

Experiences of Using WDumper to Create Topical Subsets from Wikidata

Seyed Amir Hosseini Beghaeiraveri

Alasdair Gray

Fiona McNeill

2nd International Workshop On Knowledge Graph Construction

ESWC - 2021

Outline

- Why Subsets?
- Topical Subsets
- WDumper, Easy and Accurate tool
 - Strengths and Weaknesses
- Research Challenges
- Future Works

Why Subsets?



Wikidata 2014
3.5 GB



Wikidata 2021
100 GB



Huge Size of KGs

Timed-Out Queries

Why Subsets?



Reducing the Overall Costs

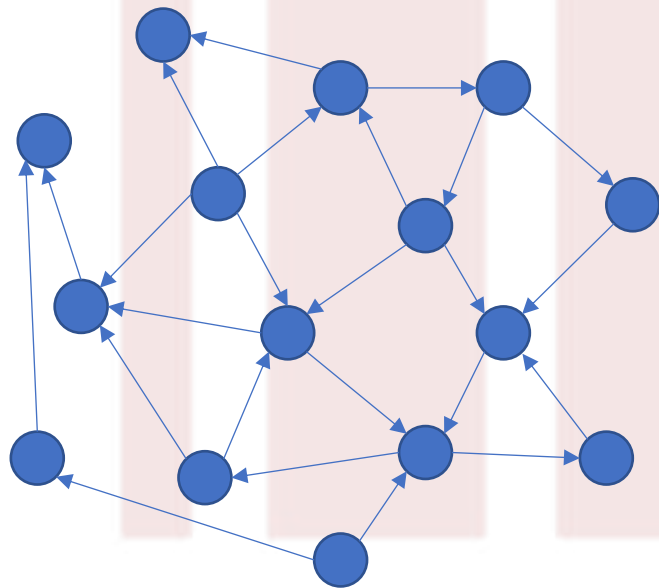
Reproducible Experiments

Topical Subsets

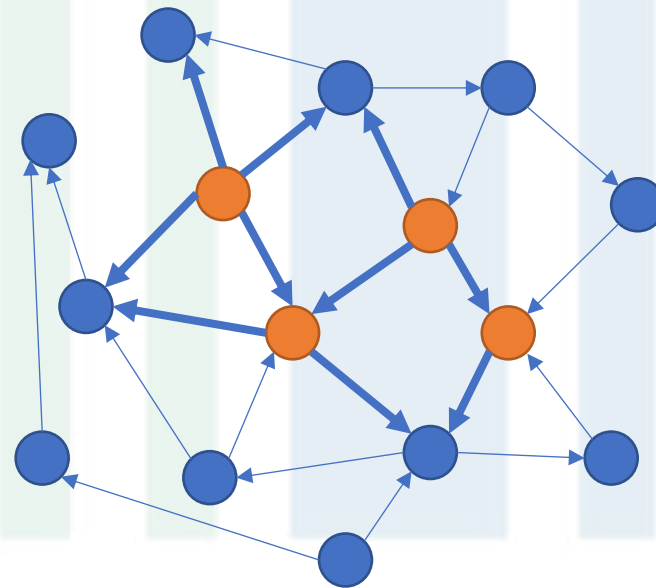
Select a set of entities, filtered based on a given topic (e.g. life science, politics, academia, etc.)

+

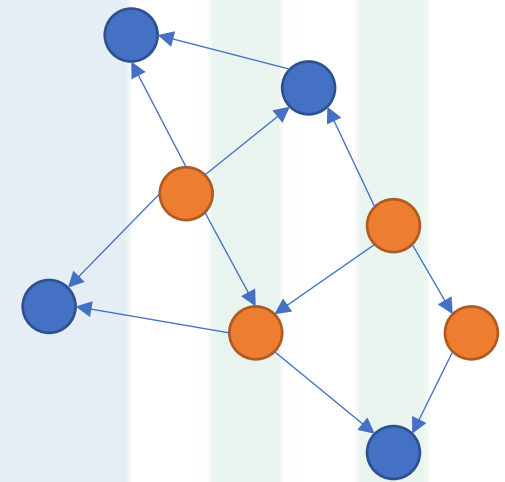
All their facts



Original Graph



**Selecting topic
related entities**

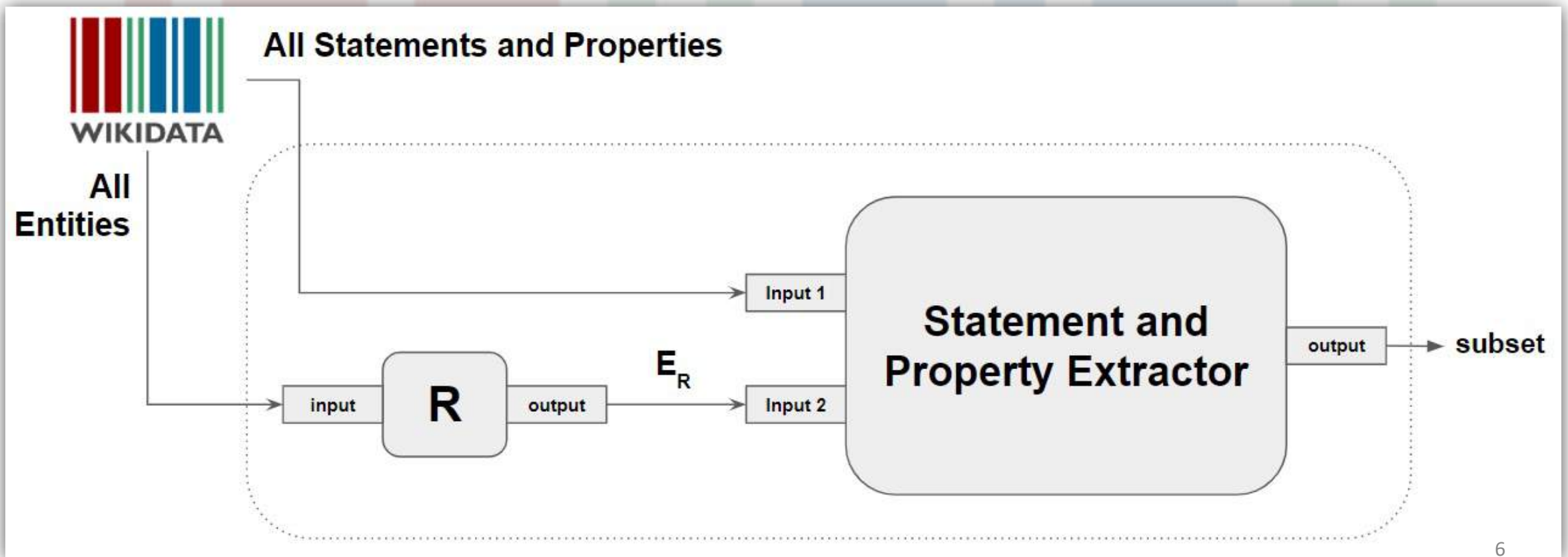


Final subset

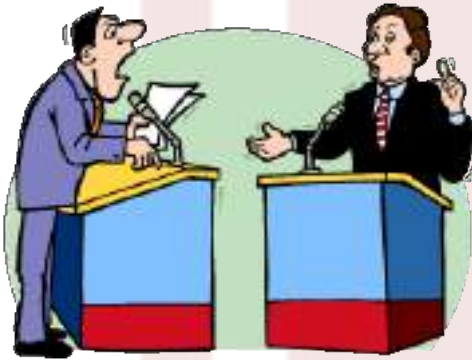
Topical Subsets

Select a set of entities, filtered based on a given topic (e.g. life science, politics, academia, etc.)

+
All their facts



Use Cases



Politicians

- One-type
- No Conditions



General(military) Politicians

- One-type
- More Conditions



UK Universities

- One-type
- Small Output



Gene Wiki*

- 17 types
- Wikipedia project

WIKIDATA

Related Works



Method	Comment	Limitation(s) regarding topical subsetting
G2GML (Matsumoto et al. 2018)	conversion of RDF graphs into Property Graphs	No RDF output
Context Graph (Mimouni et al. 2020)	Captures all adjacent nodes and edges within a given radius	Not practical for the concept of topic
ShEx	Schema Language for describing RDF graphs	Not Practical for large-scale data

WDumper

- Java-backend + Flask-frontend
- Based on Wikidata Development Kit
- Inputs:
 - A JSON specification file (Filters)
 - A JSON.gz Wikidata complete dump
- Operation:
 - Applying filters on JSON.gz dump
- Output:
 - Custom RDF dump

The screenshot shows the WDumper web interface. At the top, it says "WDumper - A tool to create custom wikidata RDF dumps" and has a link for "About Recent dumps".

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

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

Default rule: This rule is applied if no other rules matches.

How to export:

- simple statements: ON
- full statement mode: complete without references none
- export only with rank: best rank not deprecated any

+ Add custom rule

Additional settings

- labels: ON
- descriptions: ON
- aliases: ON
- sitelinks: ON
- filter languages: OFF

Dump metadata

Dump title:

Dump description:

Evaluation

- Test platform: Two different Wikidata dumps

Release Time	Size (GB)	Total Items (milion)
April 2015	~ 4.5	~ 18
November 2020	~ 90	~ 91

Evaluation

Test Conditions:



Checking the number of entities that must be on the output → COUNT queries



Checking the number of statements in each entity → DESCRIBE queries



Checking the extraction qualifiers and references



By TM/_®The Apache Software Foundation -
Vectorised by Vulphere

Evaluation

Table 4. Results of performing DESCRIBE queries on the selected entity.

Use case	Entity	2015 Dump		2020 Dump	
		Output	Input	Output	Input
Politicians	Q23	408	776	871	921
General (military) Politicians	Q355643	104	150	207	228
UK Universities	Q1094046	64	108	208	224
Gene Wiki	Q30555	12	22	30	37

Table 6. Number of qualifiers and references for the selected property of the selected entity in the output and input of WDumper (2020 dump).

Entity	Property	Qualifiers		References	
		Output	Input	Output	Input
Q23	P26	4	4	2	2
Q355643	P485	1	1	1	1
Q1094046	P355	1	1	1	1
Q17487737	P680	24	24	96	96

WDumper Strengths vs. Weaknesses



Can use any type of wikidata properties as filter (no number limit)



Can extract all desired entities, their statements and qualifiers/references



Can not use property paths like P31/P279 and path extensions like P279*



Can not detect Type Hierarchy



Can not have connections between filters

Research Challenges

- Huge size of 2021 dump
- Syntax errors in Turtle dumps
bad chars in JSON dumps
- Different implementations of
DESCRIBE in Jena vs. WDQS

```
s:Q5870-43a1689d-4352-2c50-afd9-c1a06c3f6eab a wikibase:Statement ;
v:P1621 <http://commons.wikimedia.org/wiki/Special:FilePath/Freital%20Stadtteile.svg> ;
wikibase:rank wikibase:NormalRank .
wikibase:BestRank .

ref:9715674981266bbb82725bd662B06302bd38c929 a wikibase:Reference ;
v:P854 <http://freital.de/index.phtml?La=1&object=tx|530.4293.1&NavID=530.83&kat=8&sub=0> .

ref:dd584dca2490bbbd92737fb9b28932798de694e1 a wikibase:Reference ;
v:P854 <http://freital.de/index.phtml?La=1&object=tx|530.4535.1&NavID=530.81&sub=0>

ref:20b0cf0e4b9754d9c5ab265372315f7def7a9784 a wikibase:Reference ;
v:P854 <http://freital.de/index.phtml?La=1&object=tx|530.4535.1&NavID=530.81&sub=0 > .

ref:e3e2b2b2ffec6318c97b65ca2b97821673ffc97a a wikibase:Reference ;
v:P854 <http://www.statistik.sachsen.de/regioereg/RRServlet?function=Lesen&id=21619&type=10001&param=> ;
v:P248 entity:Q17024048 .

ref:4178e3410b1cc0464222175dbdd593d210bf42fe a wikibase:Reference ;
```

\\ \n \a

Future Works

- Adding Type Hierarchy Detection to WDumper

An initial try:

https://github.com/seyed1411/wdumper/blob/master/extensions/add_subclasses.py

- Combining WDumper Core with ShEX/SPARQL for better flexibility
- How to build live subsets?
- How about topically subsetting RDF KGs?

Summary

- Subsets are useful
 - Cost reduction + Ease of access + Flexibility
- Topical subsets: Set of **entities** + their **facts** around a given **topic**
- WDumpster is a reliable tool for extracting (some) Topical Substes
 - + Good for topics with simple type structure
 - Only applicable on Wikidata
 - Not flexible