Generierung angepasster RDF-Dumps von Wikidata

Bachelorverteidigung

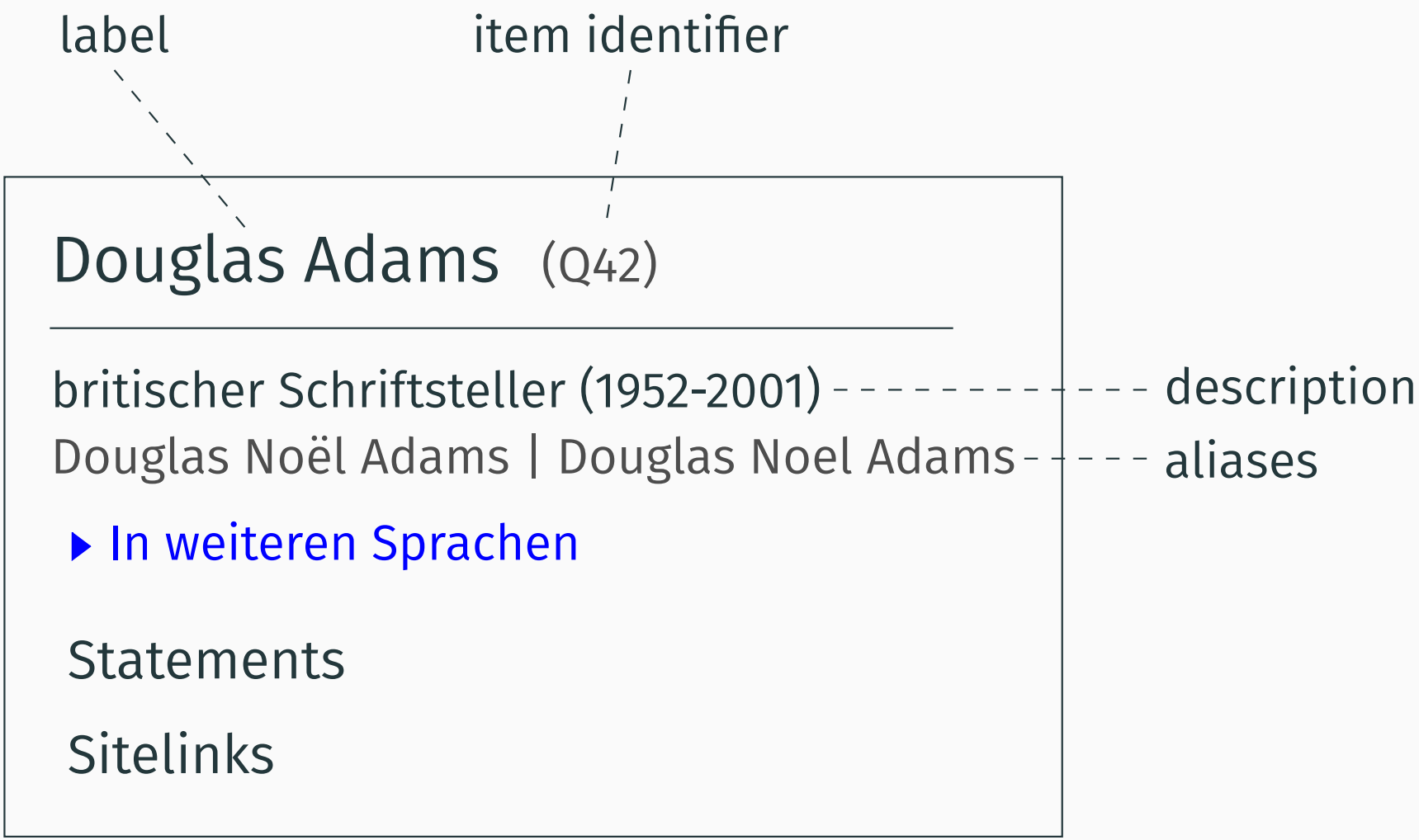
Benno Fünfstück

15. Januar 2020

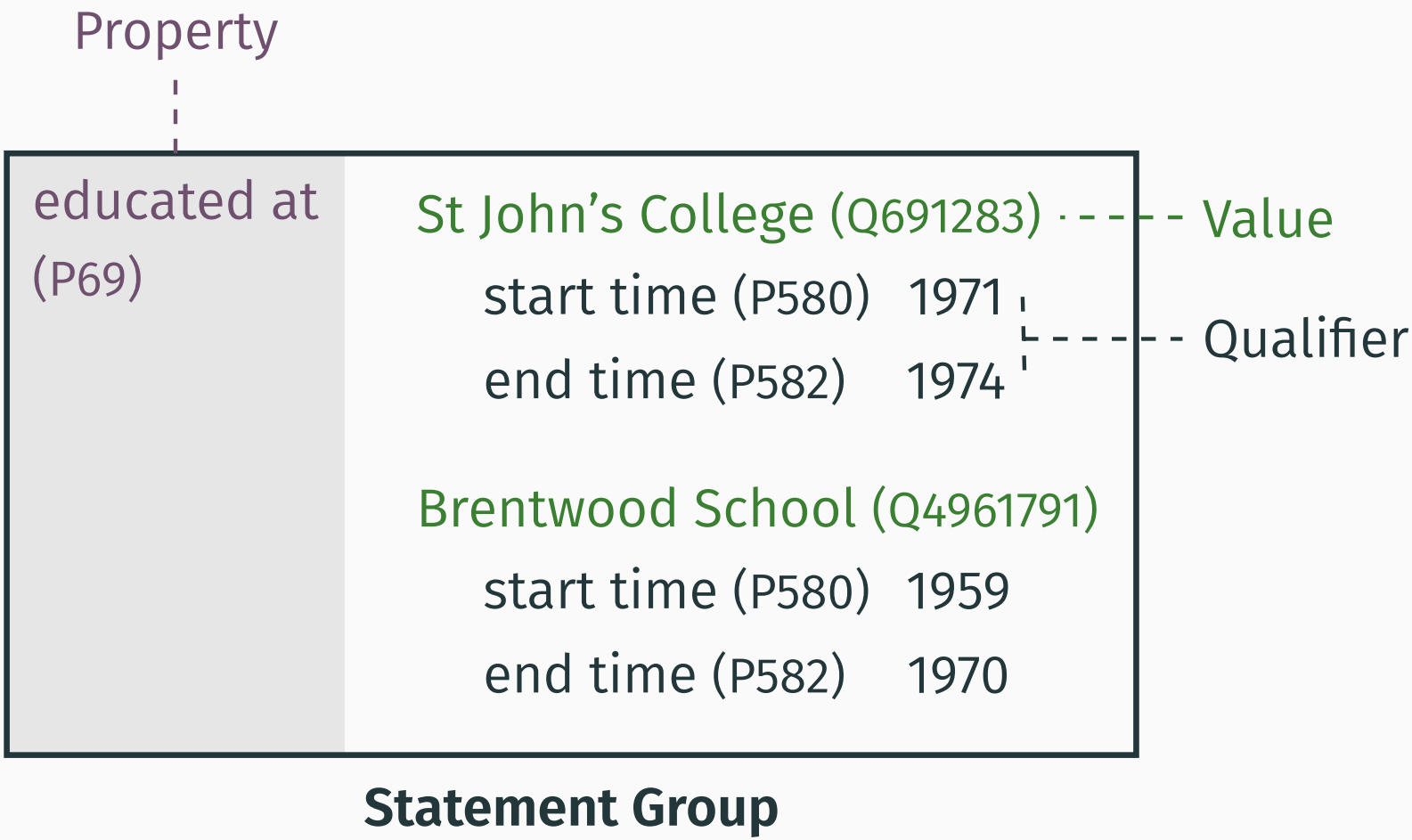
Betreuer: Prof. Dr. Markus Krötzsch
Wissensbasierte Systeme
TU Dresden

„Die freie Wissensdatenbank mit 74.018.298 Datensätzen, die jeder bearbeiten kann.“

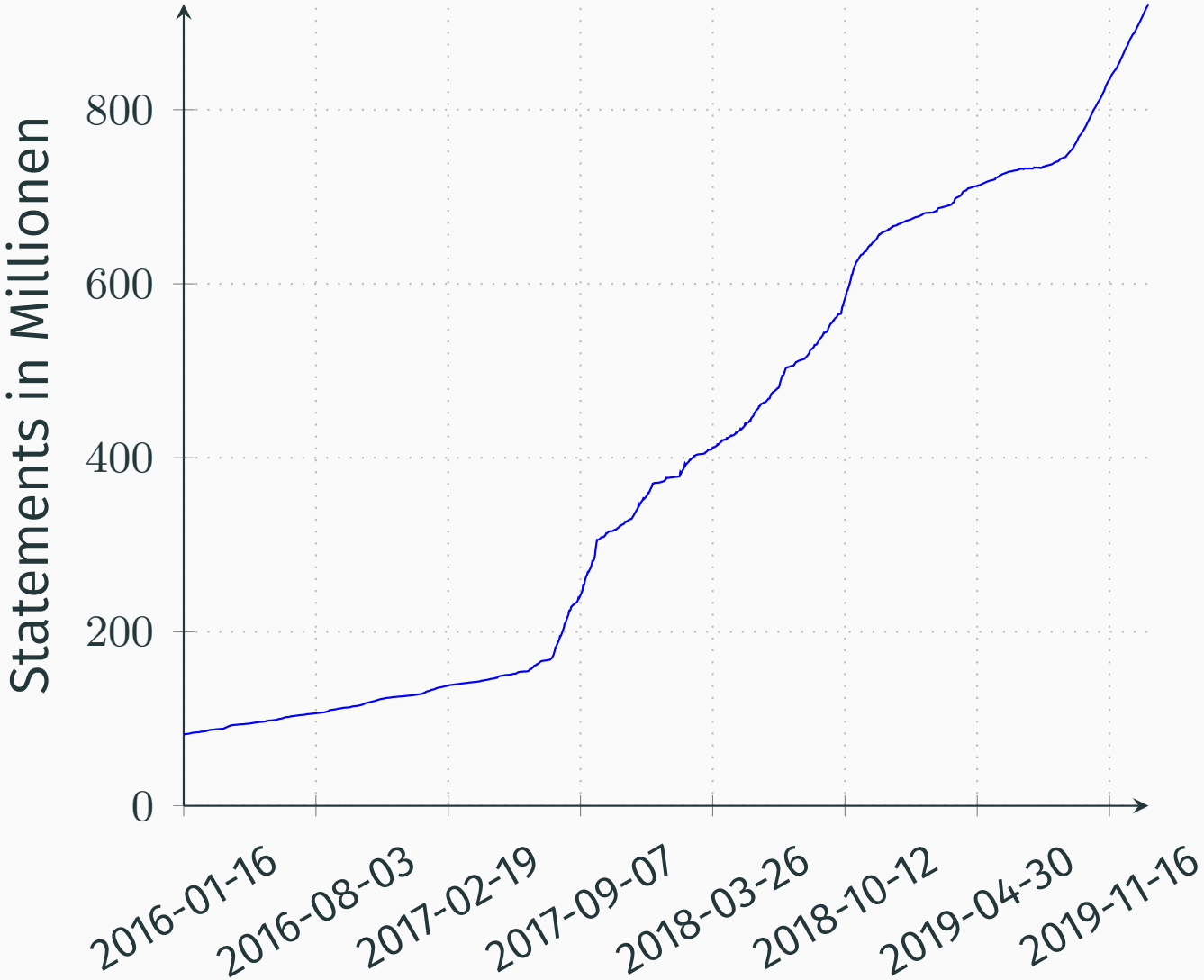
Wikidata: Items



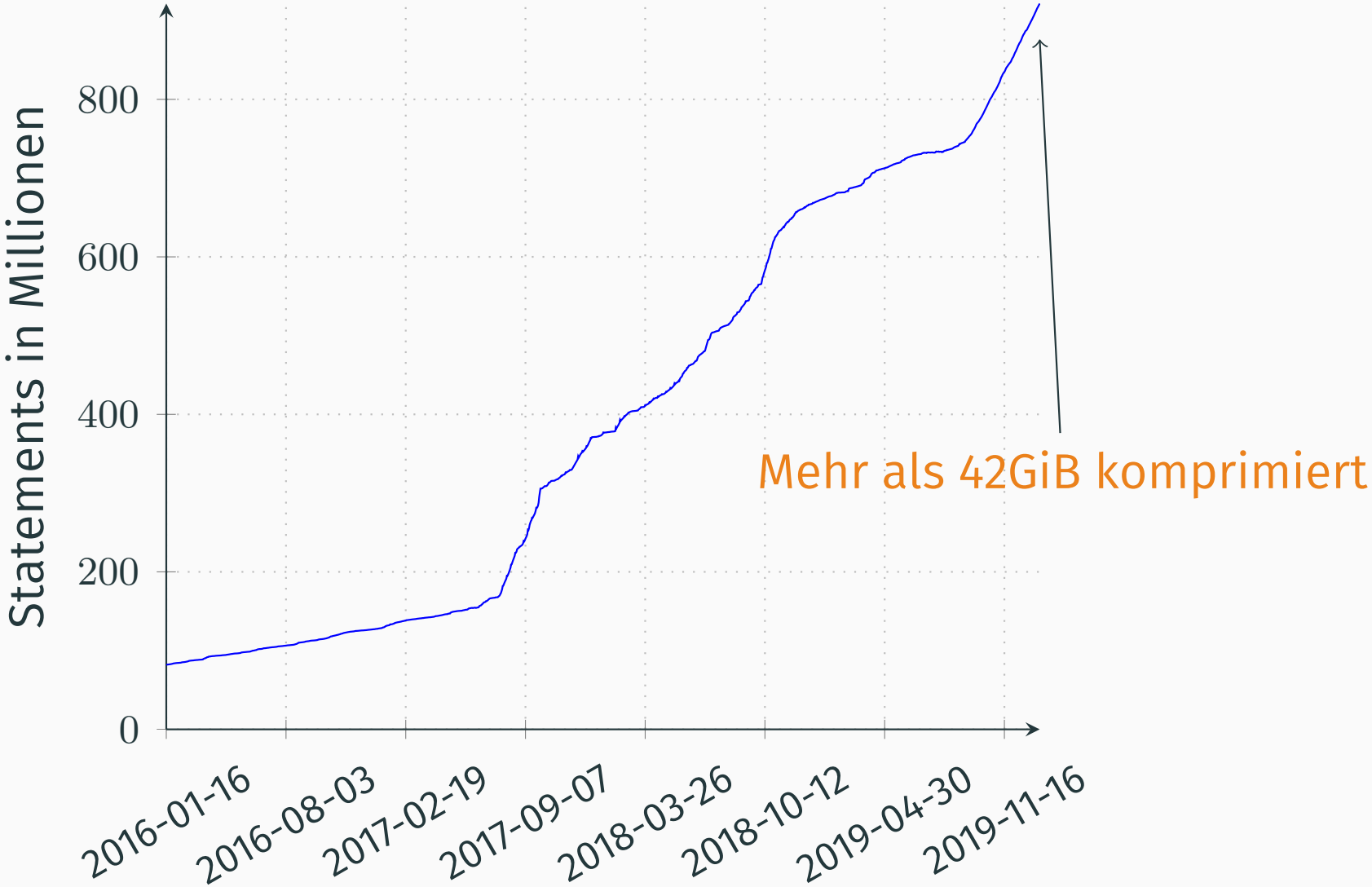
Wikidata: Statements



Wikidata: Größe



Wikidata: Größe

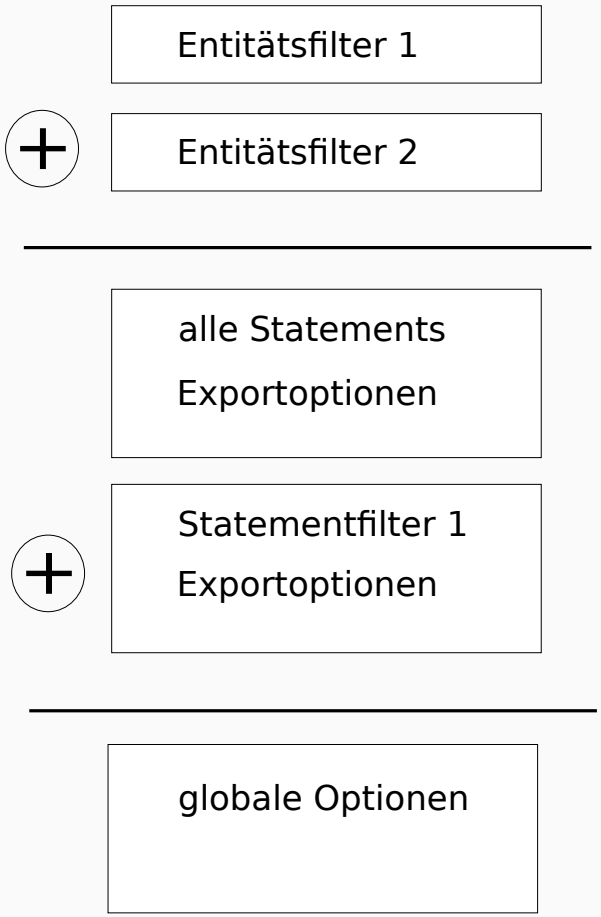


Idee: Tool zum Filtern der Daten

Anforderungen

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

UI-Konzept


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

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 .
```

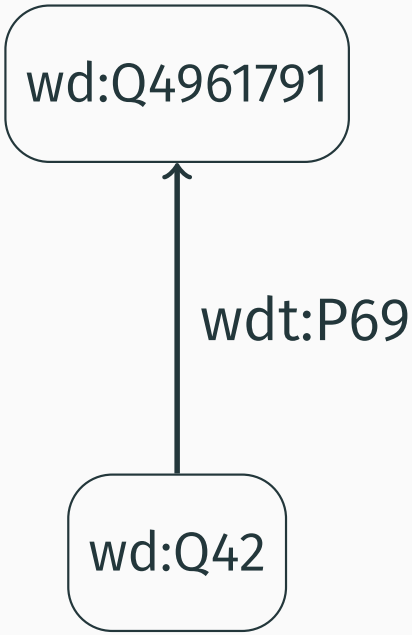
einfache Darstellung: ein Tripel für jedes Statement

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

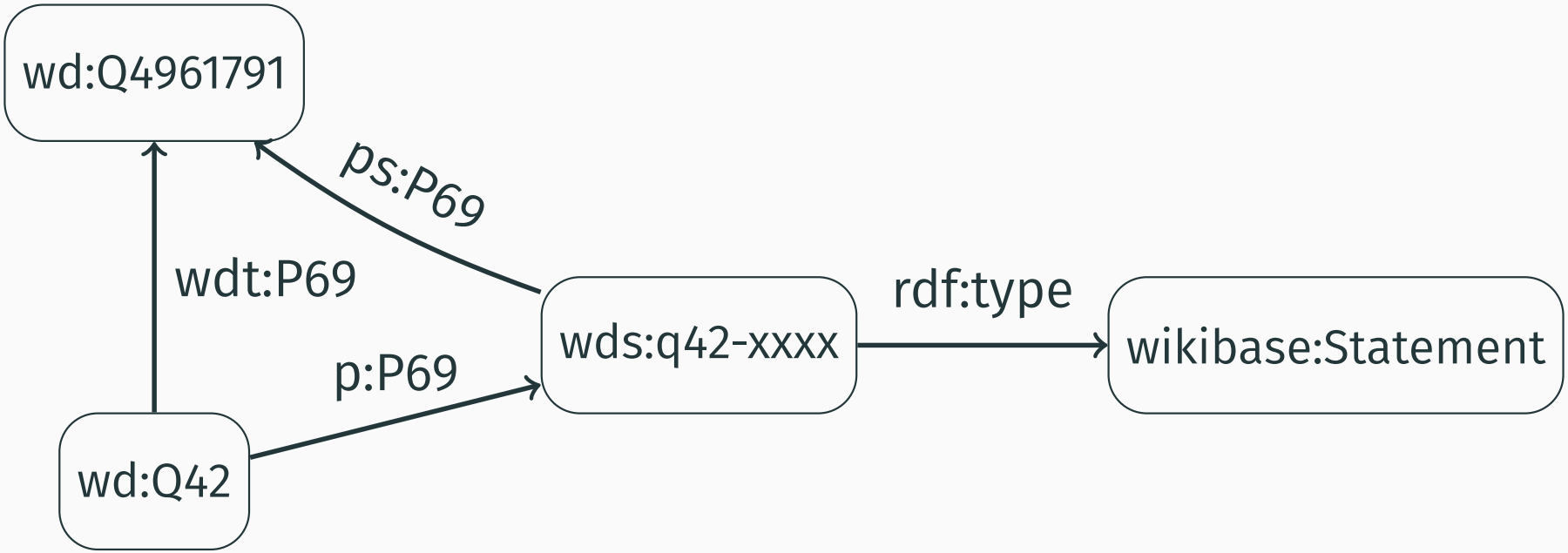
Aber: Ranks, Qualifier, komplexe Werte?

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

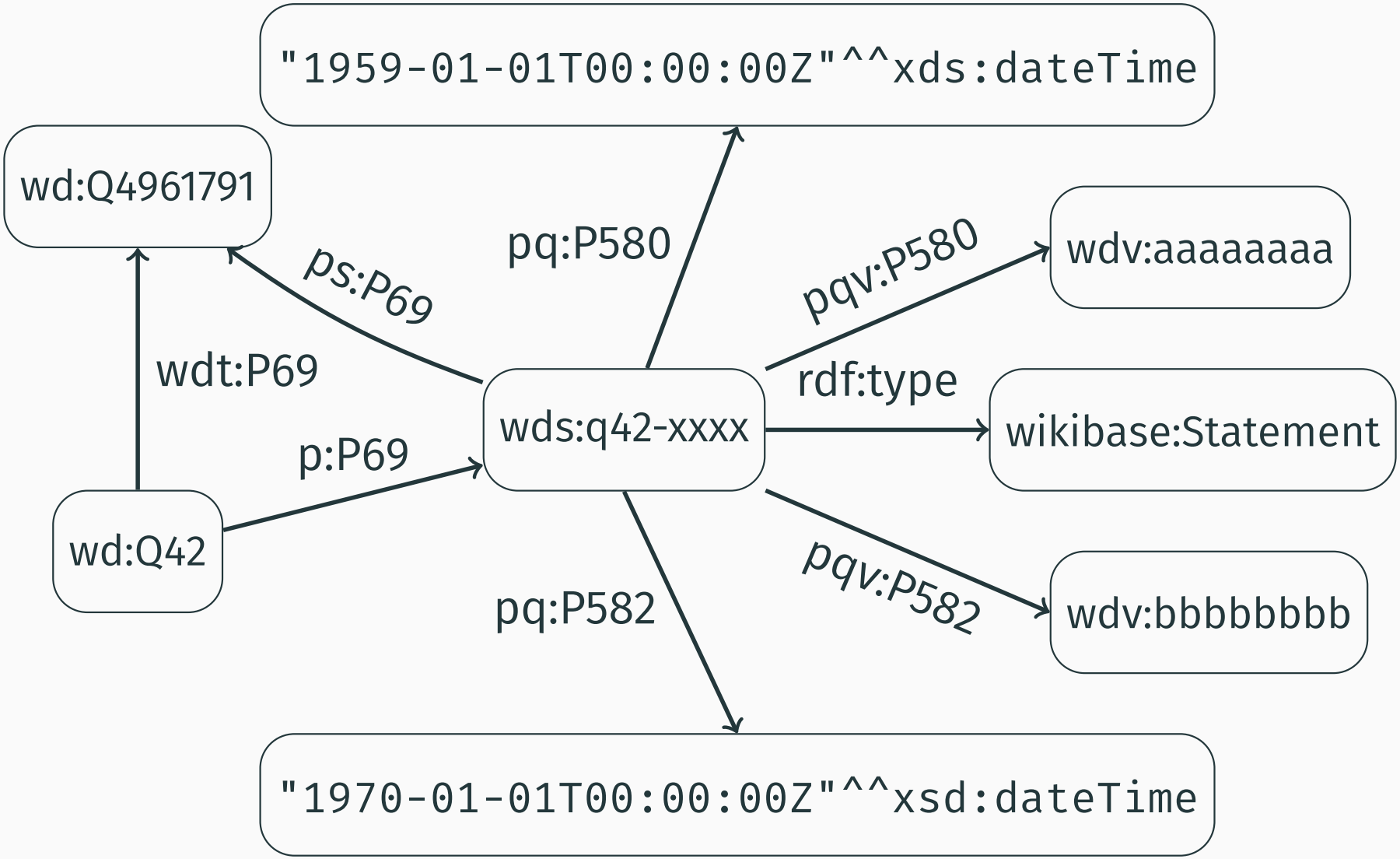
Wikidata als RDF: Reifikation



Wikidata als RDF: Reifikation



Wikidata als RDF: Reifikation

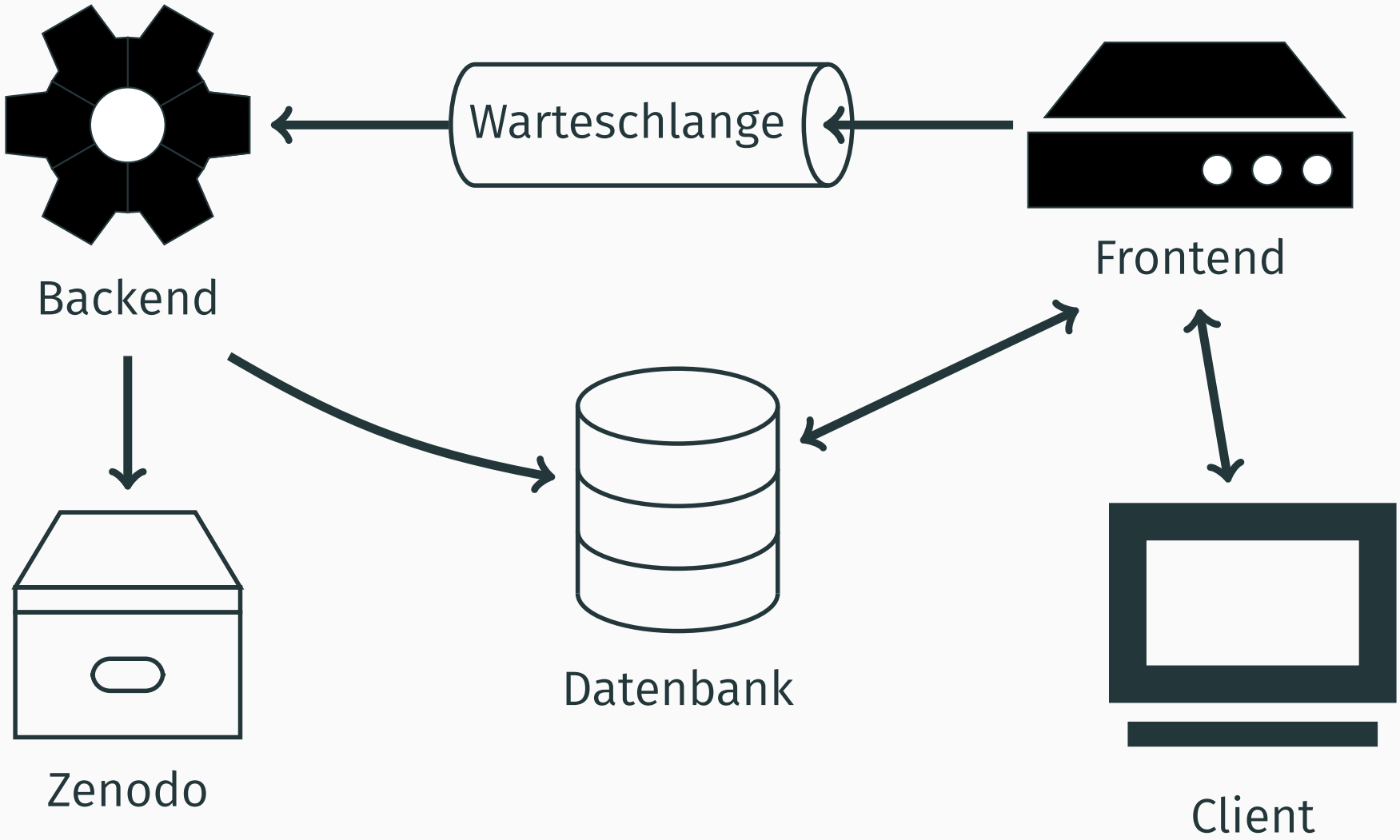


Umsetzung

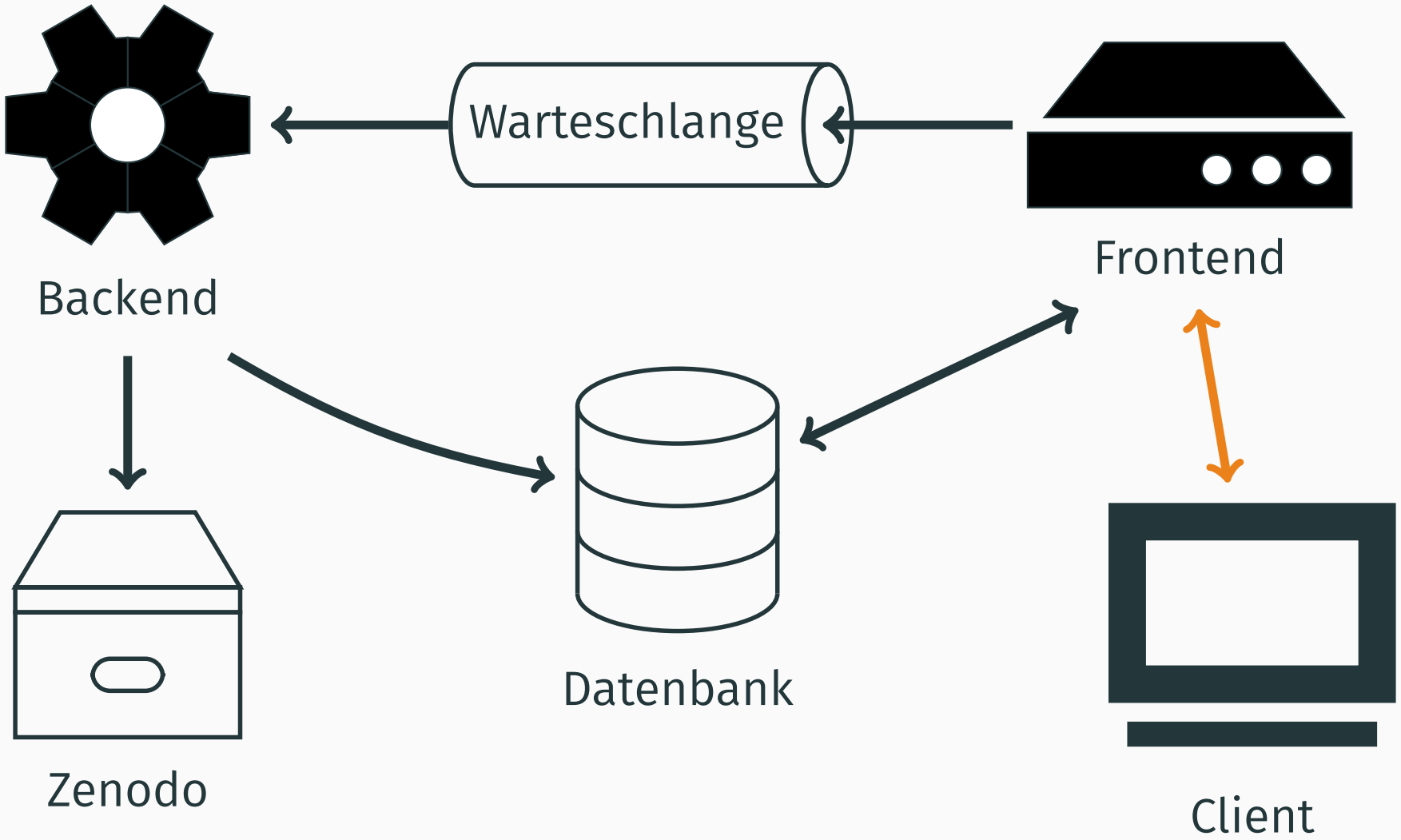
Architekturvarianten

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

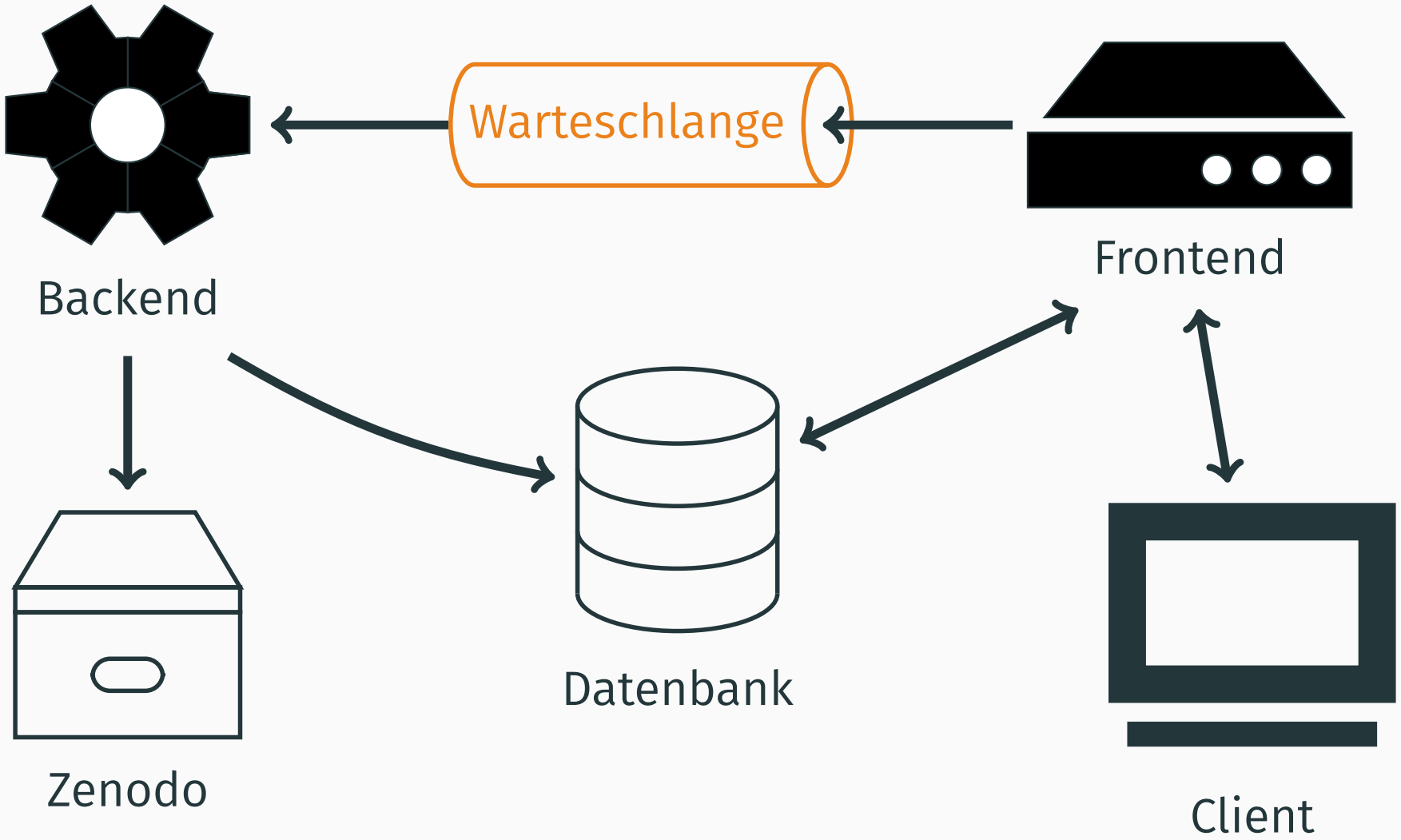
System



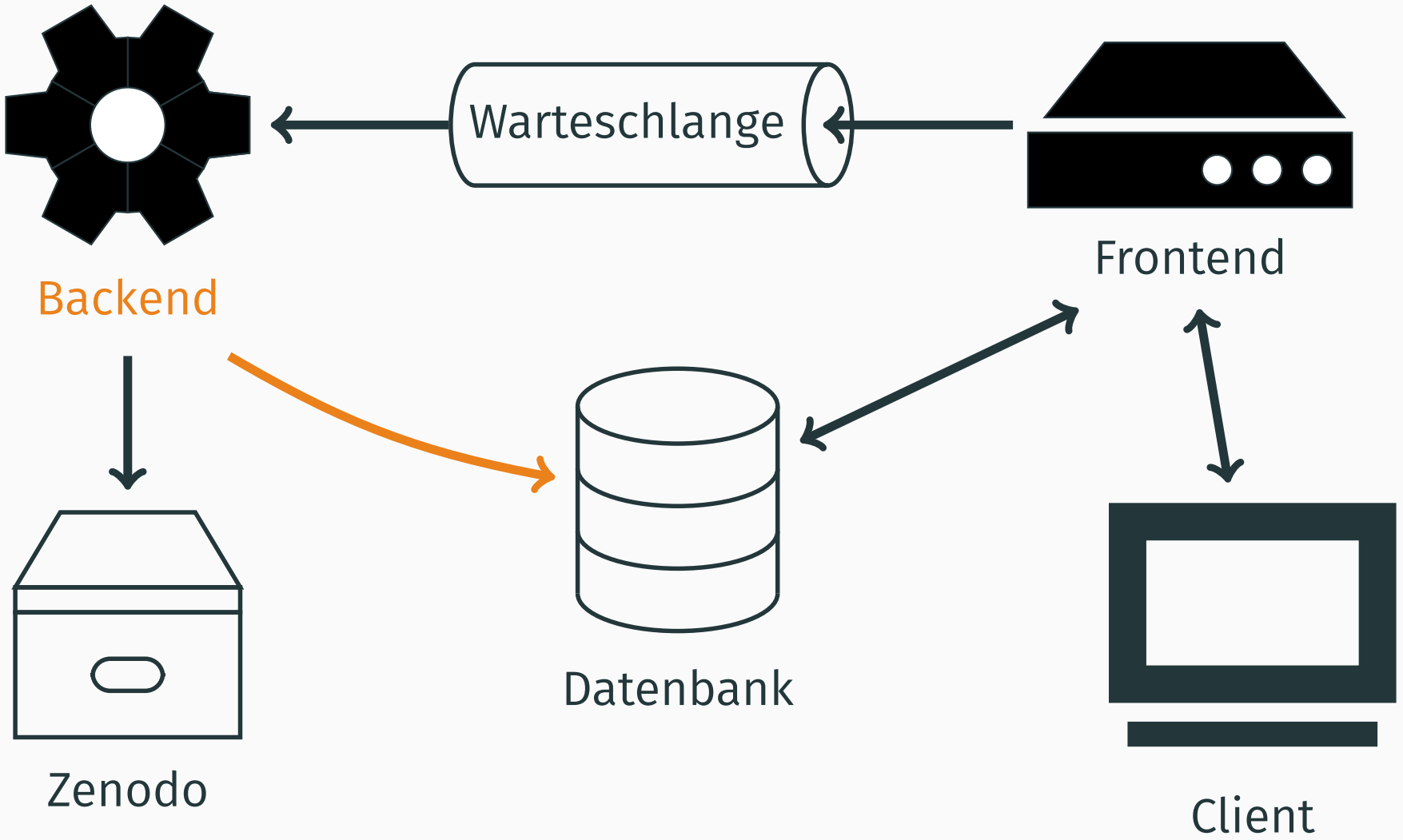
System



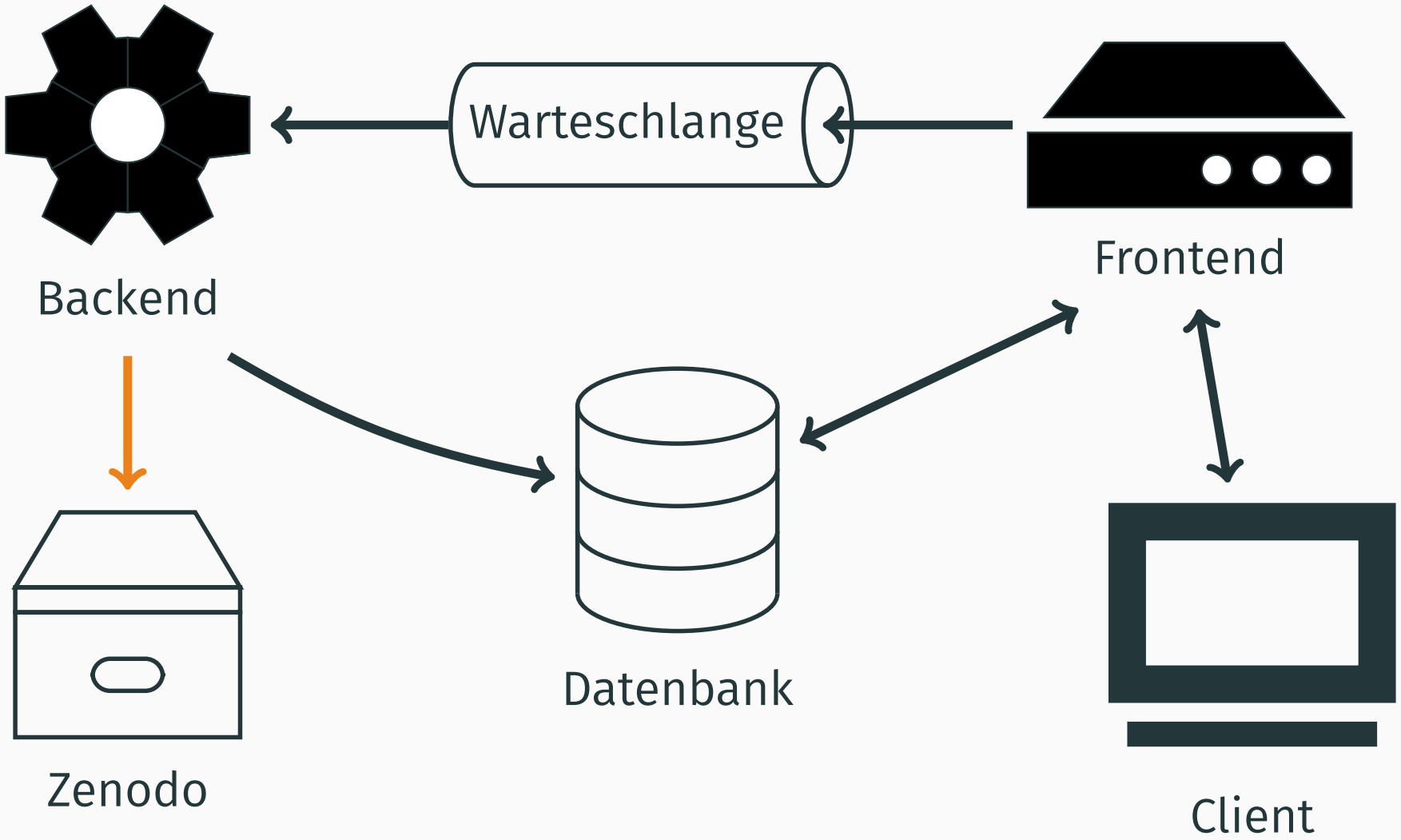
System



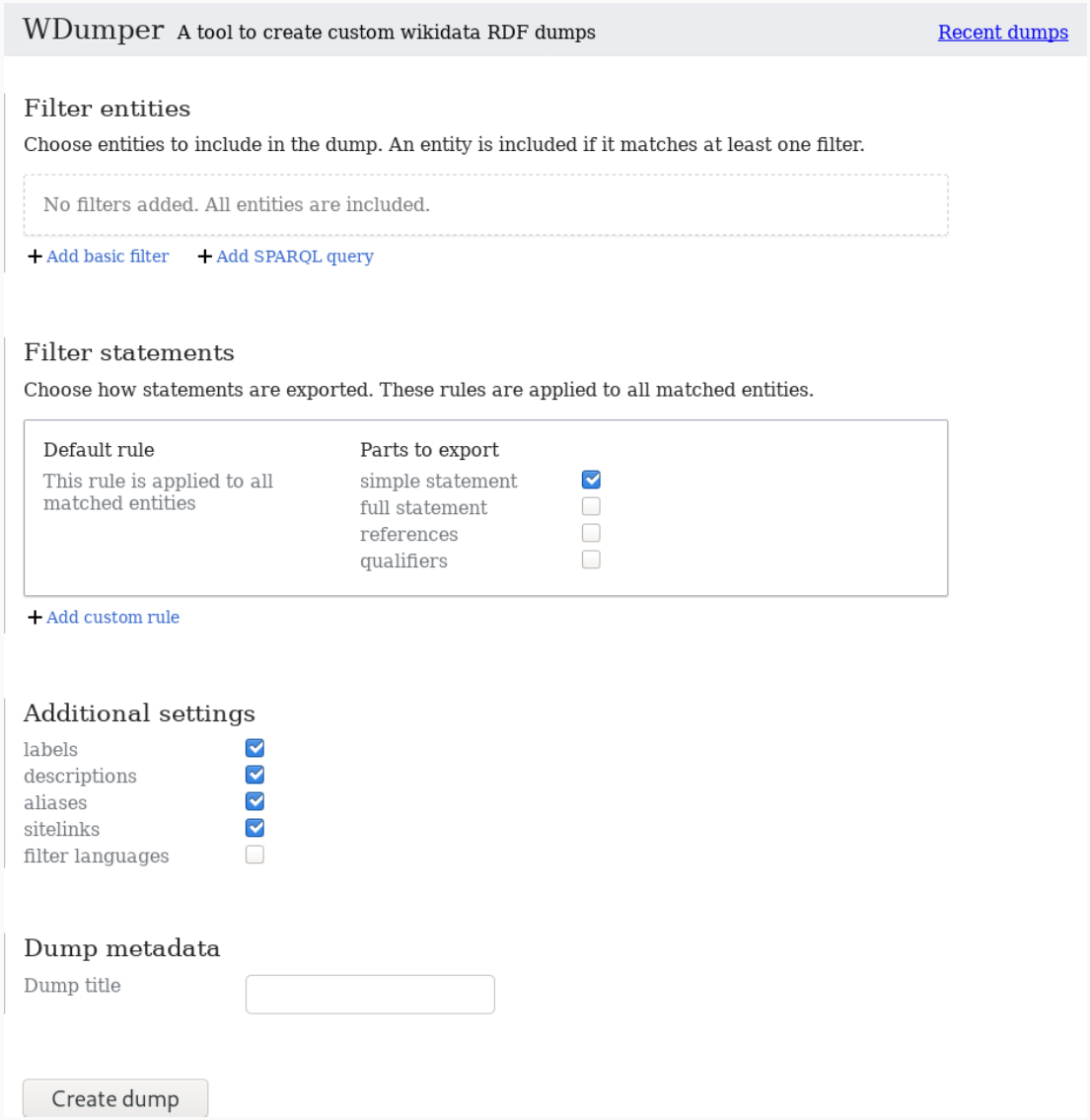
System



System



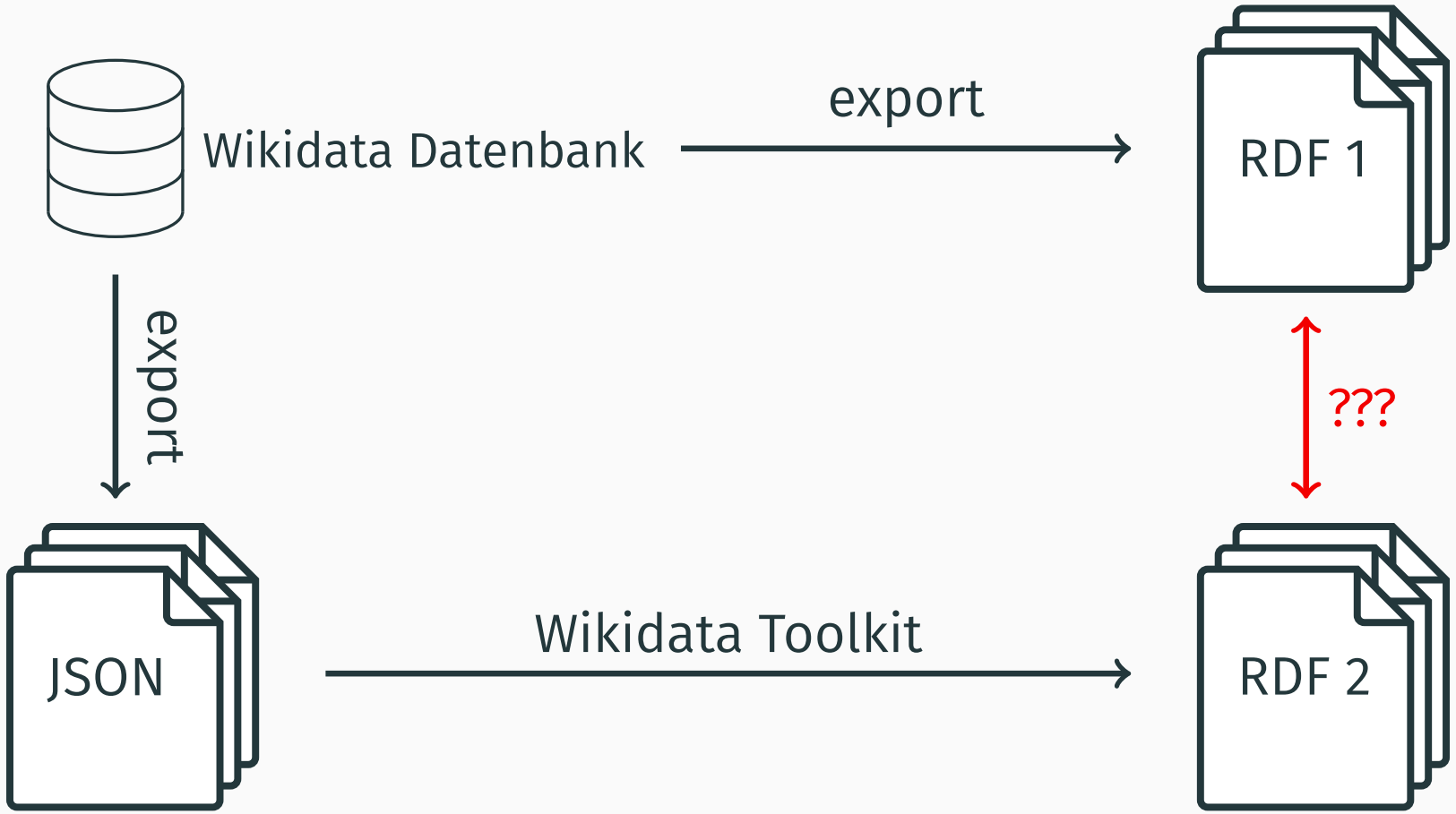
Evaluation



Screenshot der Anwendung

Korrektheit

Die Anwendung verwendet Wikidata Toolkit zur Generierung der RDF-Daten



Zwei verschiedene Wege zum Erzeugen von RDF, gibt es

Ergebnis

Viele unkritische Abweichungen:

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

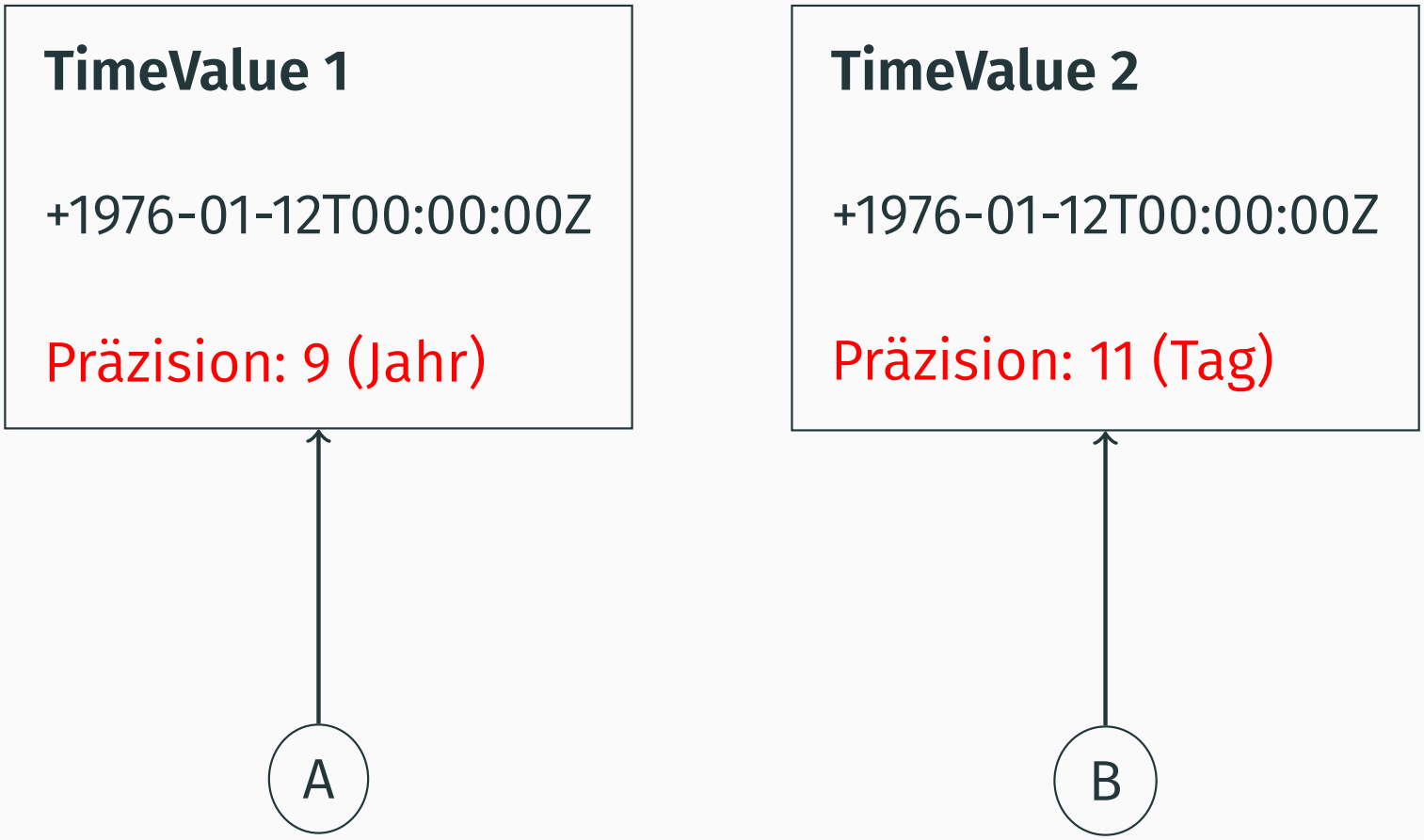
Ein paar tatsächliche Fehler:

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

Einige fehlende Features in Wikidata Toolkit

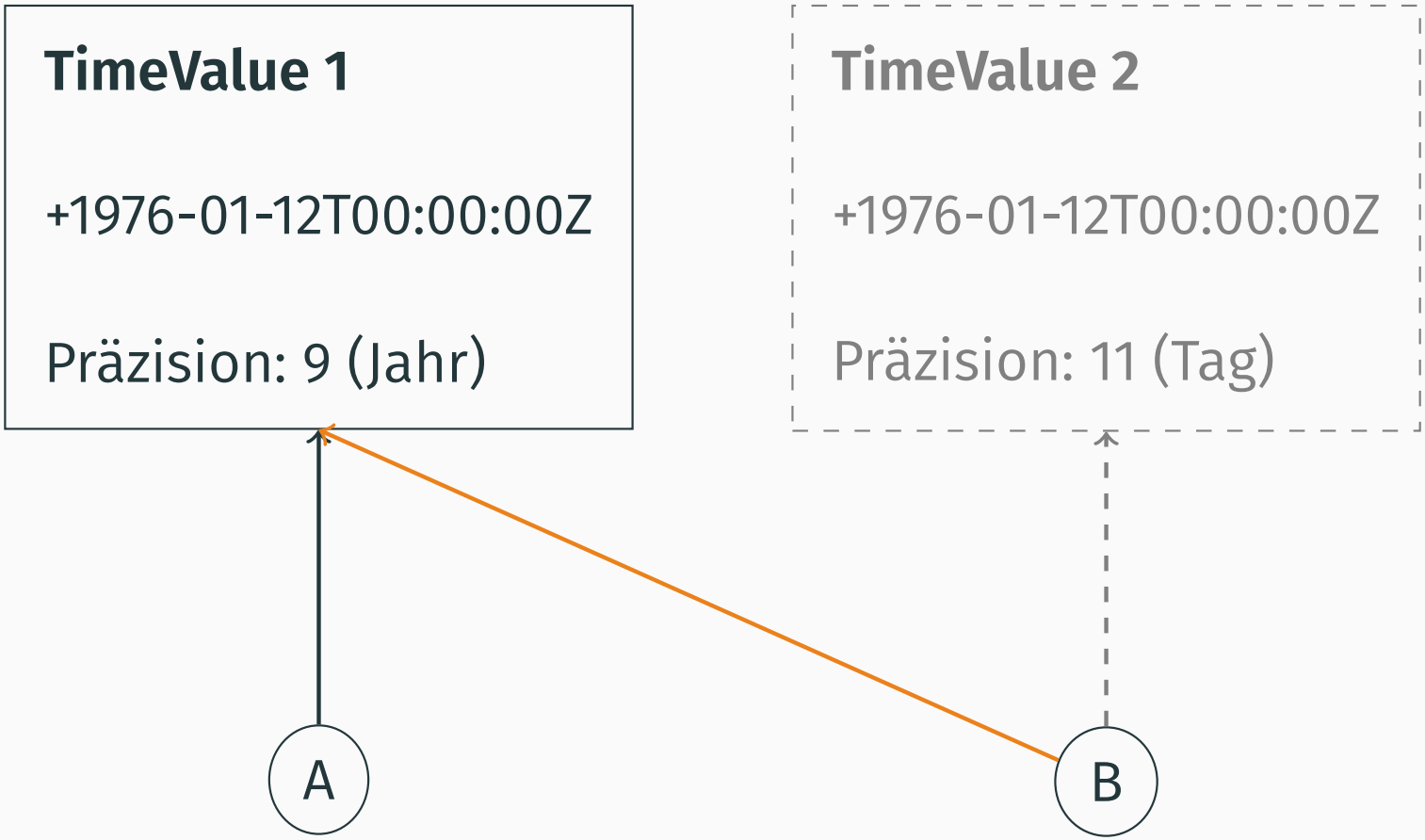
Ergebnis: falsche Deduplizierung bei Werten mit Präzisionsangabe

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?

Referenzen i