Tutorial: Linking, Extending, Exploiting and Enhancing Tabular Data with Wikidata

Part 2: Getting data from Wikidata

Daniel Garijo, **Pedro Szekely**

Information Sciences Institute and
Department of Computer Science

@dgarijov, @szeke
{dgarijo, pszekely}@isi.edu

Wikidata SPARQL Endpoint

Information Sciences Institute

Daniel Garijo and Pedro Szekely. Linking, Extending, Exploiting and Enhancing Tabular Data with Wikidata. K-CAP 2019

2

Wikidata SPARQL endpoint

https://query.wikidata.org

Wikidata Query Service — Examples — Help — More tools

Query Helper — Show help

+ Filter — Full screen

+ Show — Show prefixes

— Format query

Limit 100 — Examples

— Undo

— Clear

— Link to this query

`1 (Input a SPARQL query or choose a query example)`

Features:
- Fast
- No named graphs/reasoning
- **One client** is allowed 60 seconds of processing time each 60 seconds.
  - One client is allowed 30 error queries **per minute**

Having trouble with a query? Request it: https://www.wikidata.org/wiki/Wikidata:Request_a_query

Information Sciences Institute

Daniel Garijo and Pedro Szekely. Linking, Extending, Exploiting and Enhancing Tabular Data with Wikidata. K-CAP 2019

3

SPARQL Endpoint Helpers: Auto completion

```
1  select ?s where {
2    ?s rdfs:label "Los Angeles"@en
3  }
```

Q wd:Q65

```
1  select ?p ?o where {
2    wd:Q65 ?p ?o
3  }
```

Los Angeles (Q65) county seat of Los Angeles County, California; second largest city in the United States by population

Q65 (Q504331) Dutch band

Q65 (Q27788669) Wikimedia disambiguation page

skos:altLabel	洛杉磯市
skos:altLabel	洛杉磯市
skos:altLabel	洛杉矶市
skos:altLabel	洛杉矶市
skos:altLabel	洛杉磯市
wikibase:statements	209
wikibase:sitelinks	245
wikibase:identifiers	47
wikibase:timestamp	9 November 2019
wdt:P6	Q wd:Q5386557
wdt:P17	Q wd:Q30

Available in JSON, Turtle, JSON-LD...

More information: https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/A_gentle_introduction_to_the_Wikidata_Query_Service

Information Sciences Institute

Daniel Garijo and Pedro Szekely. Linking, Extending, Exploiting and Enhancing Tabular Data with Wikidata. K-CAP 2019

4

SERVICE wikibase:label { bd:serviceParam wikibase:language "en" . }

SPARQL Endpoint Query Service

Functionality designed to facilitate recovering basic resource metadata

```
select ?p ?o ?oLabel ?oAltLabel ?oDescription where{
  wd:Q65 ?p ?o
  SERVICE wikibase:label {
      bd:serviceParam wikibase:language "en" .
  }
}
```

wdt:P17	Q wd:Q30	United States of America	America, United States, USA, the United States, the United States of America, the US, the USA, U.S., U.S.A., US, us, US of A	federal republic in Nortl America
wdt:P18	🖼 commons:LA.jpg	http://commons.wikimedia.org/wiki/Special:FilePath/LA.jpg		
wdt:P30	Q wd:Q49	North America	NA	continent on the Earth': northwestern quadrant
wdt:P31	Q wd:Q62049	county seat	parish capitals, parish seat, parish seats, shire town	administrative center, o of government, for a co or civil parish
wdt:P31	Q wd:Q1093829	city of the United States	city of the United States, city in the United States, city in the USA, city of the USA	administrative division local government level United States

SPARQL Endpoint query service (advanced)

Geospatial search (may be slow)

```
#defaultView:Map
SELECT ?place ?placeLabel ?location
WHERE
{
    # Los Angeles coordinates
    wd:Q65 wdt:P625 ?LALoc .
    SERVICE wikibase:around {
        ?place wdt:P625 ?location .
        bd:serviceParam wikibase:center ?LALoc
        bd:serviceParam wikibase:radius "100" .
    } .
    # Is an airport
    FILTER EXISTS { ?place wdt:P31/wdt:P279* wd:Q1248784 } .
    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en" .
    }
}
```
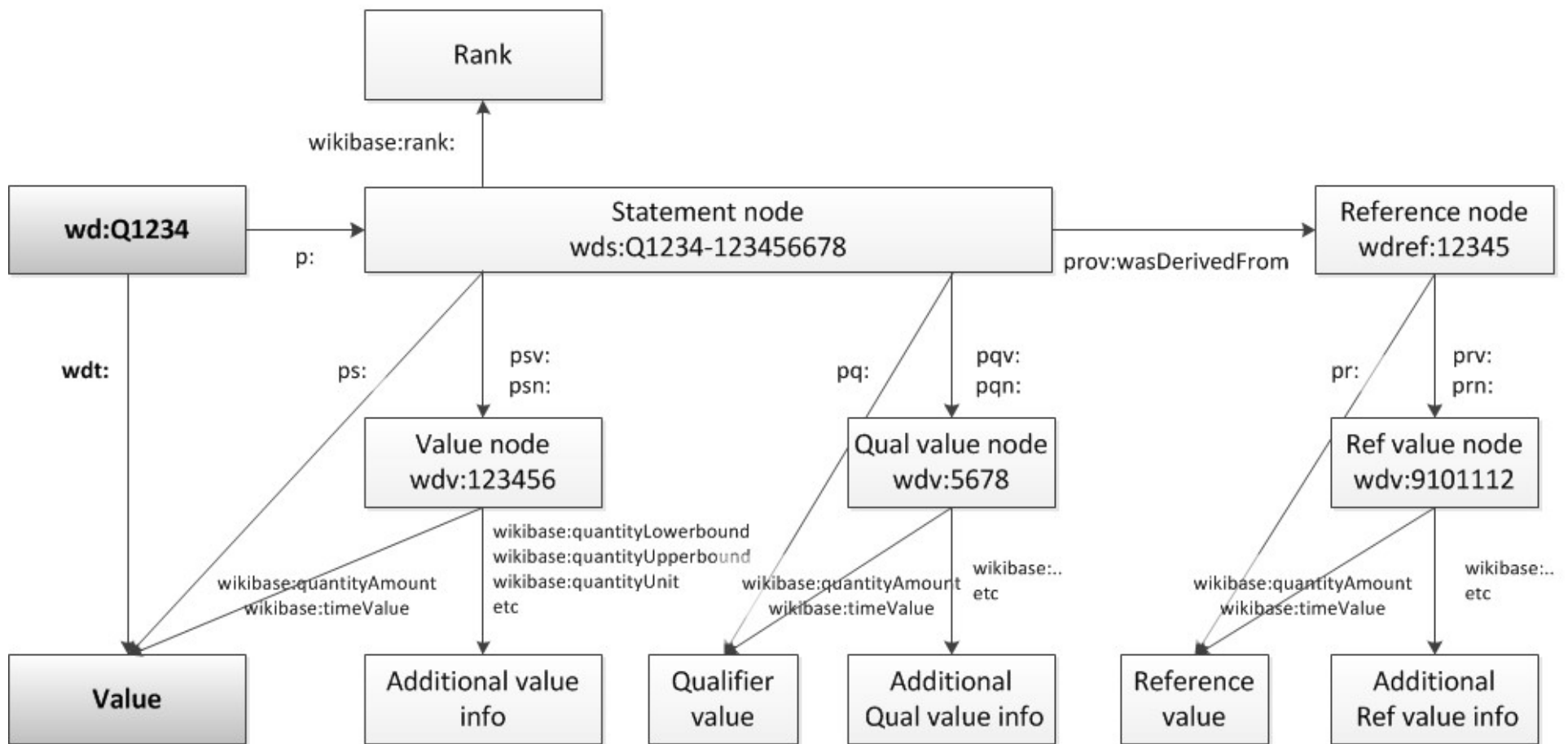
Could also be "box"
(and pass the bounding box)
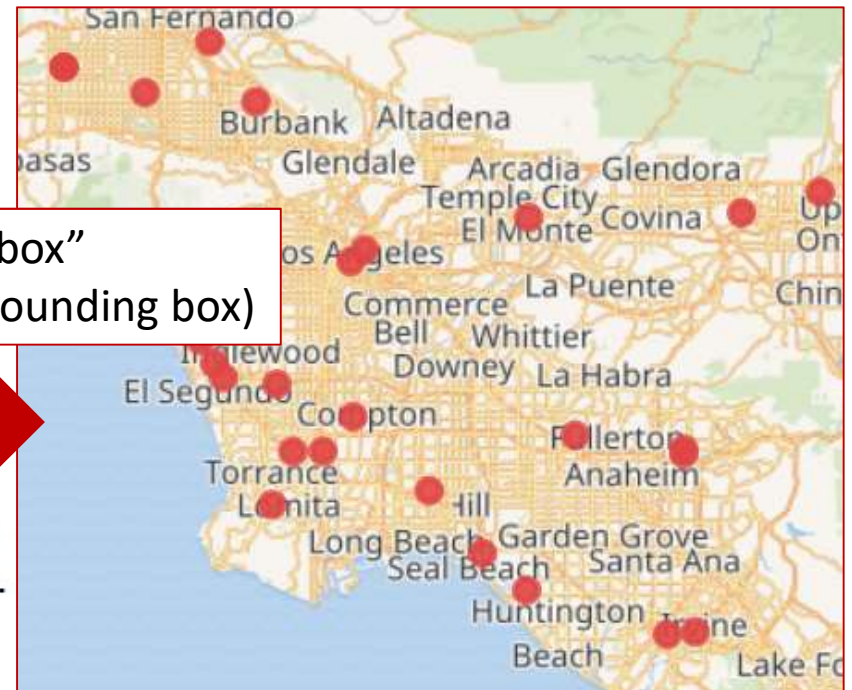
More info: https://www.mediawiki.org/wiki/Wikidata_Query_Service/User_Manual
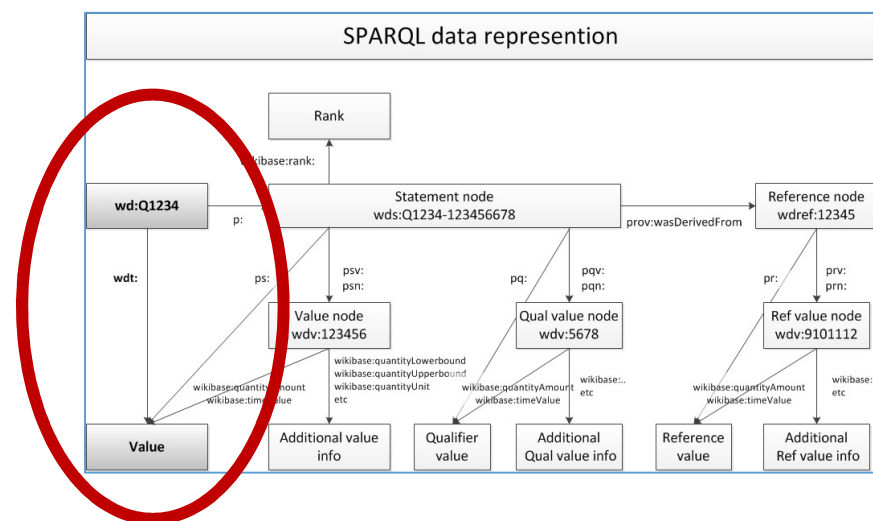
SPARQL endpoint cheat sheet


SPARQL data represention

Rank

wikibase:rank:

wd:Q1234 — p: → Statement node
wds:Q1234-123456678 — prov:wasDerivedFrom → Reference node
wdref:12345

wdt:

ps:
psv:
psn:

Value node
wdv:123456

pq:
pqv:
pqn:

Qual value node
wdv:5678

pr:
prv:
prn:

Ref value node
wdv:9101112

wikibase:quantityAmount
wikibase:timeValue

wikibase:quantityLowerbound
wikibase:quantityUpperbound
wikibase:quantityUnit
etc

wikibase:quantityAmount
wikibase:timeValue

wikibase:..
etc

wikibase:quantityAmount
wikibase:timeValue

wikibase:..
etc

Value

Additional value
info

Qualifier
value

Additional
Qual value info

Reference
value

Additional
Ref value info

Example: Simple fact from Los Angeles

Who is the governor?

SPARQL data representation

```
select ?p ?o ?oLabel ?oAltLabel ?oDescription where{
wd:Q65 wdt:P6 ?o
SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en" .
    }
}
```
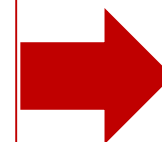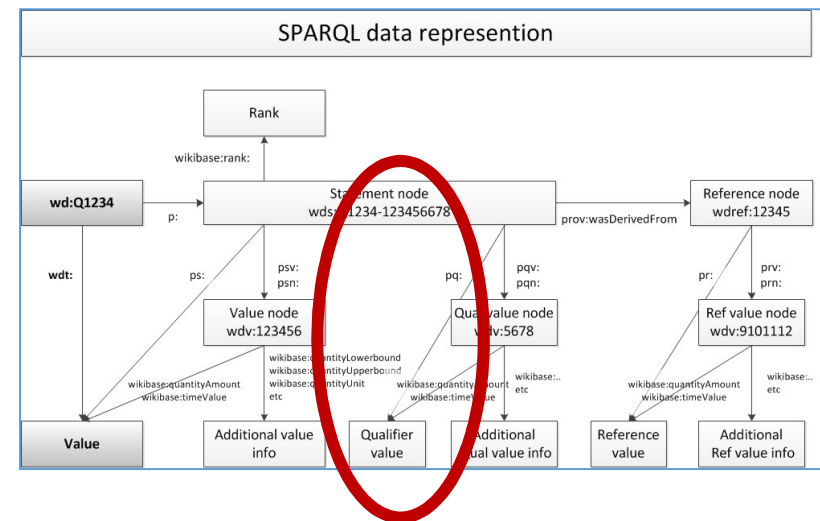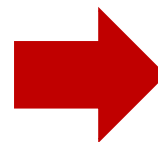
o	oLabel
Q wd:Q5386557	Eric Garcetti

Link to query: https://w.wiki/CFf

Example: Simple fact from Los Angeles: Qualifiers

How long has the governor been in office?

SPARQL data representation

```
select ?start where{
  wd:Q65 p:P6 ?statement.
  ?statement pq:P580 ?start
}
```

start

1 July 2013

Link to query: https://w.wiki/CFh

Information Sciences Institute

Daniel Garijo and Pedro Szekely. Linking, Extending, Exploiting and Enhancing Tabular Data with Wikidata. K-CAP 2019
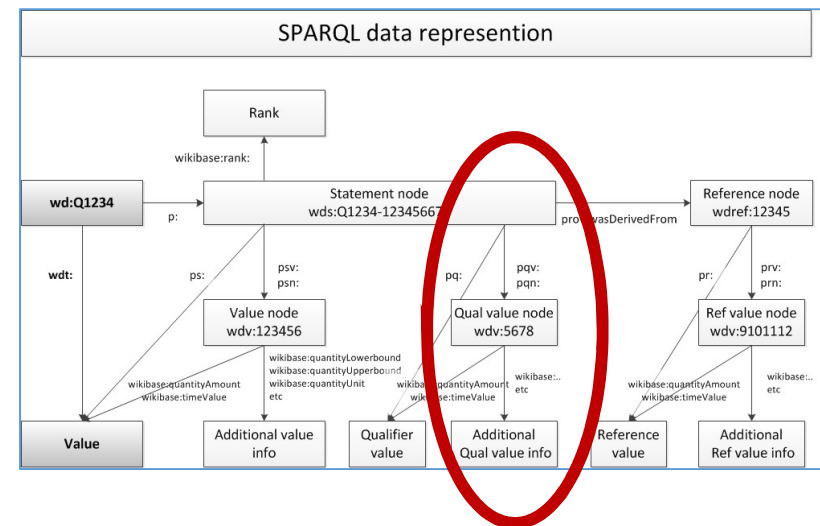
9

Example: Simple fact from Los Angeles: Additional Qualifier information

If we want to know more about how the qualifier value has been represented (e.g., calendar, etc.)


```
select ?p ?o where{
 wd:Q65 p:P6 ?statement.
 ?statement pqv:P580 ?start.
  ?start ?p ?o
}
```

p	o
wikibase:timeValue	1 July 2013
wikibase:timePrecision	11
wikibase:timeTimezone	0
wikibase:timeCalendarModel	wd:P1985727
wikibase:timeCalendarModel	wd:Q1985727
rdf:type	wikibase:TimeValue

Link to query: https://w.wiki/CFj

Information Sciences Institute

Daniel Garijo and Pedro Szekely. Linking, Extending, Exploiting and Enhancing Tabular Data with Wikidata. K-CAP 2019

10

Advanced: Truthy triples

Let's retrieve the population of LA:

```
select ?population where{
  wd:Q65 wdt:P1082 ?population.
}
```
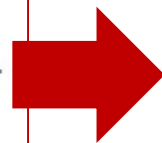
population
3976322

This returns a single result, which corresponds to the "truthy" statement (first rank). In Wikidata, the truthy statement for population is set to the latest available year:

```
select ?value ?year ?rank where{
  wd:Q65 p:P1082 ?ps. #1082 is population
  ?ps ps:P1082 ?value .
  ?ps wikibase:rank ?rank.
  ?ps pq:P585 ?year. #point in time

}order by ?year
```

Link to query: https://w.wiki/CFj, https://w.wiki/CFn
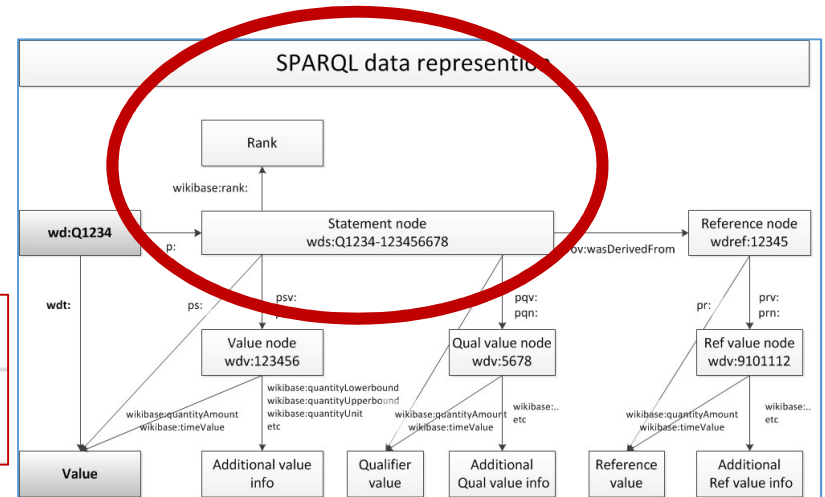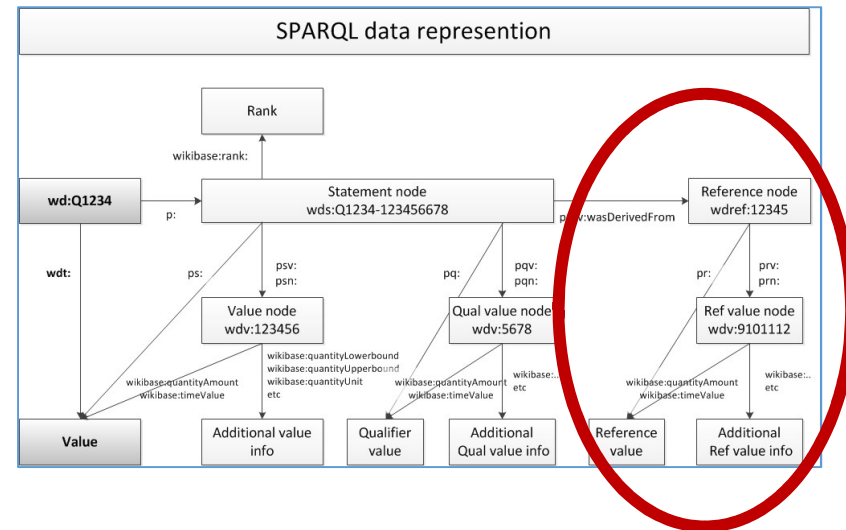
Advanced: Truthy triples - Results

```
select ?value ?year ?rank where{
 wd:Q65 p:P1082 ?ps. #1082 is population
 ?ps ps:P1082 ?value .
 ?ps wikibase:rank ?rank.
 ?ps pq:P585 ?year. #point in time

}order by ?year
```

value	year	rank
1610	1 January 1850	wikibase:NormalRank
4385	1 January 1860	wikibase:NormalRank
5728	1 January 1870	wikibase:NormalRank
11183	1 January 1880	wikibase:NormalRank
50395	1 January 1890	wikibase:NormalRank
102479	1 January 1900	wikibase:NormalRank
319198	1 January 1910	wikibase:NormalRank
3792621	1 January 2010	wikibase:NormalRank
3976322	1 January 2016	wikibase:PreferredRank

Information Sciences Institute

Daniel Garijo and Pedro Szekely. Linking, Extending, Exploiting and Enhancing Tabular Data with Wikidata. K-CAP 2019

12

Advanced: References

Population in LA: What is the source
(reference URL) for the triples?


```
select ?value ?year ?rank ?refValueLabel where{
 wd:Q65 p:P1082 ?ps. #1082 is population
 ?ps ps:P1082 ?value .
 ?ps wikibase:rank ?rank.
 ?ps pq:P585 ?year. #point in time
 ?ps prov:wasDerivedFrom/pr:P854 ?refValue #reference URL
 SERVICE wikibase:label {
    bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en" .
  }
}order by ?year
```

Link to query: https://w.wiki/CFo

Advanced: References - Results

value	year	rank	refValueLabel
11183	1 January 1880	wikibase:NormalRank	http://www.laalmanac.com/population/po02.htm
50395	1 January 1890	wikibase:NormalRank	http://physics.bu.edu/~redner/projects/population/cities/la.html
102479	1 January 1900	wikibase:NormalRank	http://www.laalmanac.com/population/po02.htm
3792621	1 January 2010	wikibase:NormalRank	http://search.census.gov/search?q=cache:b2N7EEjvBBcJ:2010.census.gov/news/xls/cb11 8&client=2010prod&proxystylesheet=2010prod&site=2010&access=p&oe=UTF-8
3976322	1 January 2016	wikibase:PreferredRank	https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk

Link to query: https://w.wiki/CFo

Information Sciences Institute

Daniel Garijo and Pedro Szekely. Linking, Extending, Exploiting and Enhancing Tabular Data with Wikidata. K-CAP 2019

14

Advanced: Units

Elevation above sea level in LA?
(with units in which it's measured)

```
select ?value ?rank ?unitLabel where{
  wd:Q65 p:P2044 ?ps. #2044 is above sea level
  ?ps ps:P2044 ?value .
  ?ps wikibase:rank ?rank.
  ?ps psv:P2044/wikibase:quantityUnit ?unit
  SERVICE wikibase:label {
      bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en" .
    }
}
```
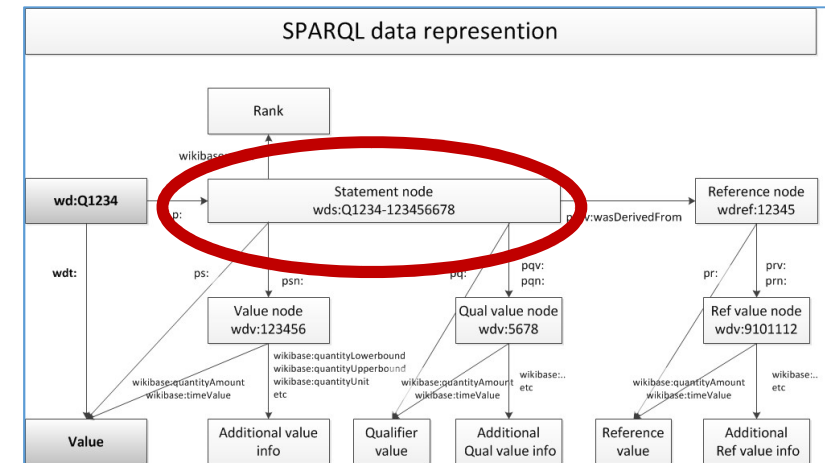

SPARQL data representation
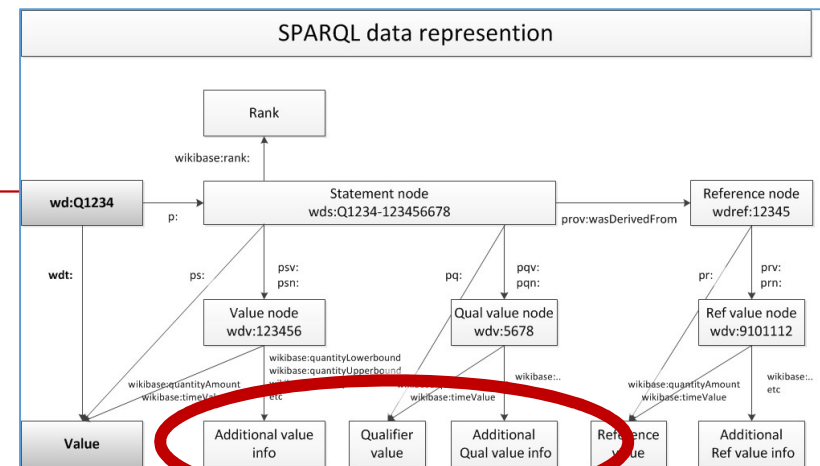
value	rank	unitLabel
93	wikibase:NormalRank	metre

Link to query: https://w.wiki/CFs

Advanced+: Properties

What if next I need to know the properties used to describe LA? (Q65)

SPARQL data representation

```
SELECT distinct ?ps ?wdLabel {

  wd:Q65 ?p ?statement .
  ?statement ?ps ?ps_ .

  ?wd wikibase:claim ?p.
  ?wd wikibase:statementProperty ?ps.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en" }
} |
```

```
#Truthy statements
SELECT distinct ?p ?wdLabel {

  wd:Q65 ?p ?statement .
  ?wd wikibase:directClaim ?p.

  SERVICE wikibase:label { bd:serviceParam wikibase:language "en" }
} |
```

Daniel Garijo and Pedro Szekely. Linking, Extending, Exploiting and Enhancing Tabular Data with Wikidata. K-CAP 2019

Advanced+: Querying for Classes

An entity can be a class and an instance in Wikidata:

```
select ?sLabel ?classLabel ?upperClassLabel where {
  ?s wdt:P31 ?class.
  ?s wdt:P279 ?upperClass
    SERVICE wikibase:label {
     bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en" .
  }
}LIMIT 100
```

sLabel	classLabel	upperClassLabel
Boeing 747-8	aircraft model	Boeing 747
Barossa German	dialect	German
Namibian Black German	pidgin	German
Franks	language	German
Franks	ethnic group	German

Link to query: https://w.wiki/CFv

Information Sciences Ins

Using SPARQL to plot results

• Plotting population by year:

Table ▾		
value		**year** ⇕
11183		1 January 1880
50395		1 January 1890
102479		1 January 1900
3792621		1 January 2010
3976322		1 January 2016

Daniel Garijo and Pedro Szekely. Linking, Extending, Exploiting and Enhancing Tabular Data with Wikidata. K-CAP 2019
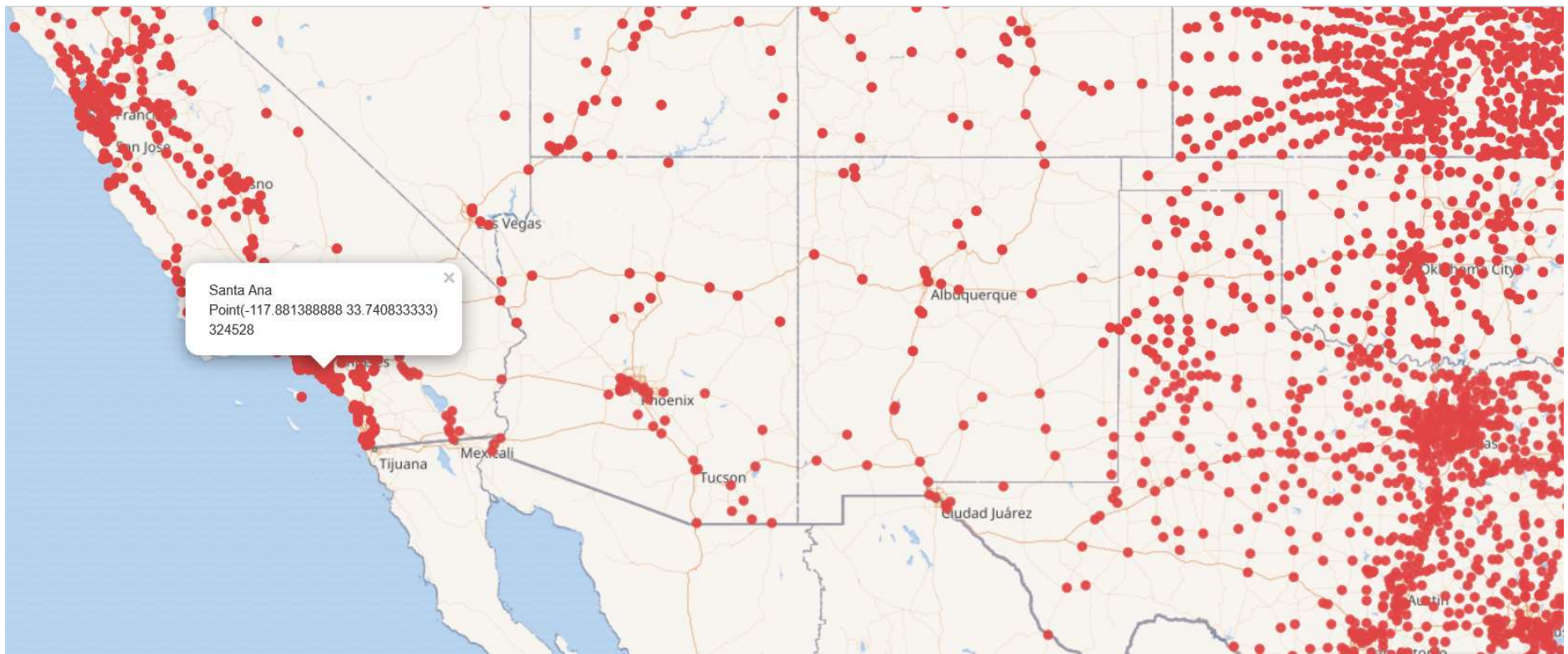
Using SPARQL to plot results
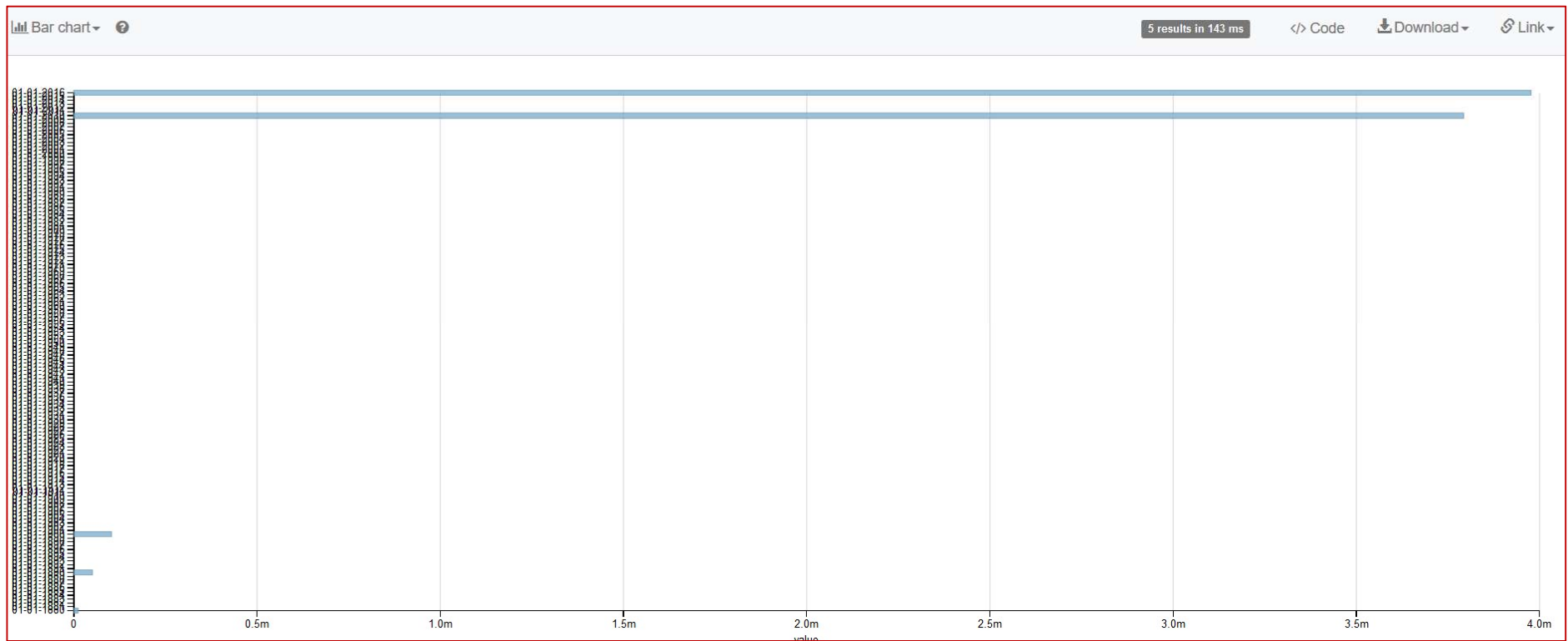
Using SPARQL to plot results – geospatial queries

If we want to know the latest known population for all cities in the US:

```
select ?cityLabel ?value ?coords  where{
 ?city wdt:P31 wd:Q1093829 .
 ?city wdt:P1082 ?value .
 ?city wdt:P625 ?coords . #coordinates |
 SERVICE wikibase:label {
    bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en" .
  }
}
```

Link to query: https://w.wiki/CFy

Daniel Garijo and Pedro Szekely. Linking, Extending, Exploiting and Enhancing Tabular Data with Wikidata. K-CAP 2019

Using SPARQL to plot results – geospatial queries

Santa Ana
Point(-117.881388888 33.740833333)
324528

Link to query: https://w.wiki/CFy

Information Sciences Institute

Daniel Garijo and Pedro Szekely. Linking, Extending, Exploiting and Enhancing
Tabular Data with Wikidata. K-CAP 2019

21