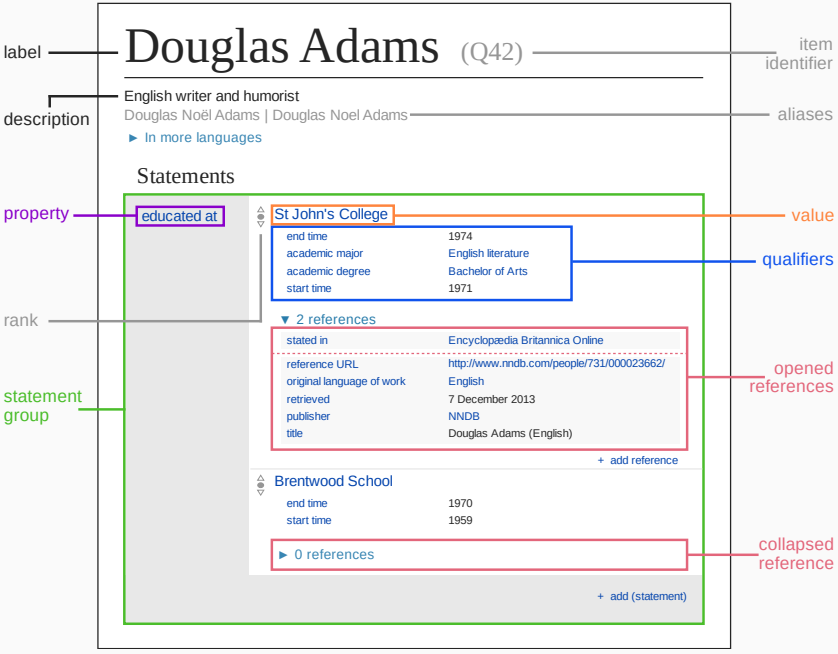
Generierung angepasster RDF-Dumps von Wikidata

Bachelorverteidigung

Benno Fünfstück

17. Dezember 2019

Betreuer: Prof. Dr. Markus Krötzsch
Wissensbasierte Systeme
TU Dresden


label
Douglas Adams
(Q42)
item identifier
description
English writer and humorist
Douglas Noël Adams | Douglas Noel Adams
aliases
► In more languages
Statements
property
educated at
St John's College
value
end time 1974
academic major English literature
academic degree Bachelor of Arts
start time 1971
qualifiers
rank
▼ 2 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)
opened references
statement group
+ add reference
Brentwood School
end time 1970
start time 1959
► 0 references
collapsed reference
+ add (statement)

Wikidata: Größe

70 Millionen Items

800 Millionen Statements

vollständige Datendumps:
mehr als 42GiB komprimiert

Idee: Tool zum Filtern der Daten

Beispiele

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

Anforderungen

- Korrektheit, Vollständigkeit
- Filterung, Archivierung, Suche
- Statistiken, Fortschritt, Parallelverarbeitung
- Nachvollziehrbarkeit, 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?

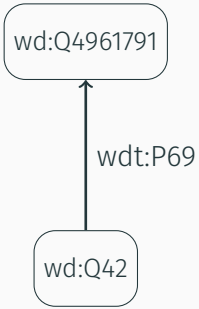
`wd:Q42` ist kurz für `<http://www.wikidata.org/entity/Q42>`

Wikidata als RDF: Reifikation

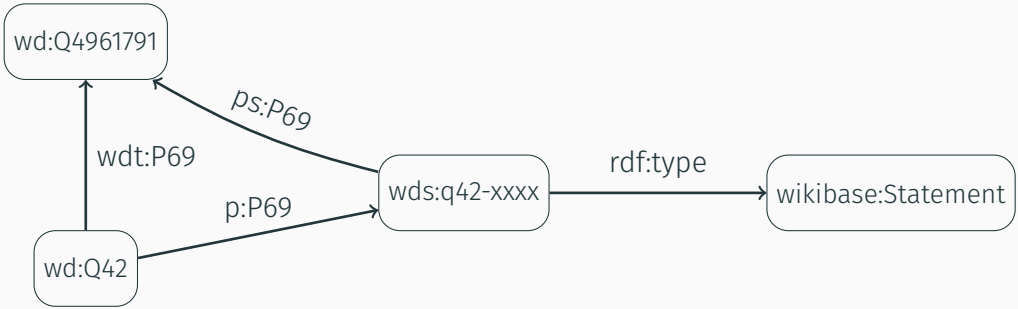
RDF

```
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"^^xds: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.
```

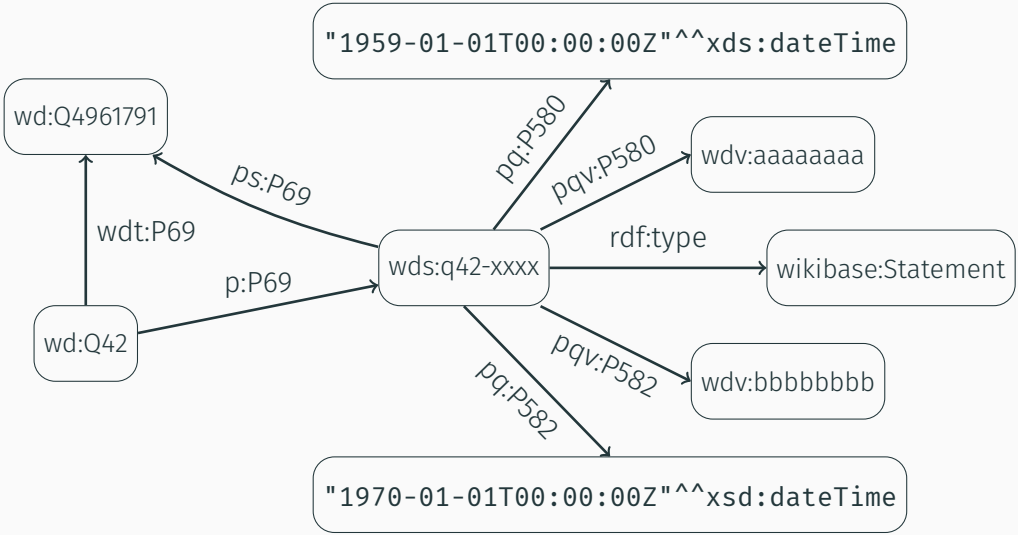
Wikidata als RDF: Reifikation



Wikidata als RDF: Reifikation



Wikidata als RDF: Reifikation

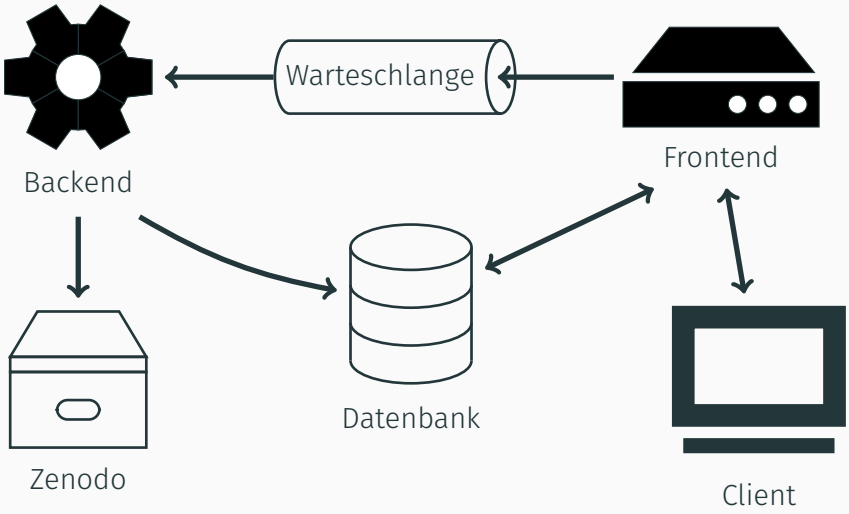


Umsetzung

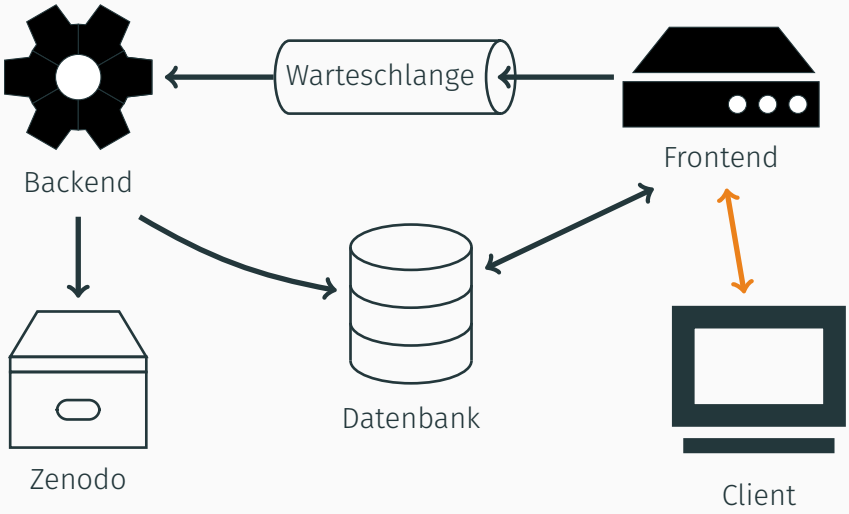
Architekturvarianten

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

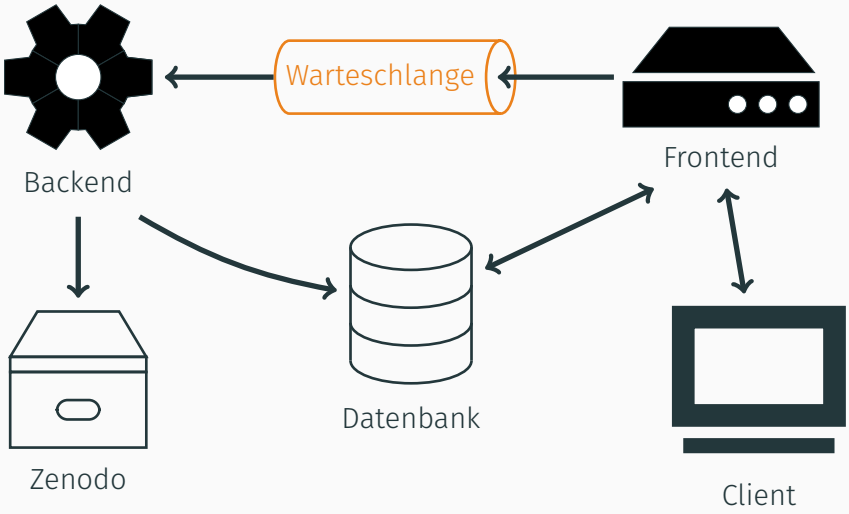
System



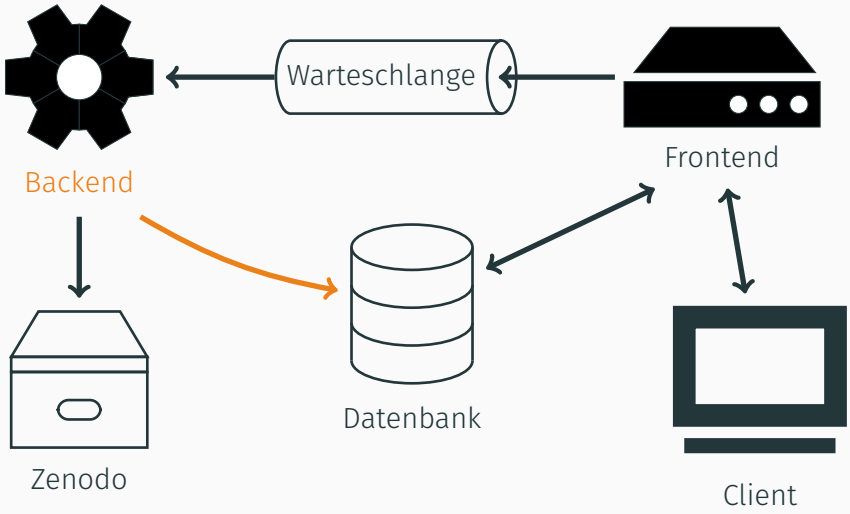
System



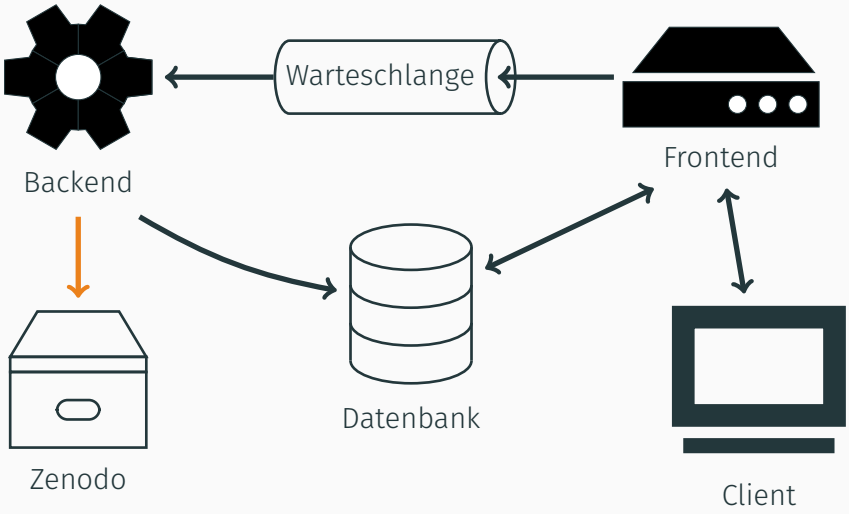
System



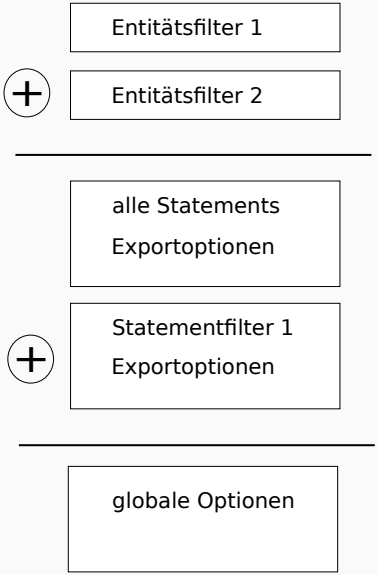
System



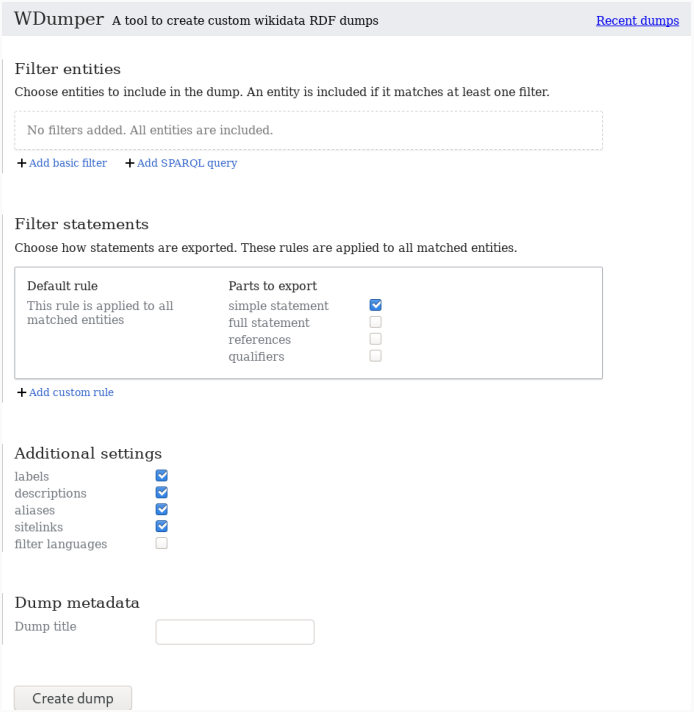
System



UI-Konzept


Entitätsfilter 1
Entitätsfilter 2
alle Statements
Exportoptionen
Statementfilter 1
Exportoptionen
globale Optionen

Evaluation

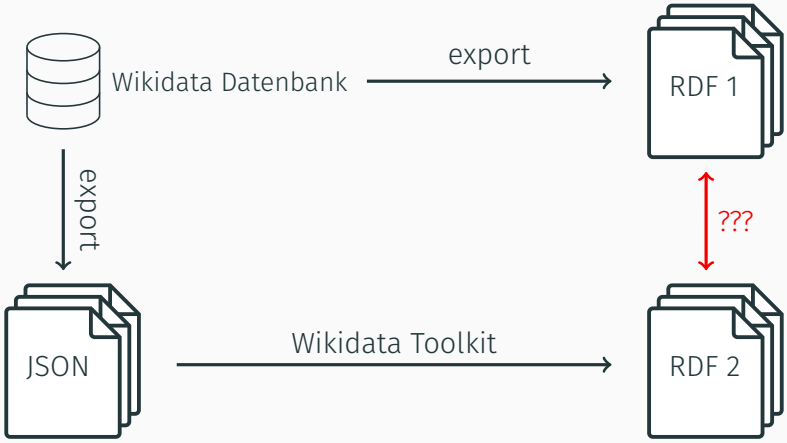


Screenshot der Anwendung

Live unter `https://tools.wmflabs.org/wdumps/`

Korrektheit

Die Anwendung verwendet Wikidata Toolkit zur Generierung der RDF-Daten



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

Ergebnis

Viele unkritische Abweichungen:

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

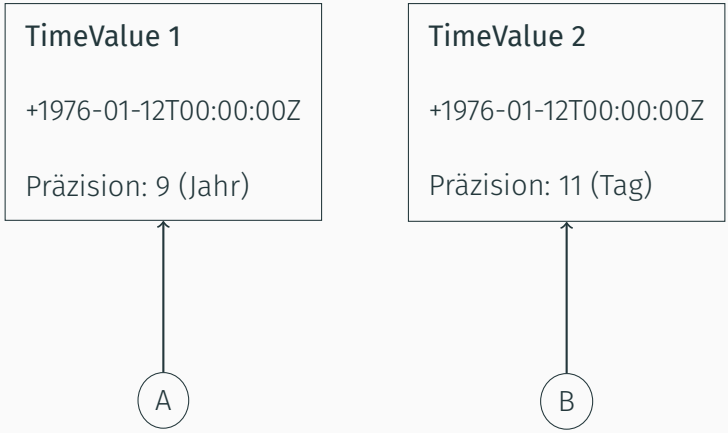
Ein paar tatsächliche Fehler:

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

Einige fehlende Features in Wikidata Toolkit

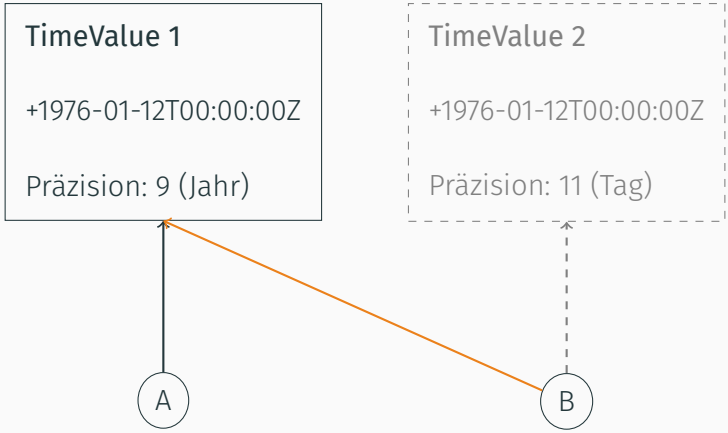
Ergebnis: falsche Deduplizierung bei Werten mit Präzisionsangabe

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!

Fazit

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

Fazit

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. URL: https://doi.org/10.1007/978-3-030-11226-4%5C_13.

Referenzen ii

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.