NECKAr: A Named Entity Classifier for Wikidata

Johanna Geiß, Andreas Spitz, Michael Gertz

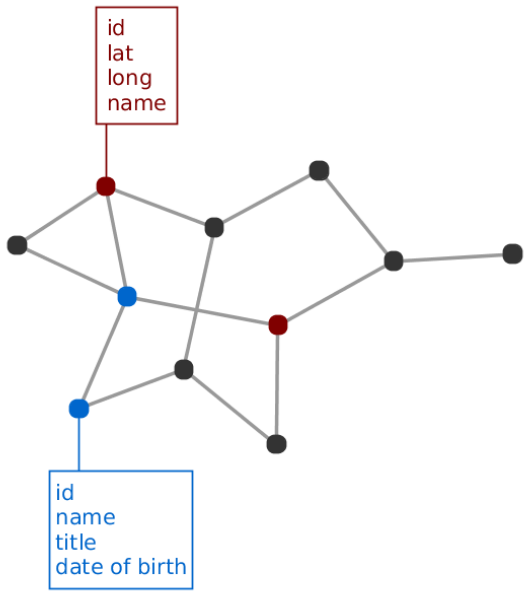
Heidelberg University, Institute of Computer Science
Database Systems Research Group

{geiss,spitz,gertz}@informatik.uni-heidelberg.de

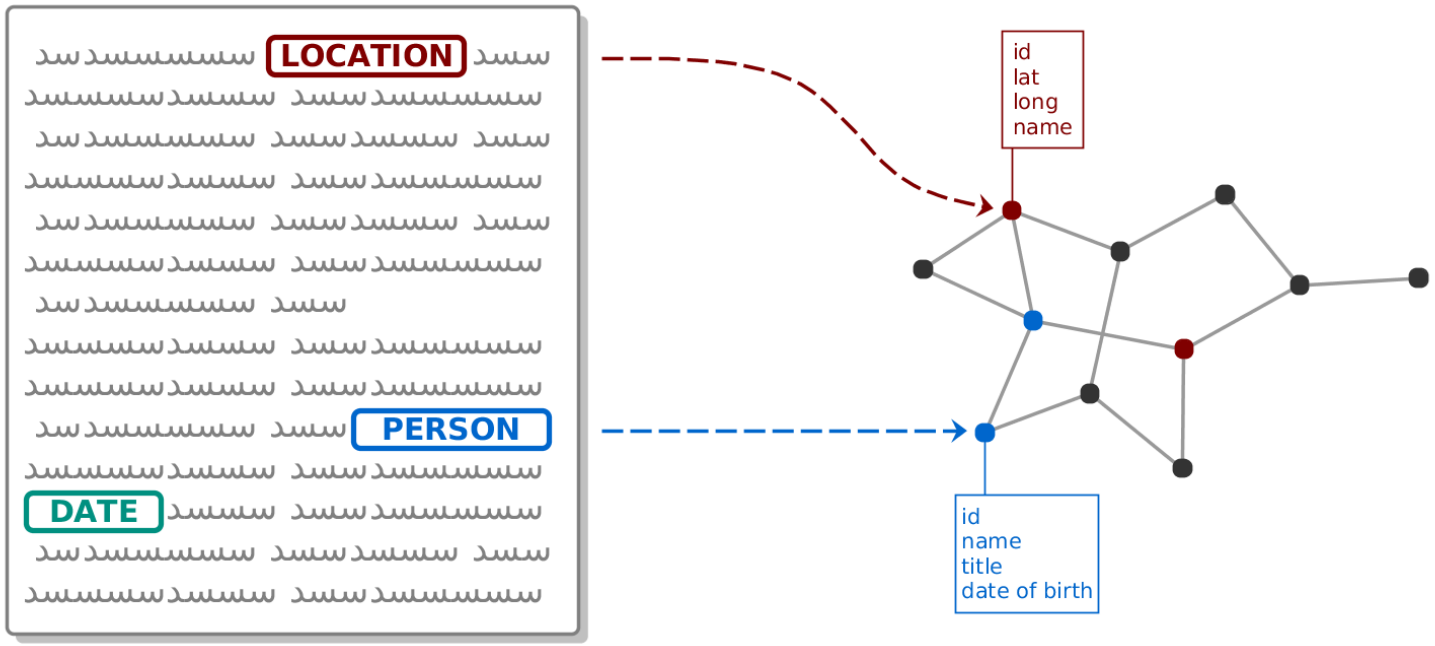
GSCL
Berlin, Sept 14, 2017

> "Knowledge is power."
>
> — Francis Bacon

Knowledge Bases and Entity Linking



Knowledge Bases and Entity Linking



Knowledge Bases in NLP & IE

Many applications are improved by using knowledge base linking

- Geolocation of documents
- Anaphora resolution
- Query expansion
- Event detection
- Entity-centric summarization
- Knowledge extraction
- ...

Prevalent Knowledge Bases

Issues of Existing KBs

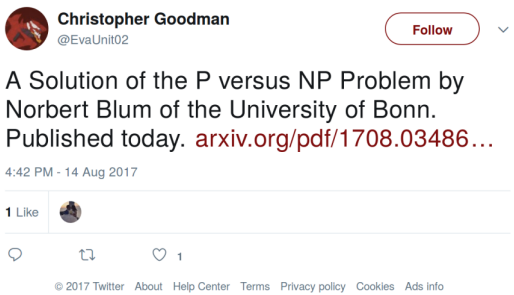
Accessibility of information:

- Google Knowledge Graph is API only

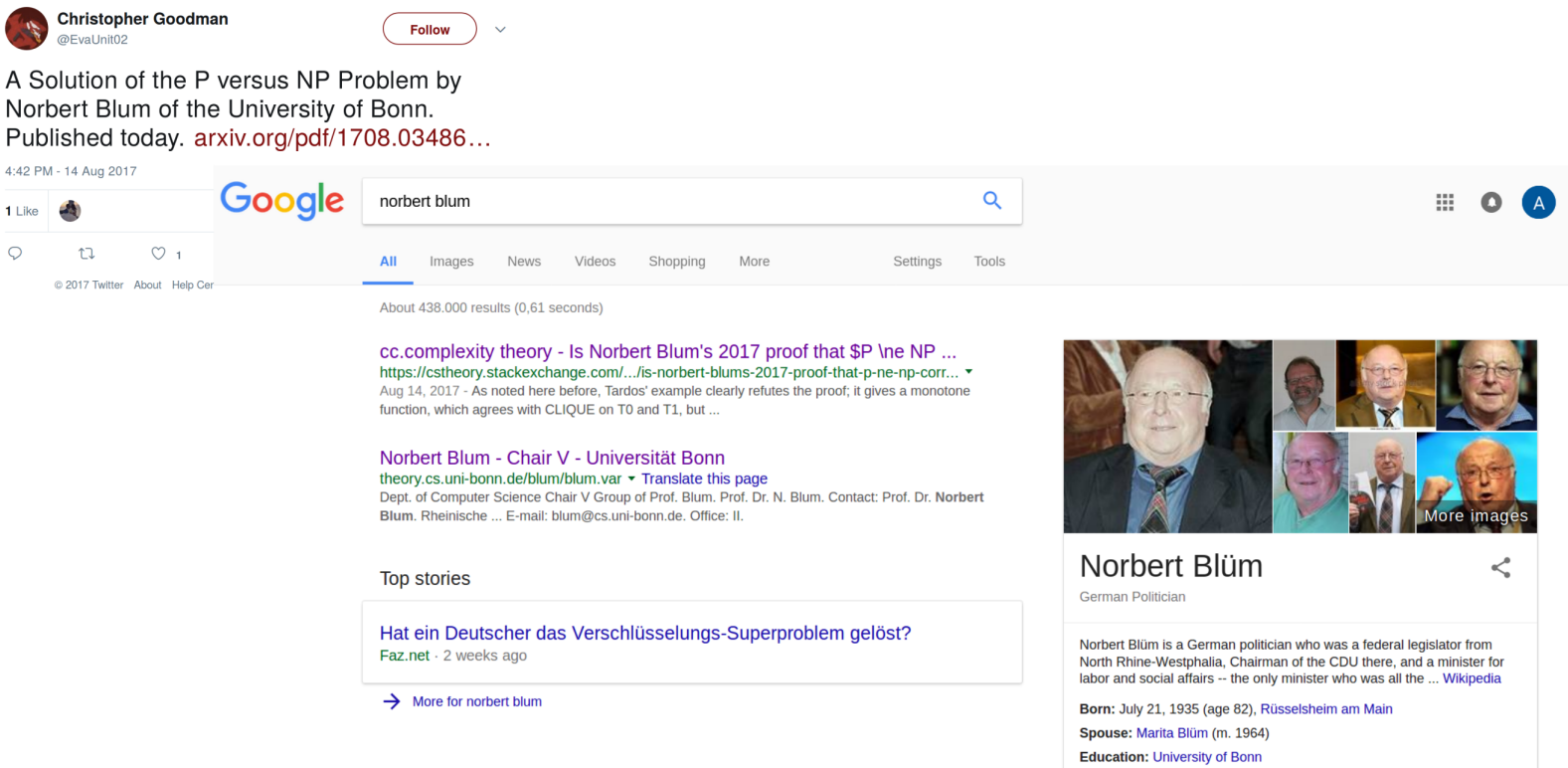
Currency of information:

- Freebase was discontinued in 2016
- DBpedia updates twice per year (2016-10, 2016-04, 2015-10, ...)
- YAGO updates irregularly (2017-05, 2014-06, 2012-11, ...)

Currency of Entities in News and Social Media



Currency of Entities in News and Social Media

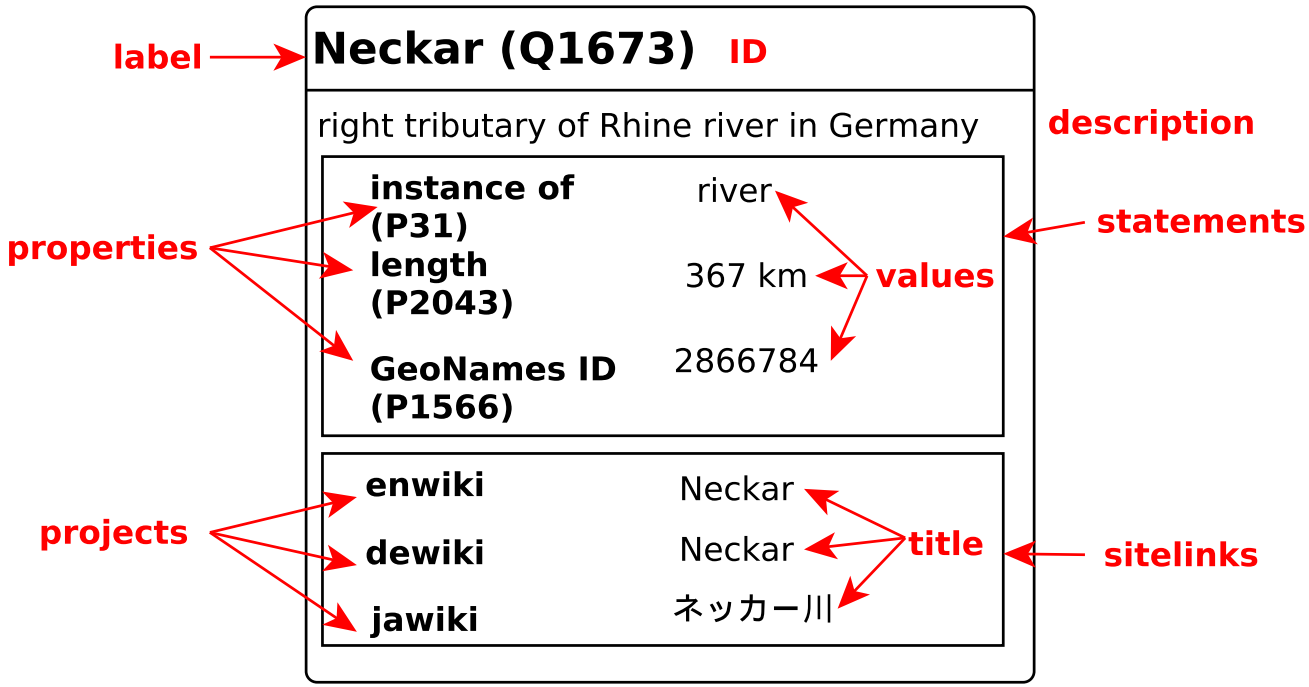


The Advantages of Wikidata

Why Wikidata is a useful resource:

- Collaboratively edited and always current
- Inherently multilingual
- Contains (multiple) claims, not facts
- Direct integration with Wikipedia
- No versioning for SPARQL access (updated incrementally)

Wikidata Item Structure



Disadvantages of Wikidata

Why Wikidata is difficult to use in research [SDR+16]:

- Convoluted, constantly evolving hierarchies
- No skeletal hierarchies
- No versioning for SPARQL access (updated incrementally)

The Importance of Entity Classification

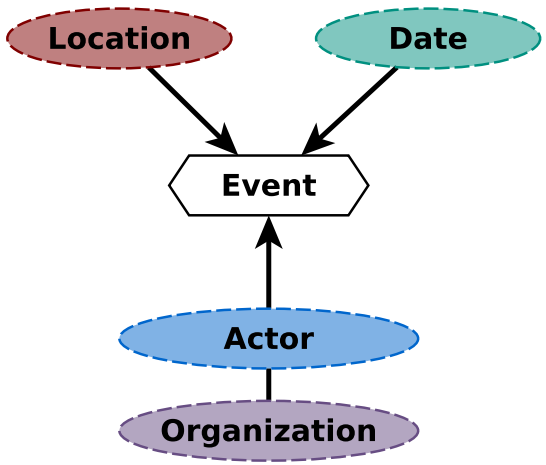
The Five Ws of information gathering:

- **Who** was involved?
- What happened?
- **When** did it take place?
- **Where** did it take place?
- Why did that happen?

The Importance of Entity Classification

The Five Ws of information gathering:

- **Who** was involved?
- What happened?
- **When** did it take place?
- **Where** did it take place?
- Why did that happen?



Definition: Event

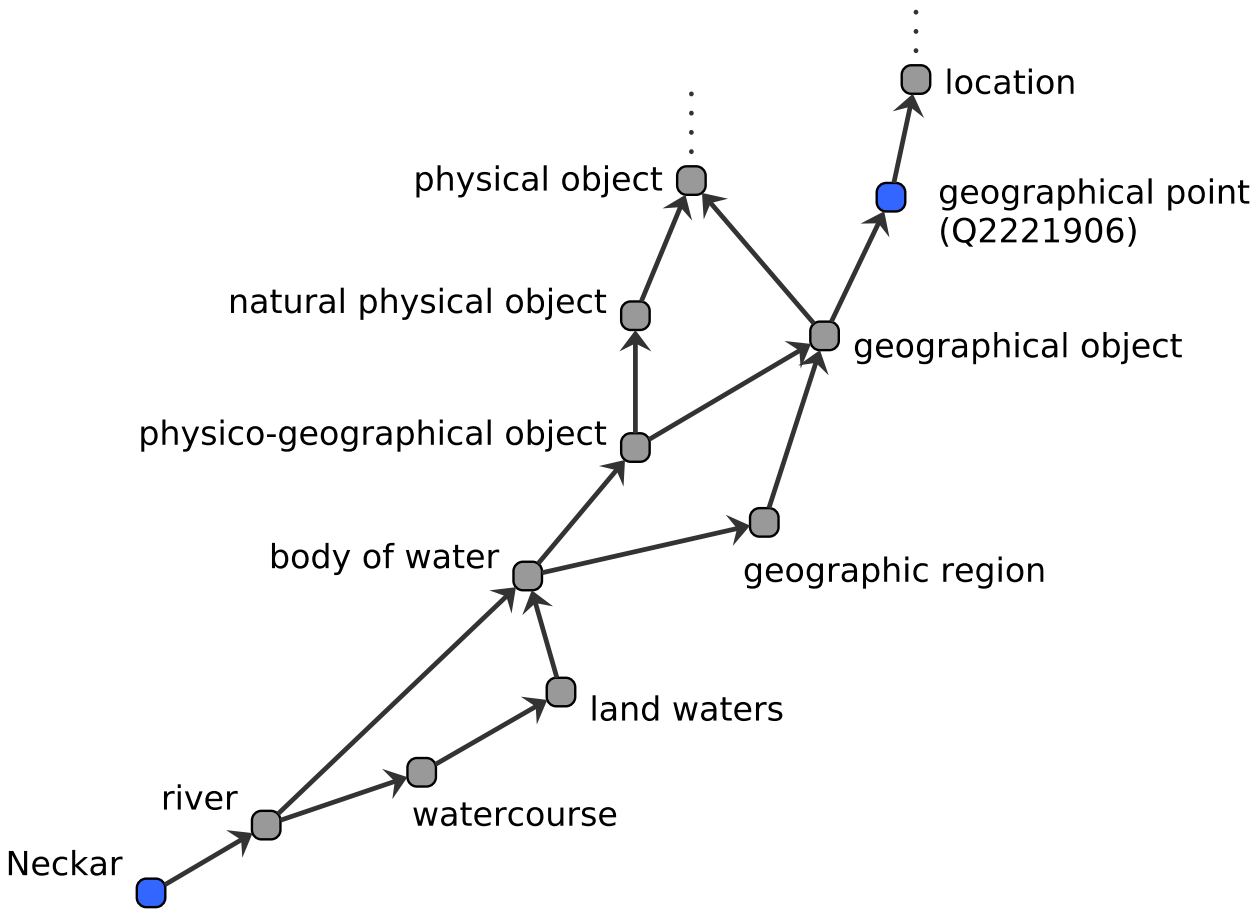
"Something that happens at a given **place** *and* **time** *between a* **group** *of* **actors**.*"* [CSG+02]

The NECKAr Classification Scheme

Contributions and purpose of NECKAr:

- Classify entities in Wikidata (PER, LOC, ORG)
- Extract easy-to-use data sets from Wikidata dumps
- Enrich entities with commonly used additional information
- Ensure reproducibility of subsequent applications

Wikidata Item Hierarchy



Location Extraction

Extract for items in the tree of **geographical point** (Q2221906):

- Coordinate location (P625)
- Population (P1082)
- Country (P17)
- Continent (P30)
- Location types (city, mountain, river, etc.)

Additionally: exclude subtree of **food**.

Organization Extraction

Extract for items in the tree of **organization** (Q43229):

- Sovereign state of (P17)
- Founder (P112)
- CEO (P169)
- Inception (P571)
- Headquarter location (P159)
- Official website (P856)
- Official language (P37)

Person Extraction

Extract for items that are instances of **human** (Q5):

- Date of birth (P569)
- Date of death (P570)
- Gender (P21)
- Occupation (P106)
- Alternative names

Note: excludes fictional characters.

NECKAr Data Set Examples

neClass	LOCATION		ORGANIZATION		PERSON
id	Q1796771	**id**	Q81230	**id**	Q76658
norm_name	Köthen	**norm_name**	Siemens	**norm_name**	Frank-Walter Steinmeier
description	capital of the district of Anhalt-Bitterfeld Saxony-Anhalt	**description**	Engineering and electronics conglomerate	**description**	politician
en Wikipedia	Köthen (Anhalt)	**en Wikipedia**	Siemens	**en Wikipedia**	Frank-Walter Steinmeier
location type	city, settlement	**instance of**	concern, bus. enterprise	**occupation**	politician, jurist, lawyer
population	26,384				
continent	Europe	**CEO**	Joe Kaeser	**gender**	male
country	Germany		Klaus Kleinfeld	**dob**	1956-01-05
coordinate	51.75 11.916666666667	**founder**	Ernst Werner von Siemens	**dod**	none
				alias	Steinmeier
GeoNames	2885237	**inception**	1847-10-01		
		HQ	Munich		
		country	Germany		
		website	www.siemens.com		

The NECKAr Named Entity Data Set

NECKAr for the Wikidata dump of December 2016:

- 8.8M extracted items
- 4.6M locations (51% with geocoordinates)
- 3.3M persons (66% with occupations)
- 900k organizations

Coverage Comparison to YAGO

neClass	NECKAr	Yago3	Yago3 ∩ Wikidata
LOC	4,582,947	1,267,402	1,250,409
PER	3,322,217	1,745,219	1,715,305
ORG	936,939	481,001	464,351

Precision Comparison to YAGO

neClass	F_1-Score	Precision	Recall
LOC	0.88	0.93	0.84
PER	0.97	0.99	0.95
ORG	0.57	0.54	0.60
combined	0.88	0.90	0.86

Summary and Outlook

NECKAr offers:

- Lightweight and multilingual set of Wikidata entities
- Large and current sets of named entities
- Links of entities to traditional knowledge bases

Outlook on upcoming changes:

- Refined class hierarchies and additional classes
- Automated process for monthly releases
- Optional use of Wikidata dump and SPARQL interface

Resources

NECKAr resources are available online:

- Named entity data sets
  (for multiple Wikidata dumps)
- Individual subsets for named entity classes
- Classification code for any Wikidata dump

`http://event.ifi.uni-heidelberg.de/`

Resources

NECKAr resources are available online:

- Named entity data sets
  (for multiple Wikidata dumps)
- Individual subsets for named entity classes
- Classification code for any Wikidata dump

`http://event.ifi.uni-heidelberg.de/`

Thank You!

Bibliography I


Christopher Cieri, Stephanie Strassel, David Graff, Nii Martey, Kara Rennert, and Mark Liberman.
Corpora for topic detection and tracking.
In *Topic Detection and Tracking*. Springer, 2002.


Andreas Spitz, Vaibhav Dixit, Ludwig Richter, Michael Gertz, and Johanna Geiß.
State of the union: A data consumer's perspective on Wikidata and its properties for the classification and resolution of entities.
In *WikiWorkshop with ICWSM*, 2016.