

# Generierung angepasster RDF-Dumps von Wikidata

Bachelorverteidigung

---

Benno Fünfstück

15. Januar 2020

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



# WIKIDATA

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

label

item identifier

Douglas Adams (Q42)

---

britischer Schriftsteller (1952-2001) ----- description

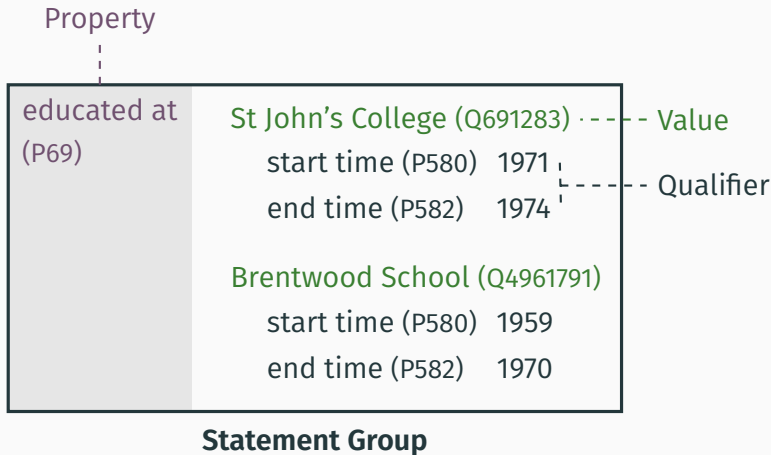
Douglas Noël Adams | Douglas Noel Adams ----- aliases

► [In weiteren Sprachen](#)

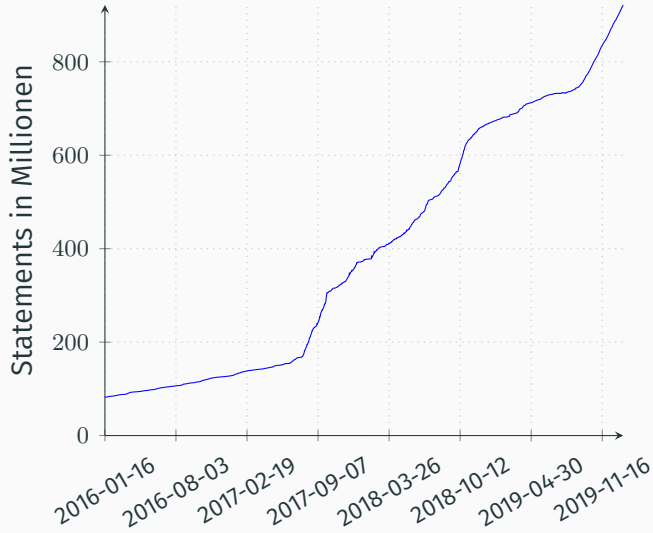
Statements

Sitelinks

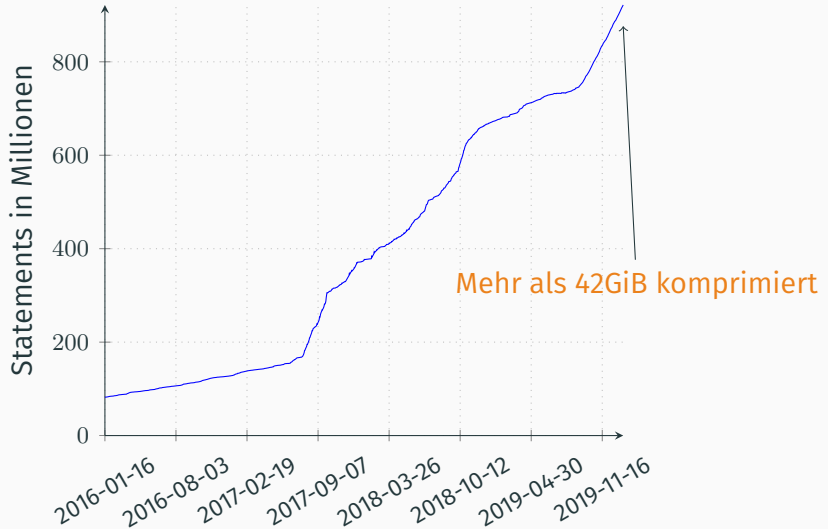
# Wikidata: Statements



# Wikidata: Größe



# Wikidata: Größe



**Idee: Tool zum Filtern der Daten**

---

**Filterung:** Einen Dump aus einer Teilmenge der Daten erzeugen, nach nutzerdefinierten *Kriterien*

Diese Dumps sollen **archiviert** werden und das Archiv **durchsuchbar** sein.



**Filterung:** Einen Dump aus einer Teilmenge der Daten erzeugen, nach nutzerdefinierten *Kriterien*

Diese Dumps sollen **archiviert** werden und das Archiv **durchsuchbar** sein.

Anforderungen an das Interface:

- Prozess ist **nachvollziehbar**: Ursprung, Inhalt des Dumps
- Feedback zu **Fortschritt**
- mehrere Nutzer können parallel Dumps erzeugen (**Parallelverarbeitung**)

**Filterung:** Einen Dump aus einer Teilmenge der Daten erzeugen, nach nutzerdefinierten *Kriterien*

Diese Dumps sollen **archiviert** werden und das Archiv **durchsuchbar** sein.

Anforderungen an das Interface:

- Prozess ist **nachvollziehbar**: Ursprung, Inhalt des Dumps
- Feedback zu **Fortschritt**
- mehrere Nutzer können parallel Dumps erzeugen (**Parallelverarbeitung**)

Format der Dumps: *RDF (Resource Description Framework)*

## Beispiel

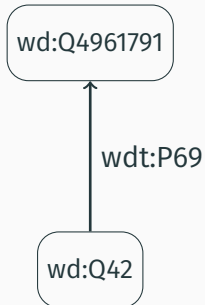
```
wd:Q42 rdfs:label "Douglas Adams"@en .  
wd:Q42 rdfs:label "Douglas Adams"@de .
```

## Einfache Darstellung: ein Tripel für jedes Statement

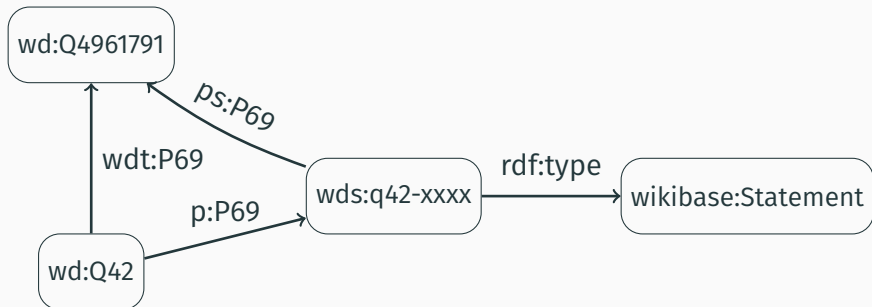
```
wd:Q42 wdt:P69 wd:Q691283 .  
wd:Q42 wdt:P69 wd:Q4961791 .
```

**Aber:** Ranks, Qualifier, komplexe Werte?

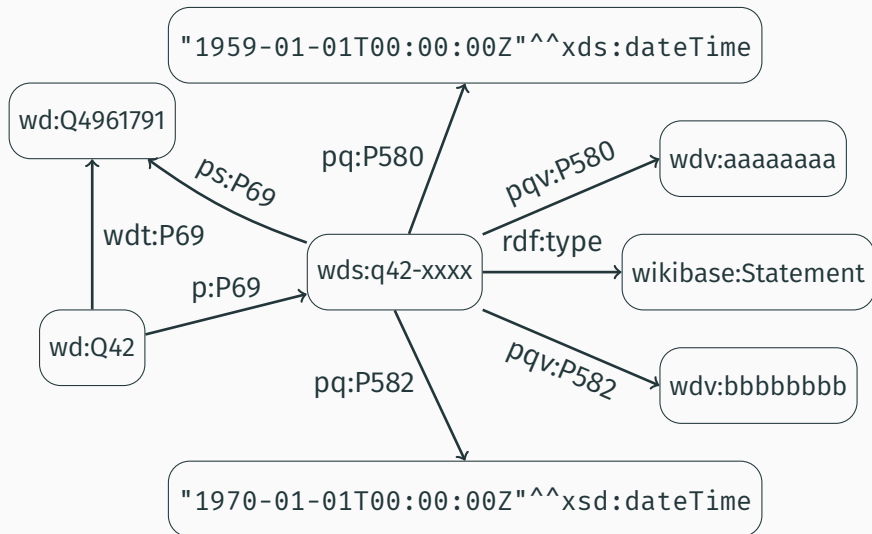
# Wikidata als RDF: Reifikation



# Wikidata als RDF: Reifikation



# Wikidata als RDF: Reifikation





# Umsetzung

---



*Wikidata Query Service* indiziert den RDF-Export von Wikidata und beantwortet SPARQL-Abfragen über diesen Datensatz.

## **Vorteile:**

- Kann einige Abfragen sehr schnell beantworten
- Verwendet aktuelle Daten

## **Nachteile:**

- Große Dumps führen zu Timeout
- Formulierung der Filter in SPARQL umständlich

# Ansatz: Wikidata Toolkit

*Wikidata Toolkit* ist eine Java-Bibliothek zum Verarbeiten der JSON-Exporte von Wikidata.

- Implementiert das RDF Dump Format
- Wird aktiv entwickelt

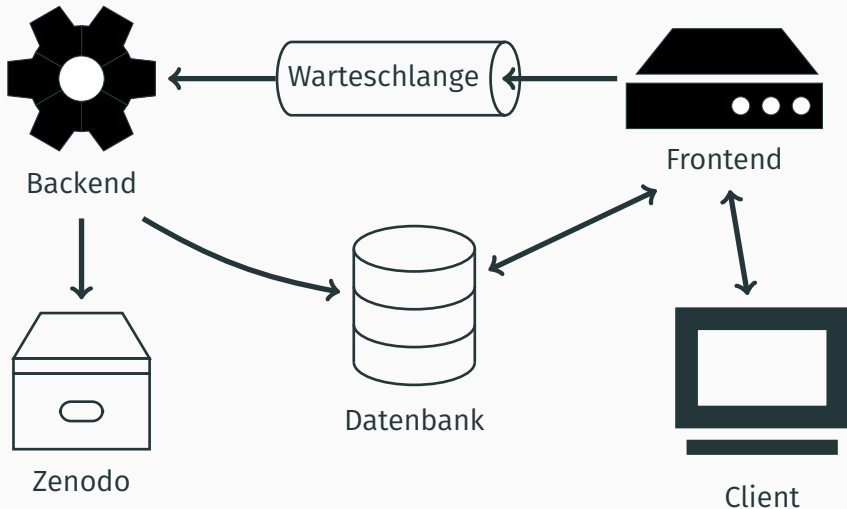
## **Vorteile:**

- Kann auch sehr große Dumps erzeugen
- Filter können in Java geschrieben werden

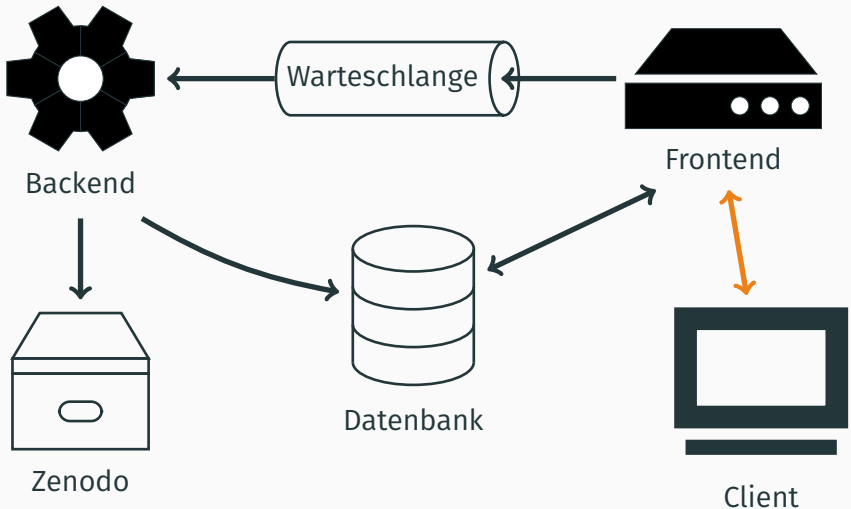
## **Nachteile:**

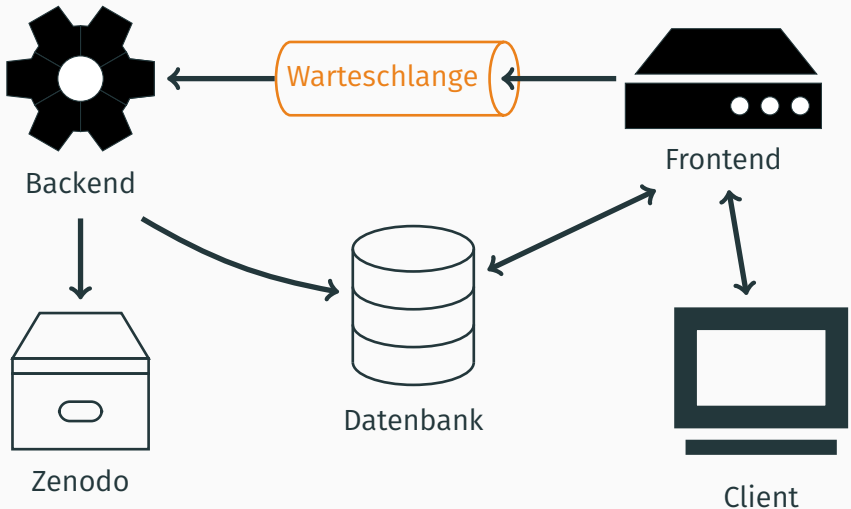
- Langsamer als index-basierte Methoden
- Verwendet nicht den „offiziellen“ RDF-Export von Wikidata

# System

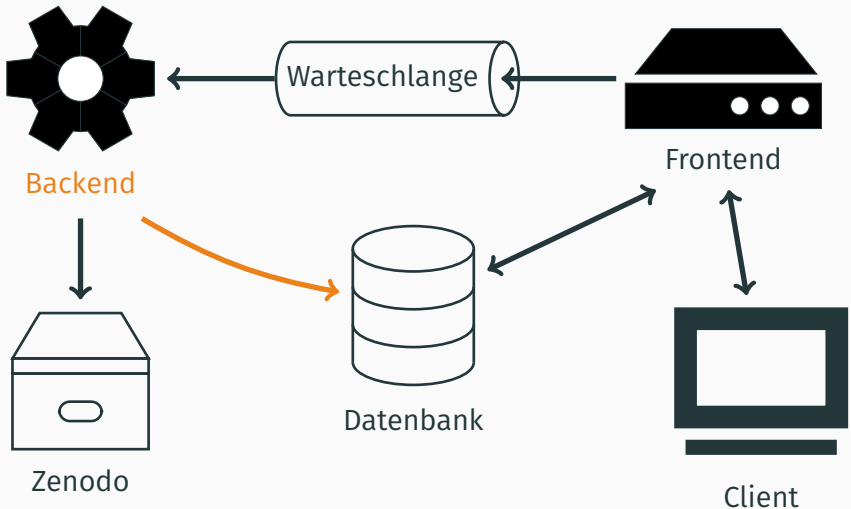


# System

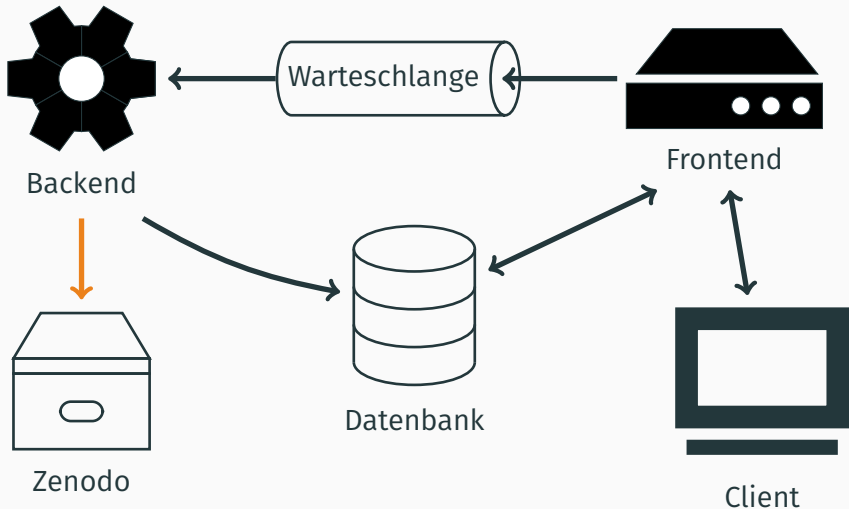




# System



# System



### 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

Deployment auf  
Toolforge

Backend in Java

Frontend in  
Python/TypeScript

MariaDB als  
Datenbank

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



# Evaluation

---

## Filter entities

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

select ☒ items ☐ properties which match all of these conditions:

**property**

instance of

**constraint**

☒ exists ☐ has value

**value**

instance of (P31)

that class of which this subject is a particular example and member (subject typically an individual member with a proper name label); different from P279; using this property as a qualifier is deprecated—use P2868 or P3831 instead

subproperty of (P1647)

all resources related by this property are also related by that

## Filter entities




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

select

☒ items ☐ properties

which match all of these conditions:

×

property	constraint	value	
<input type="text" value="P31"/>	<input type="radio"/> exists <input checked="" type="radio"/> has value	<input type="text" value="Q5"/>	
<input type="text" value="P569"/>	<input checked="" type="radio"/> exists <input type="radio"/> has value		
<input type="text" value="P570"/>	<input checked="" type="radio"/> exists <input type="radio"/> has value		

+ [Add condition](#)

## 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 ☒ only include the following languages:

de x

+

## Dump metadata

Dump title

humans of the past

## Dump 13: humans of the past

processing for 2h:24m



remaining: ~1h:55m

### Zenodo

Sandbox upload will start after dump has finished

Main

Upload to Release

### Timings

Created at 2020-01-15 07:29:14

Processing started at 2020-01-15 07:46:12

Processing finished at None

Processed items 34049735

## Dump 13: humans of the past

↓ [download](#)

### Zenodo

Sandbox [10.5072/zenodo.463471](https://zenodo.org/record/10.5072/zenodo.463471)

Main

Upload to Release

### Timings

Created at	2020-01-15 07:29:14
Processing started at	2020-01-15 07:46:12
Processing finished at	2020-01-15 10:15:14
Processed items	73108759

January 15, 2020

Dataset Open Access

# Wikidata Dump humans of the past

Benno Fünfstück

RDF Dump of wikidata produced with wdump

[Preview](#)

Files (11.4 kB)

Name	Size	
<a href="#">wdump-13.nt.gz</a>	10.9 kB	<a href="#">Download</a>
md5:29c2d7455ccba6eb19e8be68d871e2d0		
<a href="#">wdumper-spec.json</a>	502 Bytes	<a href="#">Preview</a> <a href="#">Download</a>
md5:e13a9f4b4423ff273f11dbb63fa3f690		

Beta

Citations 0

0

views

0

downloads

[See more details...](#)

Indexed in

OpenAIRE

Publication date:

January 15, 2020

DOI:

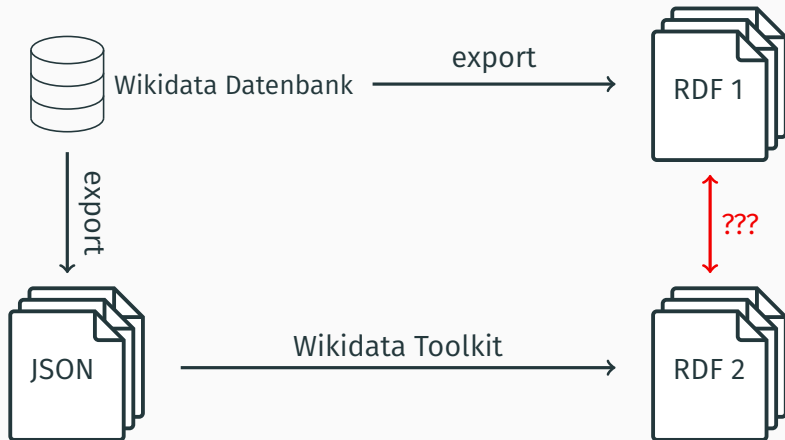
[DOI: 10.5072/zenodo.463471](#)

License (for files):

[Creative Commons Zero v1.0 Universal](#)



Die Anwendung verwendet Wikidata Toolkit zur Generierung der RDF-Daten



Es wurden eine Reihe von Fehlern im RDF-Export von Wikidata Toolkit gefunden:

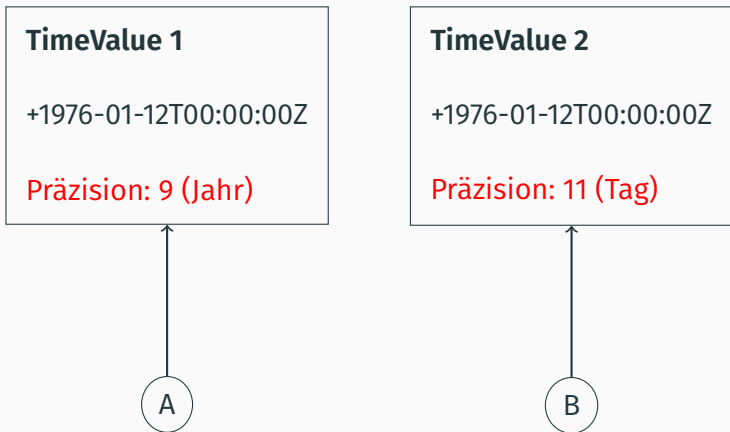
- andere Schreibweisen
- minimal andere Struktur
- besonders bei komplexen Typen (Zeitpunkte, Orte)

Beispiele für weniger offensichtliche Fehler sind:

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

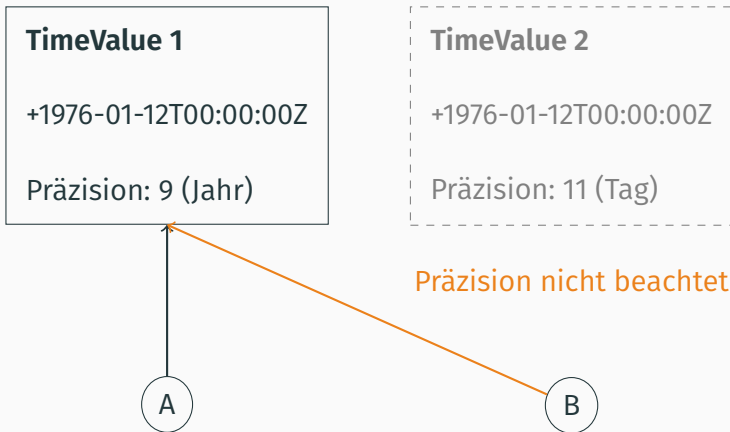
# Ergebnis: falsche Deduplizierung bei Werten mit Präzisionsangabe

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



## Ergebnis:

- 1 Erfolgreiche Implementierung eines funktionsfähigen Prototyps
- 2 RDF-Export von Wikidata Toolkit verbessert
- 3 Gesamter Quellcode ist offen verfügbar:  
<https://github.com/bennofs/wdumper>

## Ausblick:

- Weitere Kriterien zum Filtern: SPARQL, zufällige Auswahl, Anzahl an Sitelinks, ...
- Verbesserung des UI: Vorschau, vorgeschlagene Dumps, bessere Suche