

Generierung angepasster RDF-Dumps von Wikidata

Bachelorverteidigung

Benno Fünfstück

17. Dezember 2019

Betreuer: Prof. Dr. Markus Krötzsch

Wissensbasierte Systeme

TU Dresden

The diagram illustrates the structure of a Wikidata item page for **Douglas Adams (Q42)**. The page is divided into several sections, each with specific annotations:

- Label:** The main title "Douglas Adams" is annotated with the label "label".
- Item Identifier:** The identifier "(Q42)" is annotated with "item identifier".
- Description:** The text "English writer and humorist" is annotated with "description". Below it, "Douglas Noël Adams | Douglas Noel Adams" is annotated with "aliases". A link "In more languages" is also present.
- Statements:** The "Statements" section is highlighted with a green border, labeled "statement group".
 - Property:** The "educated at" property is highlighted with a purple box, labeled "property".
 - Value:** The value "St John's College" is highlighted with an orange box, labeled "value".
 - Qualifiers:** A table of qualifiers for "St John's College" is highlighted with a blue box, labeled "qualifiers".

end time	1974
academic major	English literature
academic degree	Bachelor of Arts
start time	1971
 - Rank:** A vertical axis on the left of the statements is labeled "rank".
 - Opened References:** A table of references for "St John's College" is highlighted with a red box, labeled "opened references".

stated in	Encyclopædia Britannica Online
reference URL	http://www.nndb.com/people/731/000023662/
original language of work	English
retrieved	7 December 2013
publisher	NNDB
title	Douglas Adams (English)
 - Collapsed Reference:** A reference for "Brentwood School" is highlighted with a red box, labeled "collapsed reference".

end time	1970
start time	1959

Additional UI elements include a "+ add reference" link and a "+ add (statement)" link at the bottom of the statements section.



70 Millionen Items

800 Millionen Statements

vollständige Datendumps:
mehr als 42GiB komprimiert

Idee: Tool zum Filtern der Daten

GraFa: Faceted Search & Browsing for the Wikidata Knowledge Graph

(Moreno-Vega und Hogan 2018)

nur „truthy“ Statements mit Entitäten als Objekt; Labels/Descriptions nur in Englisch und Spanisch

Populating Narratives Using Wikidata Events: An Initial Experiment

(Metilli u. a. 2019)

alle Items mit einem Statement für mindestens eine von 50 festgelegten Properties

- Korrektheit, Vollständigkeit
- Filterung, Archivierung, Suche
- Statistiken, Fortschritt, Parallelverarbeitung
- Nachvollziehbarkeit, Aktualität der Daten

Format der Dumps: RDF

Wikidata als RDF (Resource Description Framework)

Beispiel

```
wd:Q42 rdfs:label "Douglas Adams"@en .  
wd:Q42 rdfs:label "Douglas Adams"@de .
```

einfache Darstellung: ein Tripel für jedes Statement

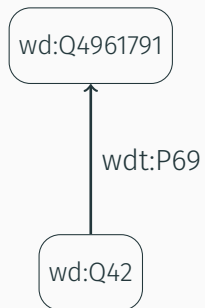
```
wd:Q42 wdt:P69 wd:Q691283 .  
wd:Q42 wdt:P69 wd:Q4961791 .
```

Aber: Ranks, Qualifier, komplexe Werte?

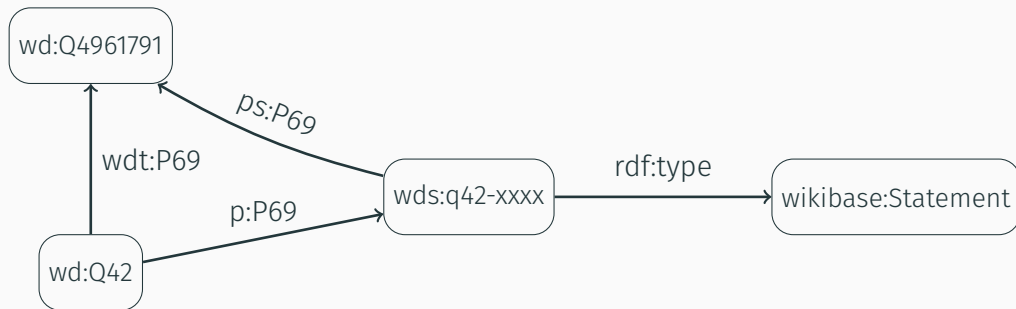
wd:Q42 ist kurz für `<http://www.wikidata.org/entity/Q42>`

RDF

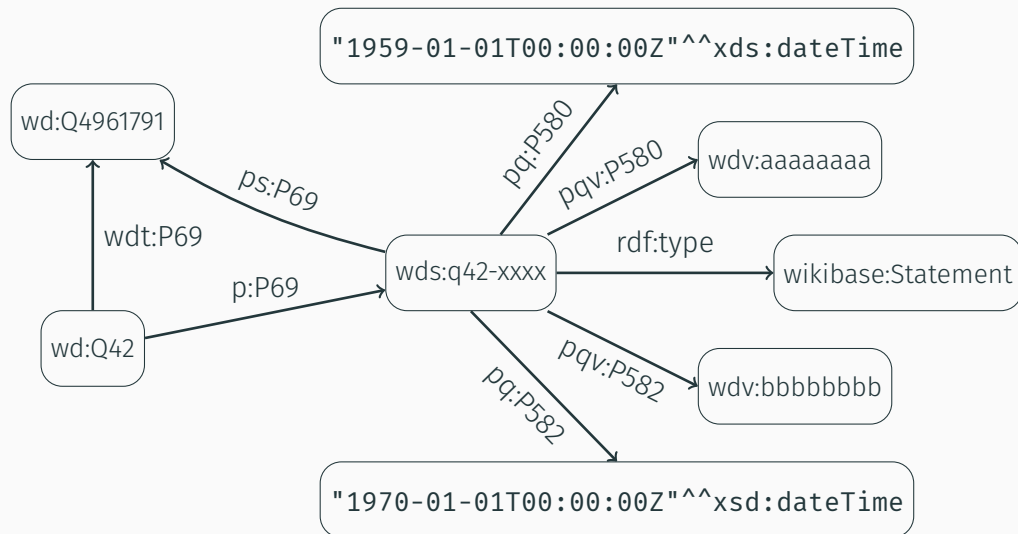
```
wd:Q42 p:P69 wds:q42-xxxx .  
wds:Q42-xxxx rdf:type wikibase:Statement .  
wds:Q42-xxxx ps:P69 wd:Q4961791 .  
wds:Q42-xxxx pq:P580 "1959-01-01T00:00:00Z"^^xsd:dateTime .  
wds:Q42-xxxx pqv:P580 wdv:aaaaaaaa .  
wds:Q42-xxxx pq:P582 "1970-01-01T00:00:00Z"^^xsd:dateTime .  
wds:Q42-xxxx pqv:P582 wdv:bbbbbbbb .  
wd:Q42 wdt:P69 wd:Q4961791.
```



Wikidata als RDF: Reifikation



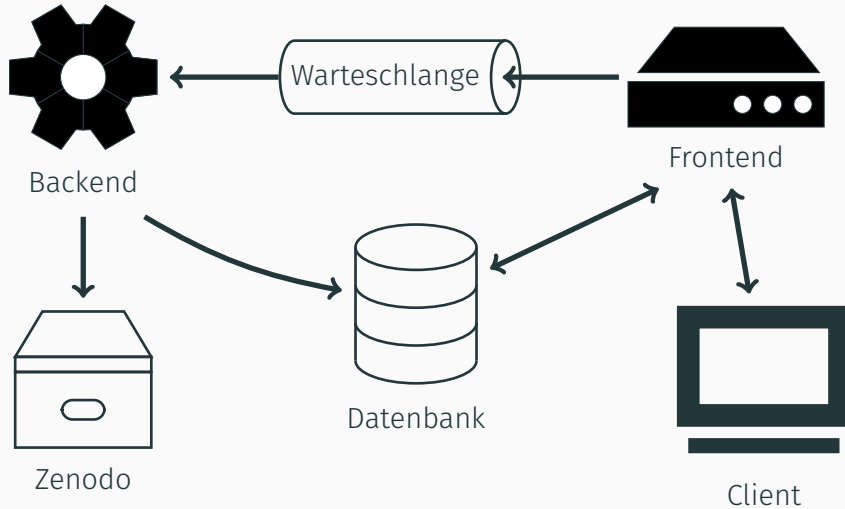
Wikidata als RDF: Reifikation



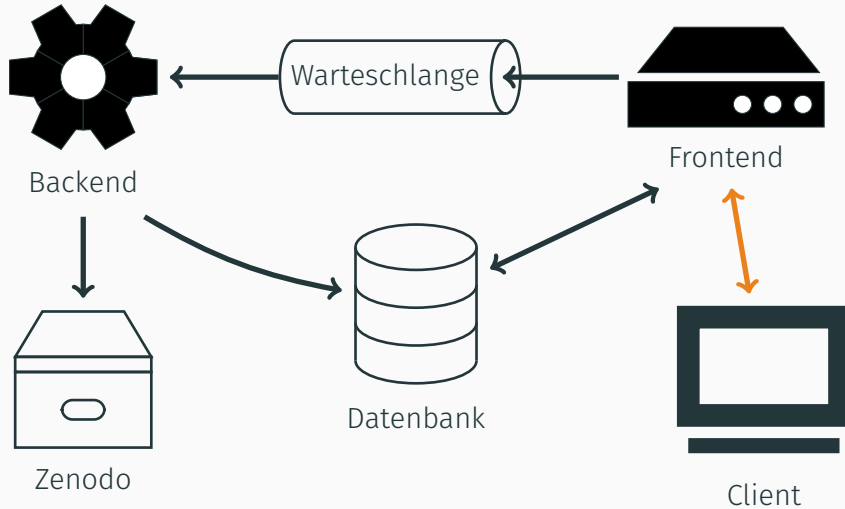
Umsetzung

	Eigener Index	SPARQL basiert	Batchverarbeitung
Geschwindigkeit	schnell	schnell	langsam
Flexibilität	gering	mittel	hoch
Implementierung	aufwändig	mittel	einfach

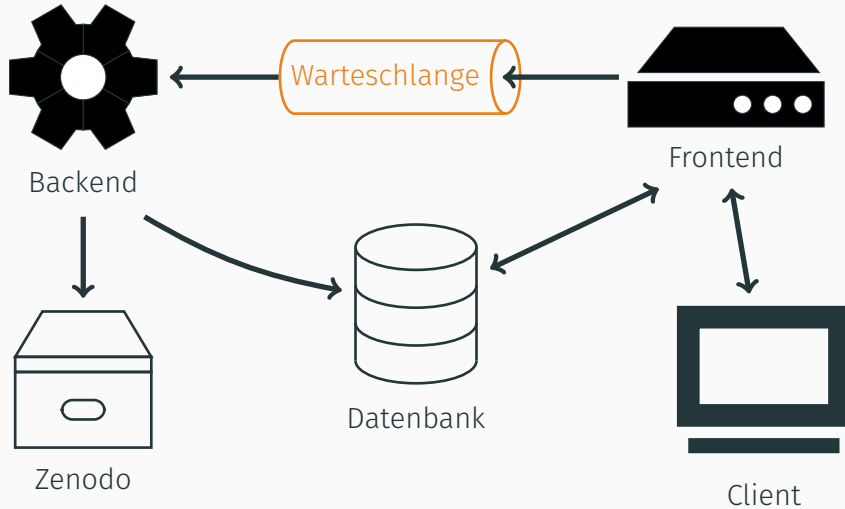
System



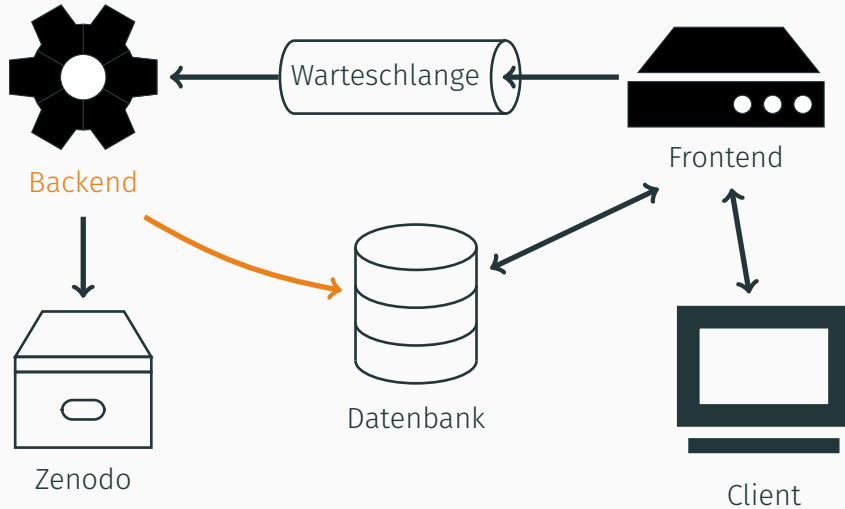
System



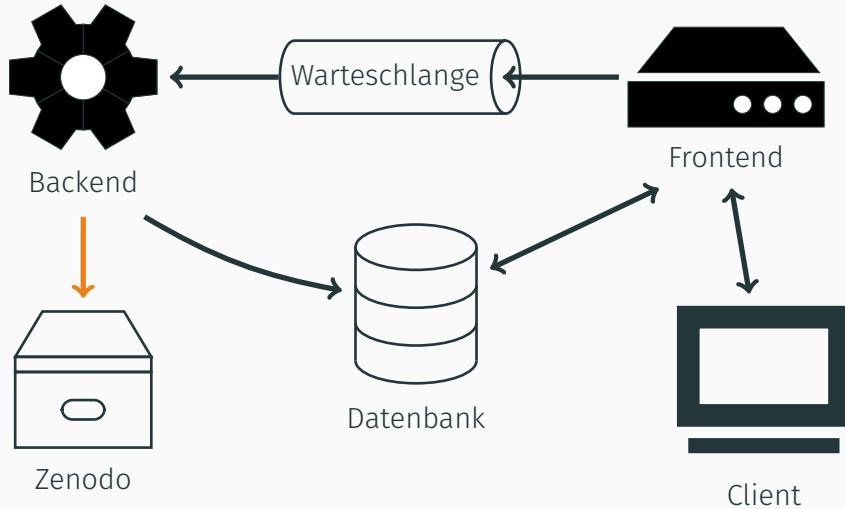
System

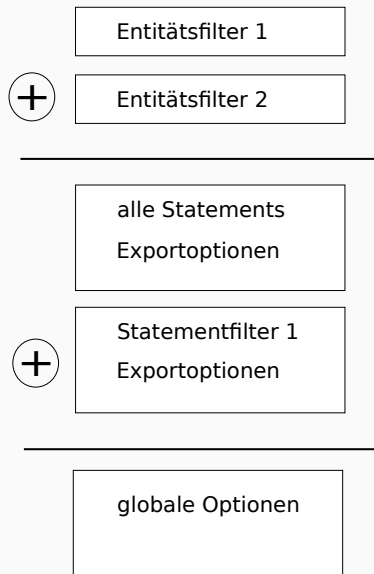


System



System





Evaluation

Filter entities

Choose entities to include in the dump. An entity is included if it matches at least one filter.

No filters added. All entities are included.

+ [Add basic filter](#) + [Add SPARQL query](#)

Filter statements

Choose how statements are exported. These rules are applied to all matched entities.

Default rule

This rule is applied to all matched entities

Parts to export

simple statement ☒

full statement ☐

references ☐

qualifiers ☐

+ [Add custom rule](#)

Additional settings

labels ☒

descriptions ☒

aliases ☒

sitelinks ☒

filter languages ☐

Dump metadata

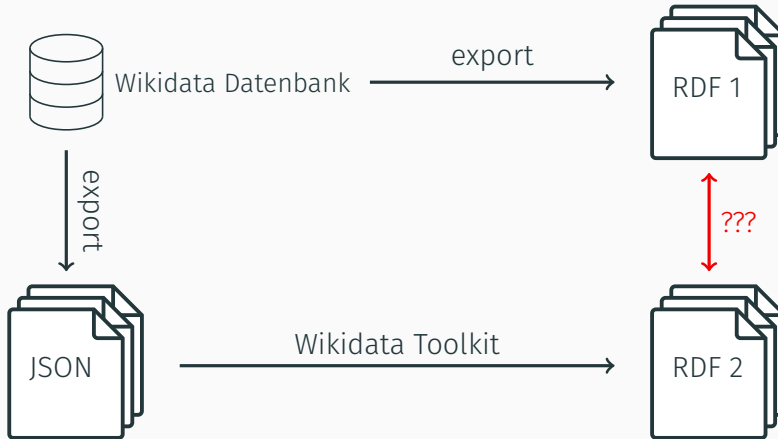
Dump title

Create dump

Screenshot der
Anwendung

Live unter <https://tools.wmflabs.org/wdumps/>

Die Anwendung verwendet Wikidata Toolkit zur Generierung der RDF-Daten



Zwei verschiedene Wege zum Erzeugen von RDF, gibt es Unterschiede?

Viele unkritische Abweichungen:

- andere Schreibweisen
- minimal andere Struktur
- besonders bei komplexen Typen (Zeitpunkte, Orte)

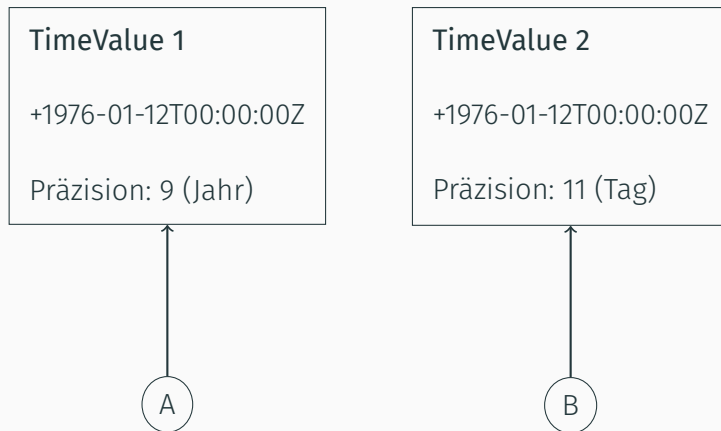
Ein paar tatsächliche Fehler:

- Vertauschung von Koordinaten bei Orten
- falsche Deduplizierung bei Value-Nodes

Einige fehlende Features in Wikidata Toolkit

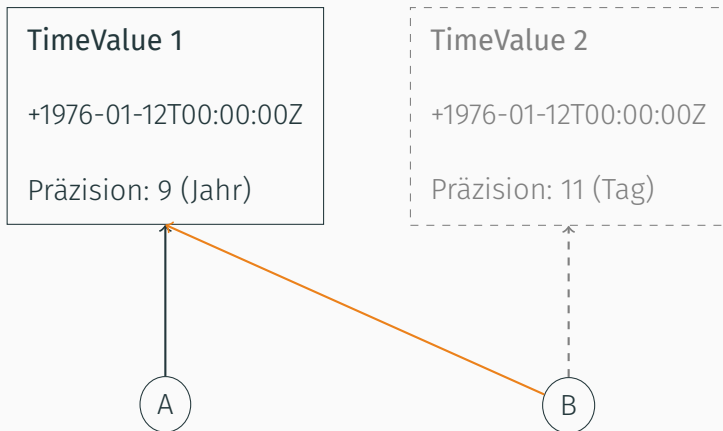
Ergebnis: falsche Deduplizierung bei Werten mit Präzisionsangabe

Für Zeit- und Ortsangaben kann die Präzision gespeichert werden



Ergebnis: falsche Deduplizierung bei Werten mit Präzisionsangabe

Für Zeit- und Ortsangaben kann die Präzision gespeichert werden



Präzision nicht beachtet!

- ① erste Version der Anwendung implementiert
- ② RDF-Export von Wikidata Toolkit verbessert
- ③ viele weitere Features denkbar: Lexeme, Parallelisierung, mehr Filter, ...

- ① erste Version der Anwendung implementiert
- ② RDF-Export von Wikidata Toolkit verbessert
- ③ viele weitere Features denkbar: Lexeme, Parallelisierung, mehr Filter, ...

Fragen?



Metilli, Daniele u. a. (2019). „Populating Narratives Using Wikidata Events: An Initial Experiment“. In: *Digital Libraries: Supporting Open Science - 15th Italian Research Conference on Digital Libraries, IRCDL 2019, Pisa, Italy, January 31 - February 1, 2019, Proceedings*. Hrsg. von Paolo Manghi, Leonardo Candela und Gianmaria Silvello. Bd. 988. Communications in Computer and Information Science. Springer, S. 159–166. ISBN: 978-3-030-11225-7. DOI: [10.1007/978-3-030-11226-4_13](https://doi.org/10.1007/978-3-030-11226-4_13). URL: https://doi.org/10.1007/978-3-030-11226-4_13.



Moreno-Vega, José und Aidan Hogan (2018). „GraFa: Faceted Search & Browsing for the Wikidata Knowledge Graph“. In: *Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, USA, October 8th - to - 12th, 2018*. Hrsg. von Marieke van Erp u. a. Bd. 2180. CEUR Workshop Proceedings. CEUR-WS.org. URL:
<http://ceur-ws.org/Vol-2180/paper-44.pdf>.