr:Madonna_(entertainer) .
```



O2 Academy, Brixton

The Astoria, Sundown Centre, The Fair Deal

Location	Brixton London, SW9 United Kingdom
Coordinates	51.465107°N 0.114922°W
Public transit	Brixton Brixton
Owner	Academy Music Group
Capacity	4,921



"Justify My Love"

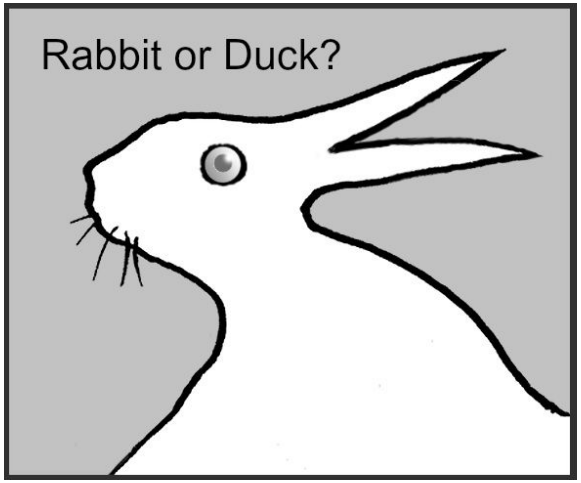
Single by Madonna

from the album *The Immaculate Collection*

B-side	"Express Yourself (1990)"
Released	October 30, 1990
Format	7" single · 12" single · CD single · maxi single · cassette single · video single

Common Errors in Knowledge Graphs

- Case 3: Ontology problems (domain/range)
- Example 1:
  - Populated places (e.g., cities) are used both as place and organization
  - For some properties, the range is either one of the two
    - e.g., *dbo:operator* (see introductory example)
  - Polysemy should be reflected in the ontology
- Example 2:
  - *dbo:architect*, *dbo:designer*, *dbo:engineer* etc. have *dbo:Person* as their range
  - Significant fractions (8.6%, 7.6%, 58.4%, resp.) have a *dbo:Organization* as object
  - Range should be broadened

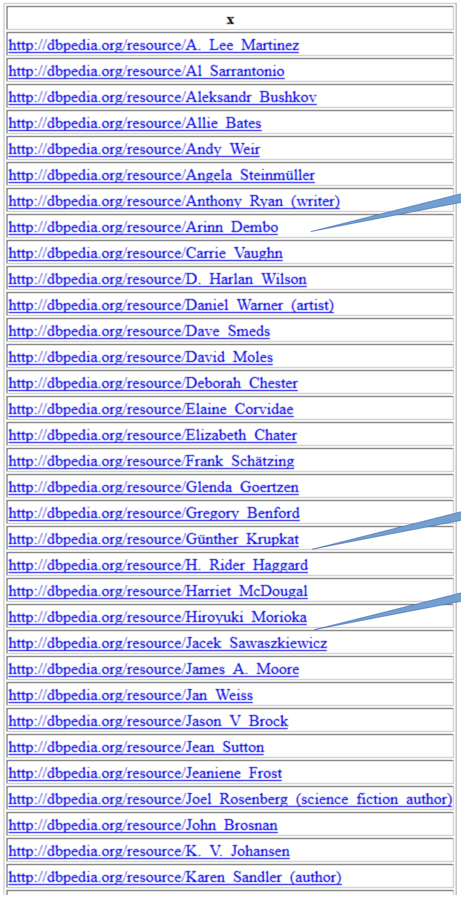


Common Errors in Knowledge Graphs

- Case 4: Missing properties
- Example 1:
  - *dbo:president* links an organization to its president
  - Majority use (8,354, or 76.2%):
    link a person to the president s/he served for
- Example 2:
  - *dbo:instrument* links an artist
    to the instrument s/he plays
  - Prominent alternative use (3,828, or 7.2%):
    links a genre to its characteristic instrument

Common Errors in Knowledge Graphs

- Introductory example:



x
http://dbpedia.org/resource/A._Lee_Martinez
http://dbpedia.org/resource/Al_Sarrantonio
http://dbpedia.org/resource/Aleksandr_Bushkov
http://dbpedia.org/resource/Allie_Bates
http://dbpedia.org/resource/Andy_Weir
http://dbpedia.org/resource/Angela_Steinmüller
http://dbpedia.org/resource/Anthony_Ryan_(writer)
http://dbpedia.org/resource/Arinn_Dembo
http://dbpedia.org/resource/Carrie_Vaughn
http://dbpedia.org/resource/D._Harlan_Wilson
http://dbpedia.org/resource/Daniel_Warner_(artist)
http://dbpedia.org/resource/Dave_Smeds
http://dbpedia.org/resource/David_Moles
http://dbpedia.org/resource/Deborah_Chester
http://dbpedia.org/resource/Elaine_Corvidae
http://dbpedia.org/resource/Elizabeth_Chater
http://dbpedia.org/resource/Frank_Schätzing
http://dbpedia.org/resource/Glenda_Goertzen
http://dbpedia.org/resource/Gregory_Benford
http://dbpedia.org/resource/Günther_Krupkat
http://dbpedia.org/resource/H._Rider_Haggard
http://dbpedia.org/resource/Harriet_McDougal
http://dbpedia.org/resource/Hiroyuki_Morioka
http://dbpedia.org/resource/Jacek_Sawaszkiewicz
http://dbpedia.org/resource/James_A._Moore
http://dbpedia.org/resource/Jan_Weiss
http://dbpedia.org/resource/Jason_V_Brock
http://dbpedia.org/resource/Jean_Sutton
http://dbpedia.org/resource/Jeaniene_Frost
http://dbpedia.org/resource/Joel_Rosenberg_(science_fiction_author)
http://dbpedia.org/resource/John_Brosnan
http://dbpedia.org/resource/K._V._Johansen
http://dbpedia.org/resource/Karen_Sandler_(author)

Arthur C. Clarke?

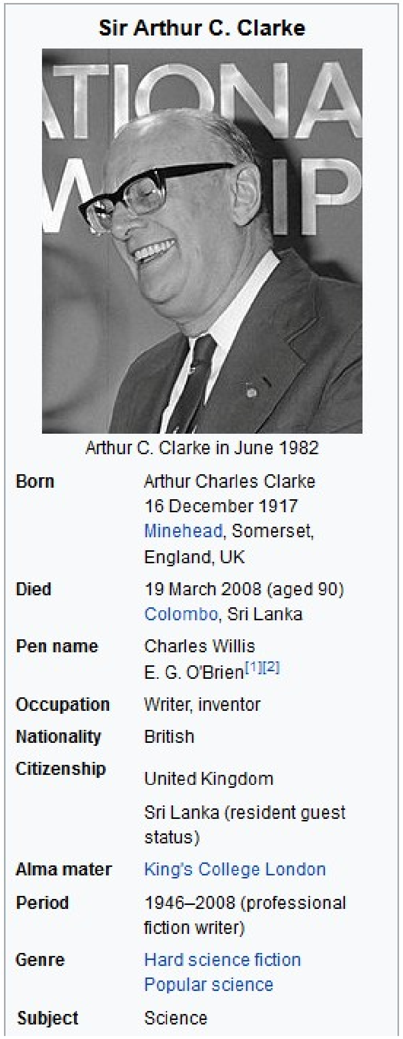
```
select ?x where
      {?x a dbo:Writer .
       ?x dbo:genre dbr:Science_Fiction}
order by ?x
```

H.G. Wells?

Isaac Asimov?

Common Errors in Knowledge Graphs

- Incompleteness in relation assertions
- Example: Arthur C. Clarke, Isaac Asimov, ...
  - There is no explicit link to Science Fiction in the infobox
  - i.e., the statement for
    `... dbo:genre dbr:Science_Fiction`
    is not generated



Sir Arthur C. Clarke

Arthur C. Clarke in June 1982

Born	Arthur Charles Clarke 16 December 1917 Minehead, Somerset, England, UK
Died	19 March 2008 (aged 90) Colombo, Sri Lanka
Pen name	Charles Willis E. G. O'Brien[1][2]
Occupation	Writer, inventor
Nationality	British
Citizenship	United Kingdom Sri Lanka (resident guest status)
Alma mater	King's College London
Period	1946–2008 (professional fiction writer)
Genre	Hard science fiction Popular science
Subject	Science

Common Errors in Knowledge Graphs

- Example for recent work (ISWC 2017):
  heuristic relation extraction from Wikipedia abstracts
- Idea:
  - There are probably certain patterns:
    - e.g., all genres linked in an abstract about a writer are that writer’s genres
    - e.g., the first place linked in an abstract about a person is that person’s birthplace
  - The types are already in DBpedia
  - We can use existing relations as training data
    - Using a local closed world assumption for negative examples
  - Learned models can be evaluated and only used at a certain precision

Heist & Paulheim: *Language-agnostic relation extraction from Wikipedia Abstracts*. ISWC 2017

Common Errors in Knowledge Graphs

- Results:
  - 1M additional assertions can be learned for 100 relations at 95% precision

- Additional consideration:
  - We use only links, types from DBpedia, and positional features
  - No language-specific information (e.g., POS tags)
  - Thus, we are not restricted to English!

Heist & Paulheim: *Language-agnostic relation extraction from Wikipedia Abstracts*. ISWC 2017

Common Errors in Knowledge Graphs

- Cross-lingual experiment:
  - Using the 12 largest language editions of Wikipedia
  - Exploiting inter-language links

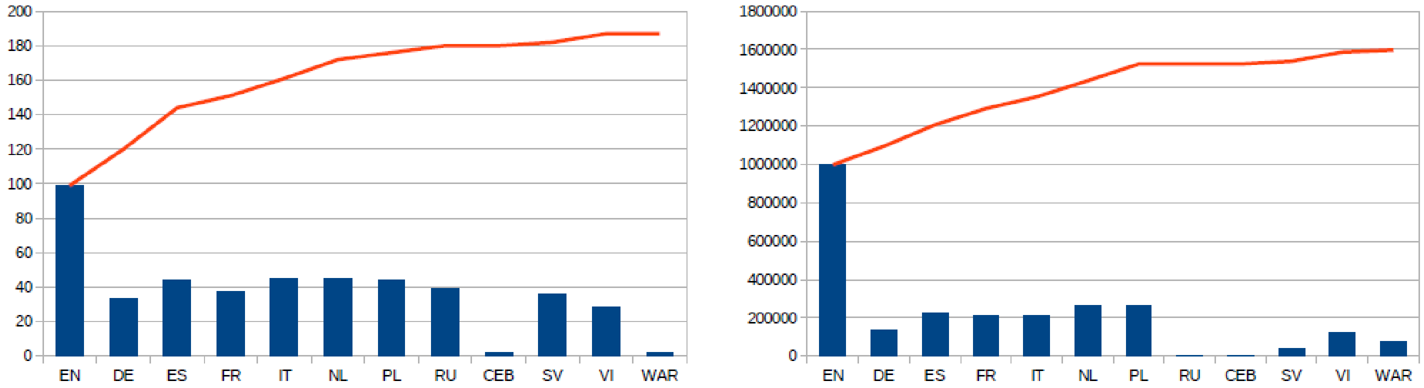
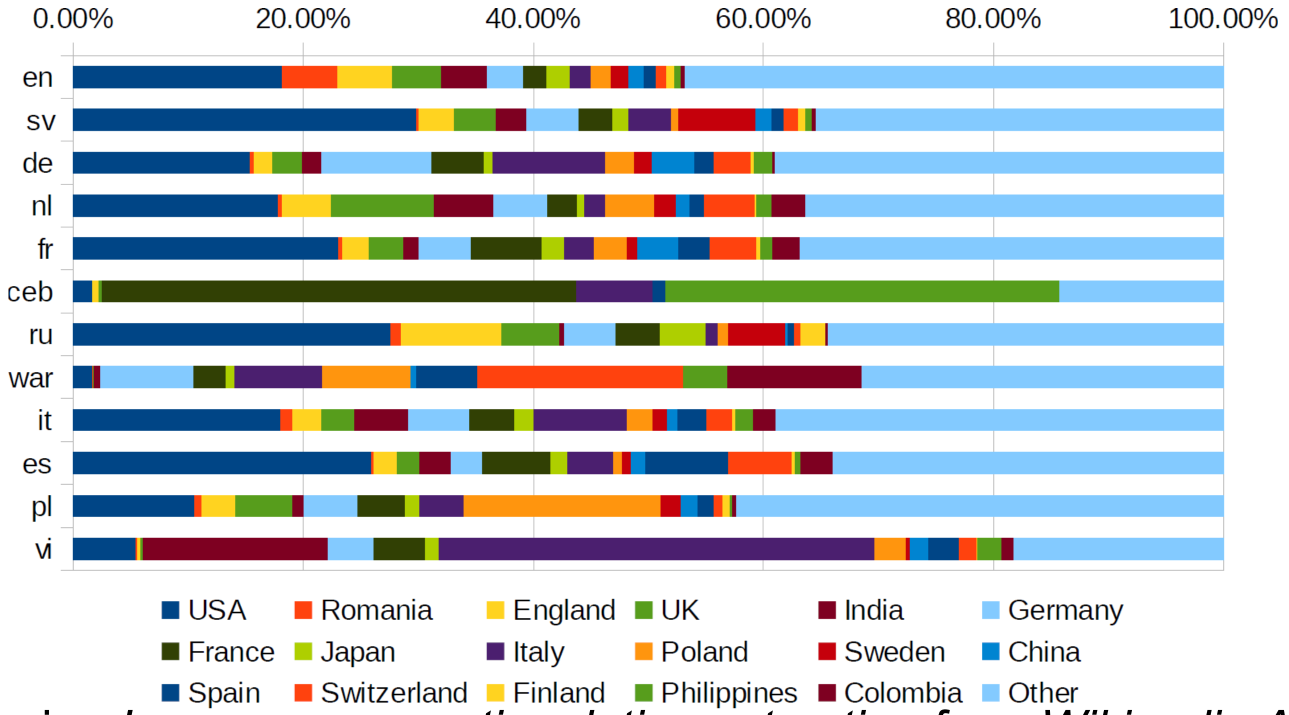


Fig. 3: Number of relations (left) and statements (right) extracted at 95% precision in the top 12 languages. The bars show the number of statements that could be extracted for the given language, the line depicts the accumulated number of statements for the top N languages.

Heist & Paulheim: *Language-agnostic relation extraction from Wikipedia Abstracts*. ISWC 2017

Common Errors in Knowledge Graphs

- Analysis
  - Is there a relation between the language and the the country (dbo:country) of the entities for which information is extracted?



Heist & Paulheim: *Language-agnostic relation extraction from Wikipedia Abstracts*. ISWC 2017

Common Errors in Knowledge Graphs

- So far, we have looked at relation assertions
- Numerical values can also be problematic…
  - Recap: Wikipedia is made for human consumption

- The following are all valid representations of the same height value (and perfectly understandable by humans)
  - 6 ft 6 in, 6ft 6in, 6'6'', 6'6”, 6´6´´, ...
  - 1.98m, 1,98m, 1m 98, 1m 98cm, 198cm, 198 cm, ...
  - 6 ft 6 in (198 cm), 6ft 6in (1.98m), 6'6'' (1.98 m), ...
  - 6 ft 6 in[1], 6 ft 6 in [citation needed], ...
  - ...

Wienand & Paulheim: *Detecting Incorrect Numerical Data in DBpedia.* ESWC 2014
Fleischhacker et al.: *Detecting Errors in Numerical Linked Data Using Cross-Checked Outlier Detection.* ISWC 2014

Common Errors in Knowledge Graphs

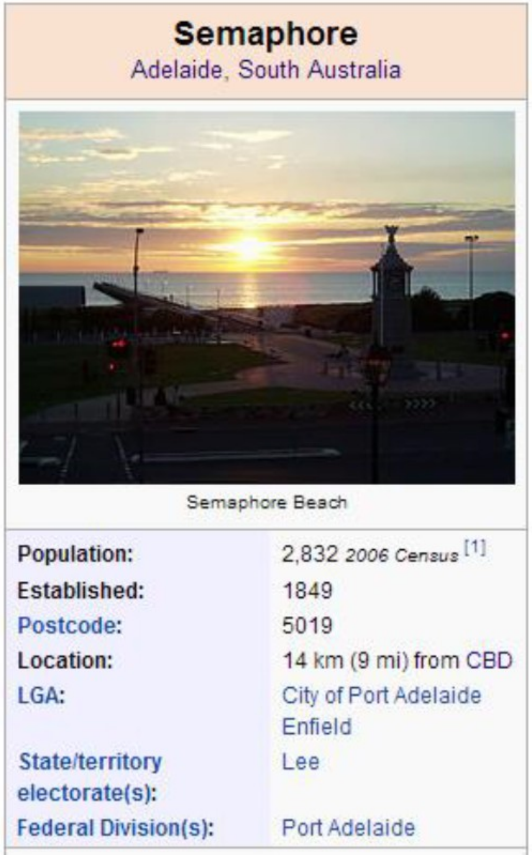
- Approach: outlier detection
  - With preprocessing: finding meaningful subpopulations
  - With cross-checking: discarding natural outliers

- Findings: 85%-95% precision possible
  - depending on predicate
  - Identification of typical parsing problems

Wienand & Paulheim: *Detecting Incorrect Numerical Data in DBpedia.* ESWC 2014
Fleischhacker et al.: *Detecting Errors in Numerical Linked Data Using Cross-Checked Outlier Detection.* ISWC 2014

Common Errors in Knowledge Graphs

- Errors include
  - Interpretation of imperial units
  - Unusual decimal/thousands separators
  - Concatenation (population 28,322,006)



Semaphore
Adelaide, South Australia

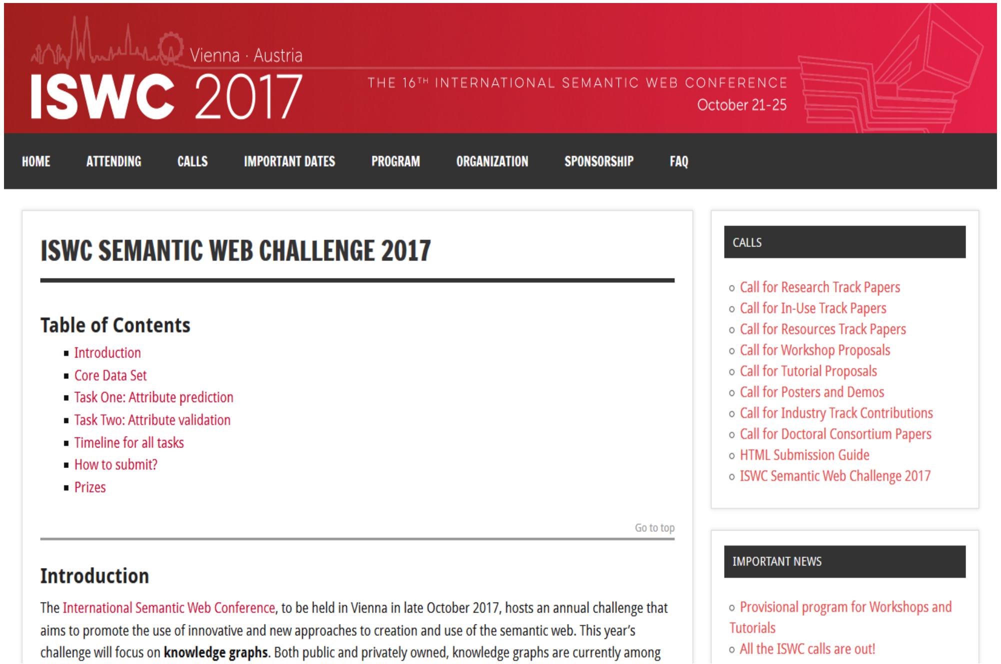
Semaphore Beach

Population:	2,832 *2006 Census* [1]
Established:	1849
Postcode:	5019
Location:	14 km (9 mi) from CBD
LGA:	City of Port Adelaide Enfield
State/territory electorate(s):	Lee
Federal Division(s):	Port Adelaide

Wienand & Paulheim: *Detecting Incorrect Numerical Data in DBpedia.* ESWC 2014
Fleischhacker et al.: *Detecting Errors in Numerical Linked Data Using Cross-Checked Outlier Detection.* ISWC 2014

Common Errors in Knowledge Graphs

- Got curious? Want to get your hands dirty?
  - 2017 Semantic Web Challenge revolves around knowledge graph completion and correction
  - Using permid.org



ISWC SEMANTIC WEB CHALLENGE 2017

Table of Contents

- Introduction
- Core Data Set
- Task One: Attribute prediction
- Task Two: Attribute validation
- Timeline for all tasks
- How to submit?
- Prizes

Introduction

The International Semantic Web Conference, to be held in Vienna in late October 2017, hosts an annual challenge that aims to promote the use of innovative and new approaches to creation and use of the semantic web. This year's challenge will focus on **knowledge graphs**. Both public and privately owned, knowledge graphs are currently among

CALLS

- Call for Research Track Papers
- Call for In-Use Track Papers
- Call for Resources Track Papers
- Call for Workshop Proposals
- Call for Tutorial Proposals
- Call for Posters and Demos
- Call for Industry Track Contributions
- Call for Doctoral Consortium Papers
- HTML Submission Guide
- ISWC Semantic Web Challenge 2017

IMPORTANT NEWS

- Provisional program for Workshops and Tutorials
- All the ISWC calls are out!

https://iswc2017.semanticweb.org/calls/iswc-semantic-web-challenge-2017/

New Kids on the Block

Subjective age:
Measured by the fraction
of the audience
that understands a reference
to your young days'
pop culture...

New Kids on the Block

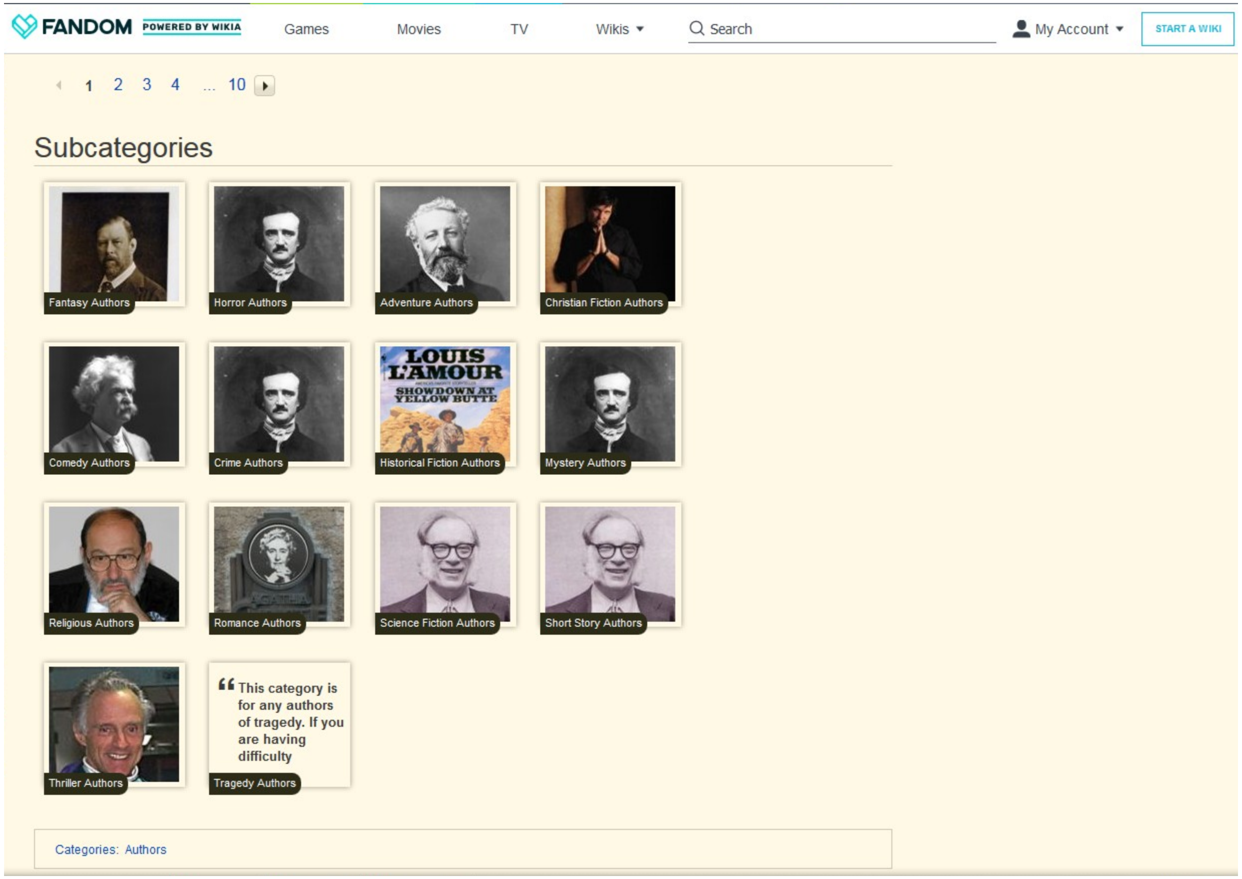
- Wikipedia-based Knowledge Graphs will remain an essential building block of Semantic Web applications
- But they suffer from...
  - ...a coverage bias
  - ...limitations of the creating heuristics

Work in Progress: 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

Work in Progress: DBkWik

- Back to our original example...



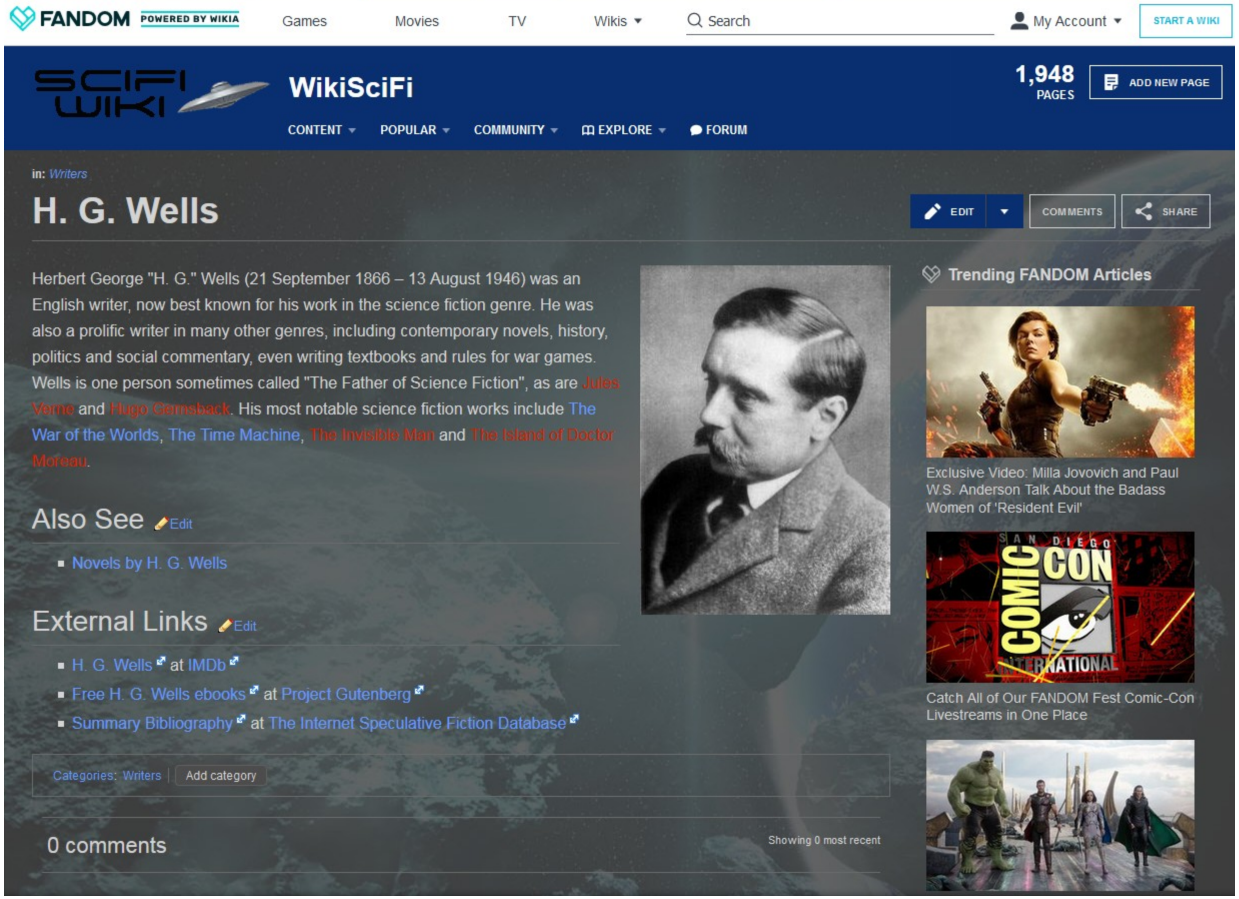
Subcategories

Fantasy Authors
Horror Authors
Adventure Authors
Christian Fiction Authors
Comedy Authors
Crime Authors
Historical Fiction Authors
Mystery Authors
Religious Authors
Romance Authors
Science Fiction Authors
Short Story Authors
Thriller Authors
“This category is for any authors of tragedy. If you are having difficulty
Tragedy Authors

Categories: Authors

Work in Progress: DBkWik

- Back to our original example...



in: Writers

H. G. Wells

Herbert George "H. G." Wells (21 September 1866 – 13 August 1946) was an English writer, now best known for his work in the science fiction genre. He was also a prolific writer in many other genres, including contemporary novels, history, politics and social commentary, even writing textbooks and rules for war games. Wells is one person sometimes called "The Father of Science Fiction", as are Jules Verne and Hugo Gernsback. His most notable science fiction works include The War of the Worlds, The Time Machine, The Invisible Man and The Island of Doctor Moreau.

Also See Edit

- Novels by H. G. Wells

External Links Edit

- H. G. Wells at IMDb
- Free H. G. Wells ebooks at Project Gutenberg
- Summary Bibliography at The Internet Speculative Fiction Database

Categories: Writers | Add category

0 comments

Showing 0 most recent

Exclusive Video: Milla Jovovich and Paul W.S. Anderson Talk About the Badass Women of 'Resident Evil'

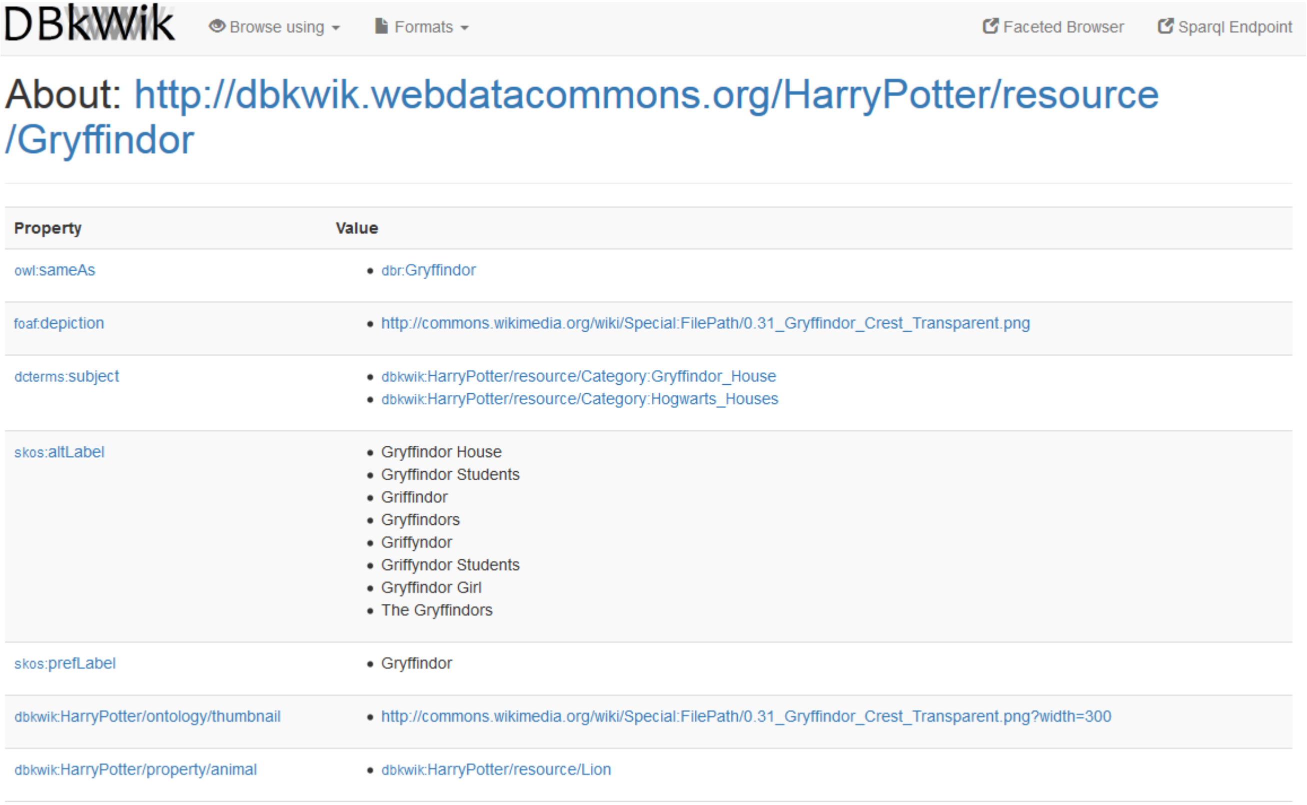
Catch All of Our FANDOM Fest Comic-Con Livestreams in One Place

Work in Progress: DBkWik

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

Work in Progress: 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 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.webdatacommons.org/

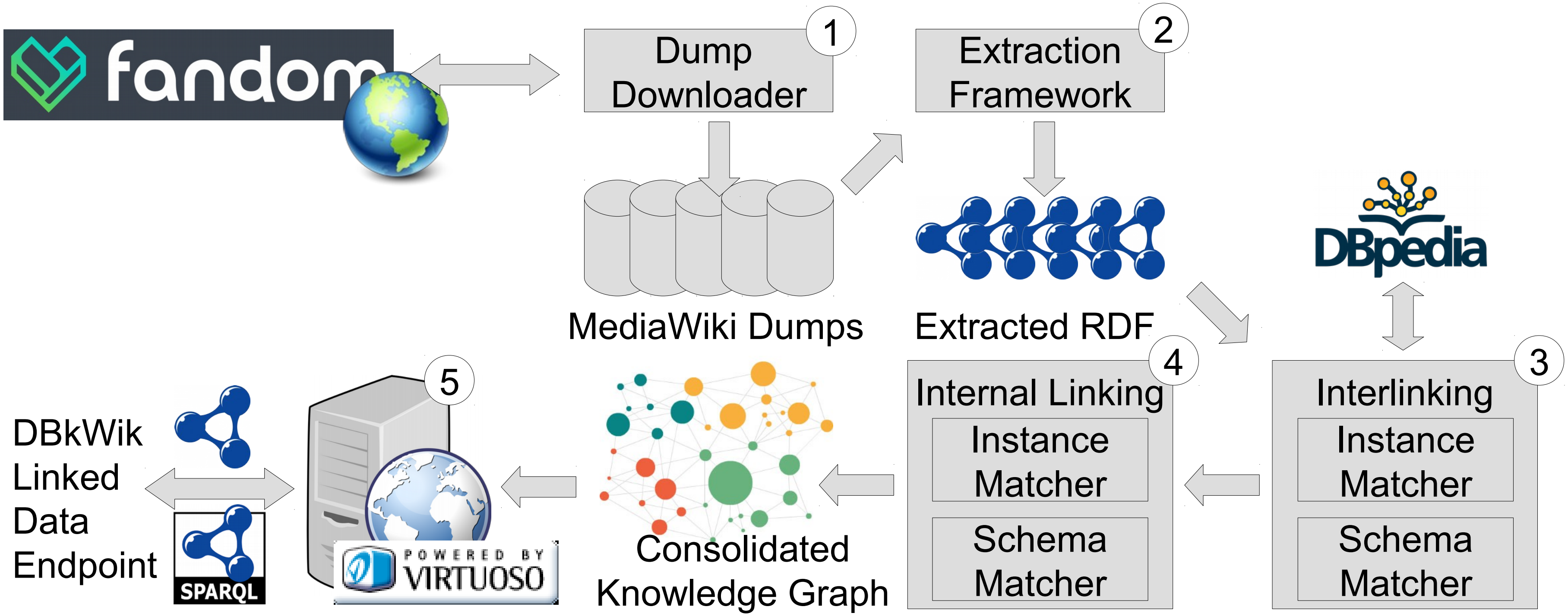
Work in Progress: DBkWik

- 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

Work in Progress: DBkWik

- Avoiding O(n²) internal linking:
  - Match to DBpedia first
  - Use common links to DBpedia as blocking keys for internal matching



Work in Progress: DBkWik

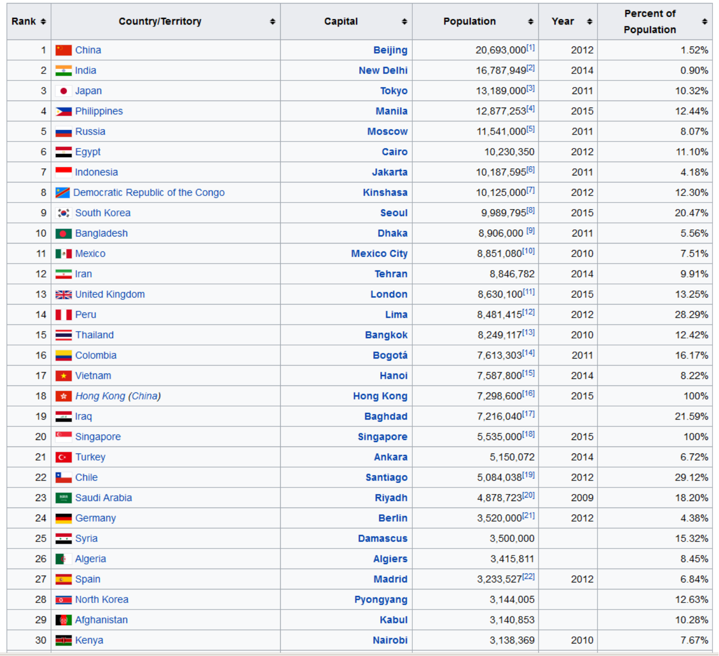
- Downloaded ~15k Wiki dumps from Fandom
  - 52.4GB of data, roughly the size of the English Wikipedia

- 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)

Work in Progress: 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%

Work in Progress: WebIsALOD

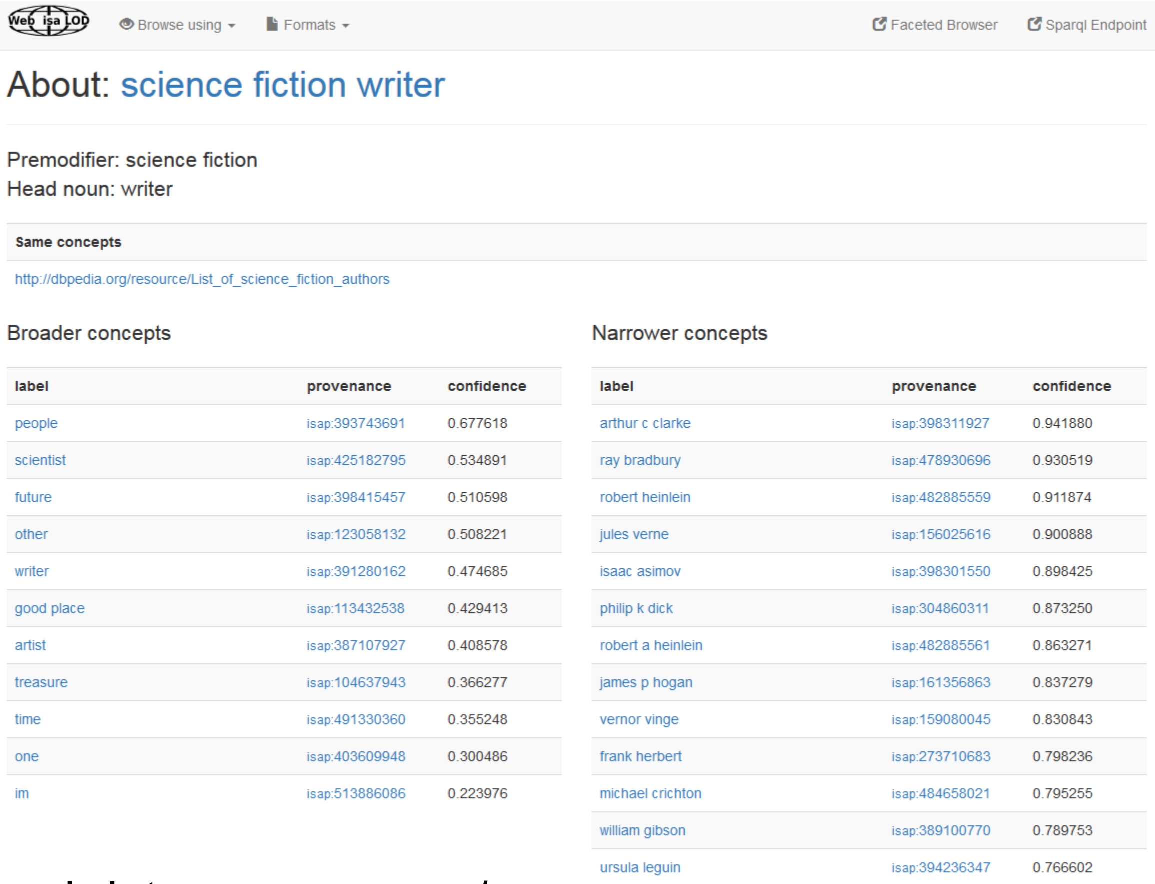
- Extraction of type information using Hearst-like patterns, e.g.,
  - T, such as X
  - X, Y, and other T
- Text corpus: common crawl
  - ~2 TB crawled web pages
  - Fast implementation: regex over text
  - “Expensive” operations only applied once regex has fired
- Resulting database
  - 400M hypernymy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

Common Crawl

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

Work in Progress: WebIsALOD

- Back to our original 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

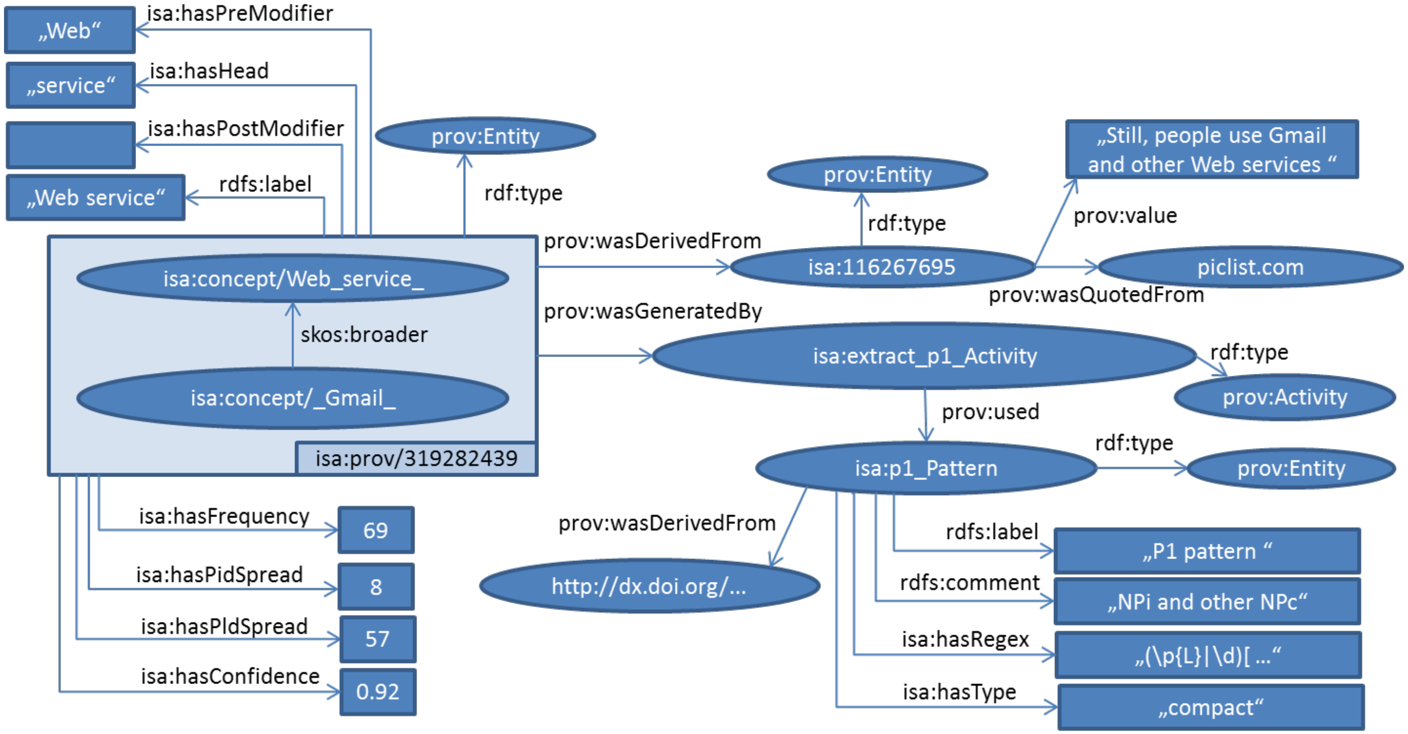
Narrower concepts

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/

Work in Progress: WebIsALOD

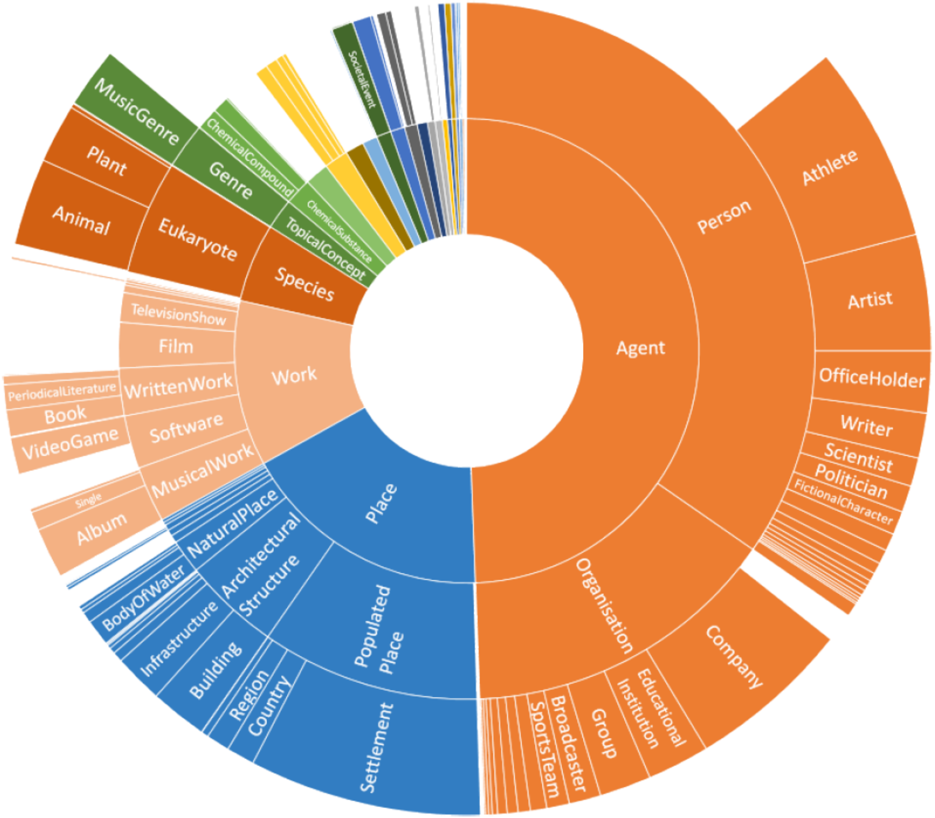
- Initial effort: transformation to a LOD dataset
  - including rich provenance information



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

Work in Progress: WebIsALOD

- Estimated contents breakdown



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

Work in Progress: 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

Work in Progress: WebIsALOD

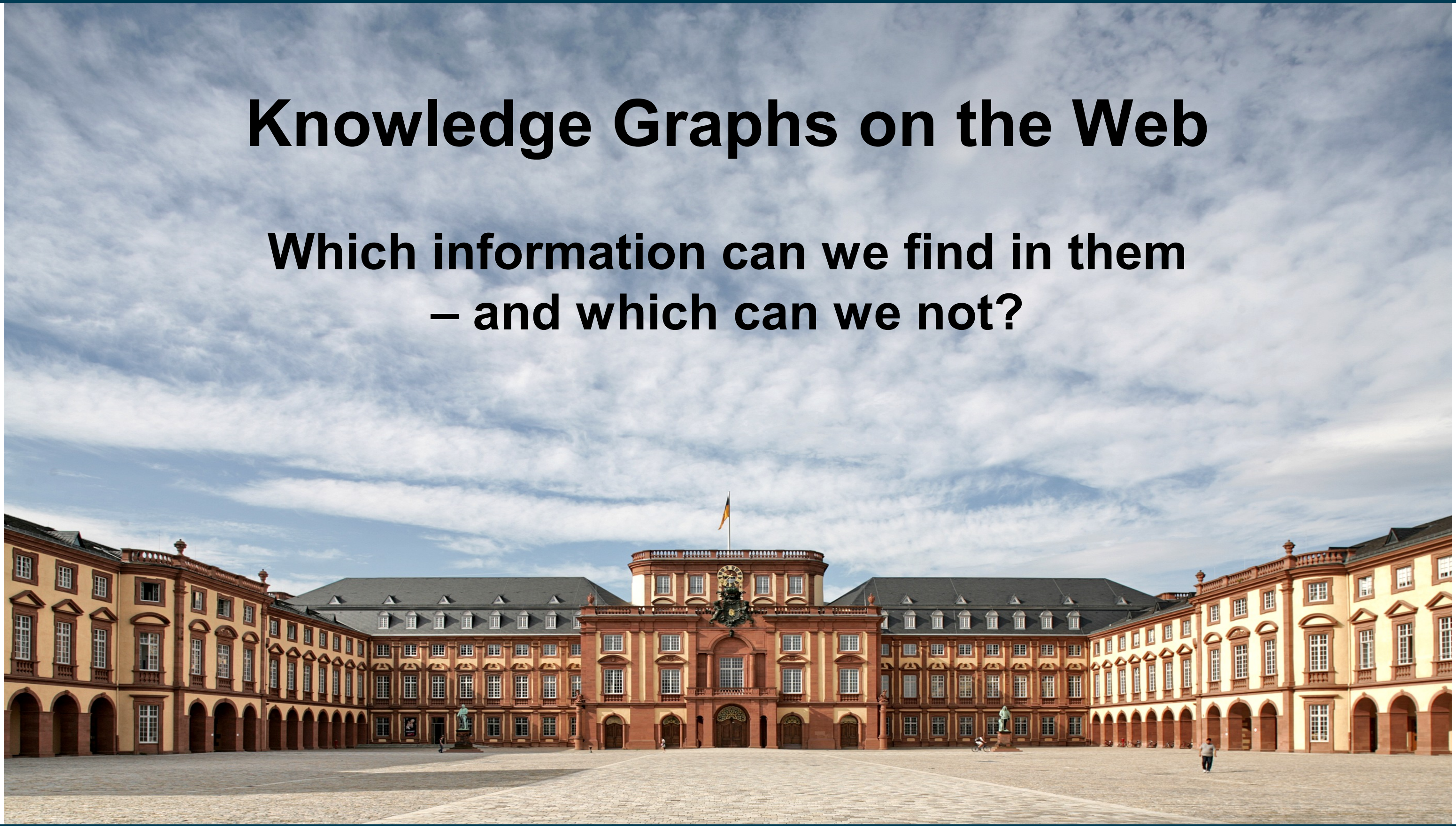
- Current challenges and works in progress
  - Distinguishing instances and classes
    - i.e.: subclass vs. instance of relations
  - Splitting instances
    - *Bauhaus is a goth band*
    - *Bauhaus is a German school*
  - Knowledge extraction from pre and post modifiers
    - *Bauhaus is a goth band* → 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

Take Aways

- Knowledge Graphs contain a massive amount of information
  - Various trade offs in their creation
- We can find it if...
  - ...it is in there
  - ...the clues we need to find it are in it and correct
- Various methods exist for
  - ...completing knowledge graphs
  - ...identifying errors
  - ...lately also: identifying the roots of errors
- New kids on the block
  - DBkWik and WebIsALOD
  - Focus on long tail entities

Credits & Contributions



Knowledge Graphs on the Web

Which information can we find in them – and which can we not?

Heiko Paulheim