Efficient Keyword Search For The QLever SPARQL Engine

…or how we improved the QLever text search.

Nick Göckel

What is SPARQL?

- query language
- used for querying RDF data
  - common format to store data
  - uses only triples

Example query:

```
SELECT ?scientist WHERE {
    ?scientist <Award_Won> <Nobel_Prize_in_Physics> .
}
```

Run on database:

Subject	Predicate	Object
<Albert_Einstein>	<Award_Won>	<Nobel_Prize_in_Physics>
<Carl_Bosch>	<Award_Won>	<Nobel_Prize_in_Chemistry>
<Charles_Darwin>	<Award_Won>	<Royal_Medal>
<Marie_Curie>	<Award_Won>	<Nobel_Prize_in_Chemistry>
<Marie_Curie>	<Award_Won>	<Nobel_Prize_in_Physics>

Example query:

```
SELECT ?scientist WHERE {
    ?scientist <Award_Won> <Nobel_Prize_in_Physics> .
}
```

Run on database:

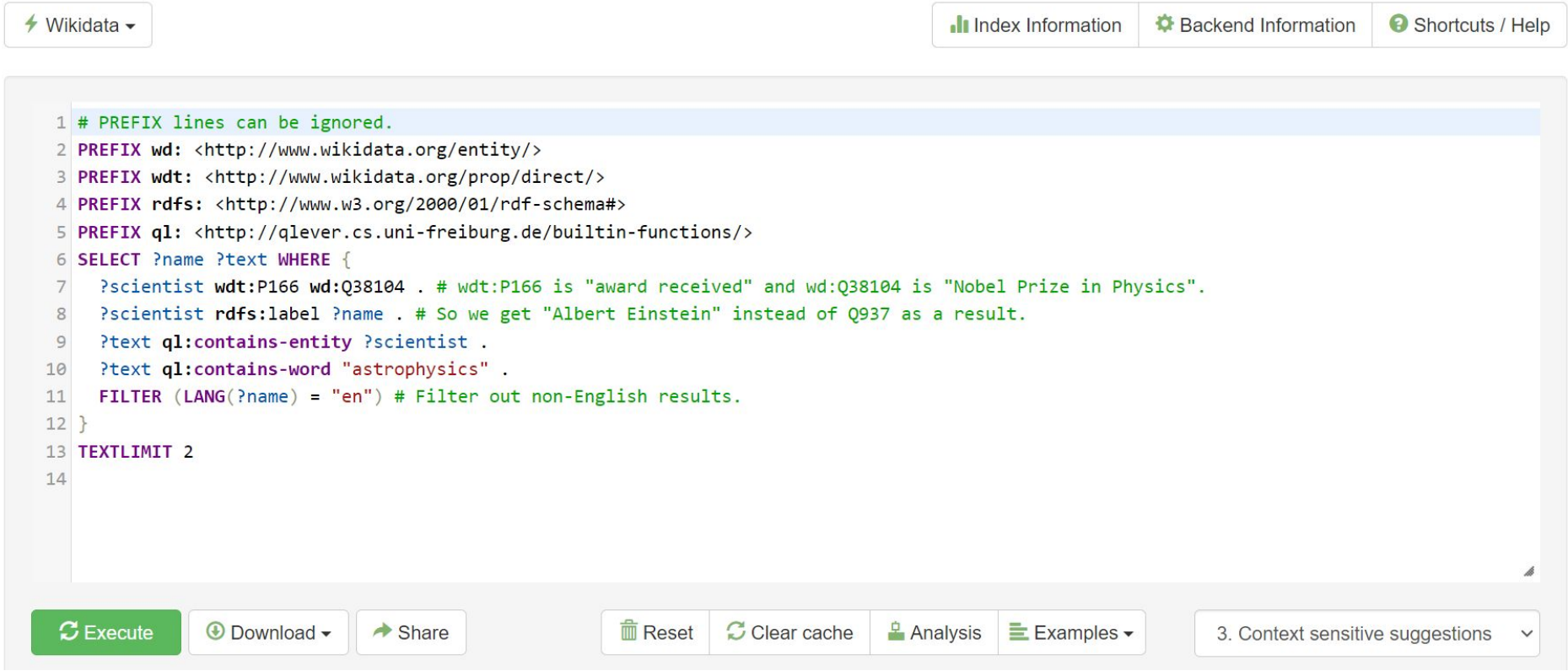
Subject	Predicate	Object
<Albert_Einstein>	<Award_Won>	<Nobel_Prize_in_Physics>
<Carl_Bosch>	<Award_Won>	<Nobel_Prize_in_Chemistry>
<Charles_Darwin>	<Award_Won>	<Royal_Medal>
<Marie_Curie>	<Award_Won>	<Nobel_Prize_in_Chemistry>
<Marie_Curie>	<Award_Won>	<Nobel_Prize_in_Physics>

Has result:

?scientist
<Albert_Einstein>
<Marie_Curie>

What is QLever?

- SPARQL query engine
  - runs SPARQL queries on RDF data bases
- allows for combined search
  - search on structural data from an RDF knowledge graph
  - but also on textual information from a collection of texts



https://qlever.cs.uni-freiburg.de/wikidata/v42wB4

How did we improve the QLever text search?

1. Implemented new feature, that extends the text search
2. Improved the code structure of the text search

Query results:

143 lines found | 2ms in total | 1ms for computation | 1ms for resolving and sending

Query Virtuoso | Query WDQS | Limited to 100 results; show all 143 results

	?name	?text	?ql_matchingword_text_astrophy
1	John C. Mather	John Cromwell Mather (born August 7, 1946, Roanoke, Virginia) is an American astrophysicist, cosmologist and Nobel Prize in Physics laureate for his work on the Cosmic Background Explorer Satellite (COBE) with George Smoot.	astrophysicist
2	John C. Mather	Mather is a senior astrophysicist at the NASA Goddard Space Flight Center (GSFC) in Maryland and adjunct professor of physics at the University of Maryland College of Computer, Mathematical, and Natural Sciences.	astrophysicist
3	Vitaly Ginzburg	He also headed the Academic Department of Physics and Astrophysics Problems, which Ginzburg founded at the Moscow Institute of Physics and Technology in 1968.	astrophysics
4	Vitaly Ginzburg	Soviet astrophysicist Vitaly Ginzburg said that ideologically the "Bolshevik communists were not merely atheists, but, according to Lenin's terminology, militant atheists" in excluding religion from the social mainstream, from education and from government.	astrophysicist
5	Adam Riess	Adam Guy Riess (born December 16, 1969) is an American astrophysicist and Bloomberg Distinguished Professor at Johns Hopkins University and the Space Telescope Science Institute.	astrophysicist
6	Adam Riess	In astrophysics, Press is best known for his discovery, with Paul Schechter, of the Press–Schechter formalism, which predicts the distribution of masses of galaxies in the Universe; and for his work with Adam Riess and Robert Kirshner on the calibration of distant supernovas as "standard candles".	astrophysics
7	Saul Perlmutter	The newest supercomputer Perlmutter, is named after Saul Perlmutter, an astrophysicist at Berkeley Lab who shared the 2011 Nobel Prize in Physics for his contributions to research showing that the expansion of	astrophysicist

Why do we want this feature?

- we can use information like normal variable
  - group by
  - join
  - filter
  - etc.
- allows user to run new set of queries
  - for example: https://qlever.cs.uni-freiburg.de/wikidata/onBH1f

Search @ Satzungen

Powered by AD Freiburg

staat

Zoomed in on 32 text records

Refine by WORD

☐	staatlichen	6
☐	staatlich	6
☐	staatsangehörigen	1
☐	staatsexamensstudiengängen	4
☐	mitgliedstaates	2

1 − 20 of 20

Refine by TITLE

☐	Corona-Satzung Uni Freiburg	10

Number of hits: 32, showing: 1 −

Corona-Satzung Uni F
staatenloser Studienbe

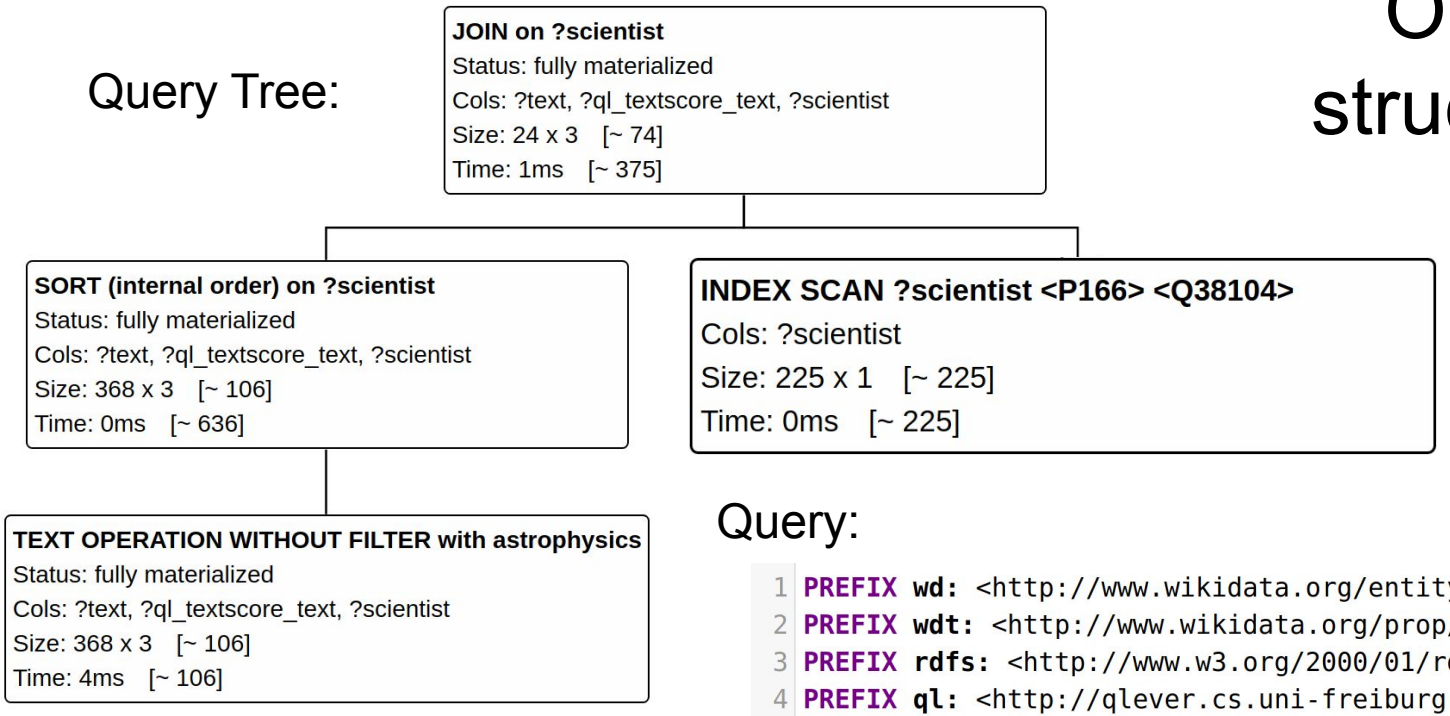
Abweichend von § 9 Absat
(ZImmO) der Universität Fr
ausländischen oder staate
Erscheinen und auf die Vor
zichtet werden, wenn er/sie
2021 beziehungsweise im
Bundesgebiet einreist. Nac
seiner/ihrer Einreise in das
Krankenkasse bestehende
über eine Aufenthaltserlaub
unverzüglich, ist der/die St

MSc Rahmenprüfungs
Studienleistungen und

(1) Studienzeiten, Studien-

OLD structure

Query Tree:



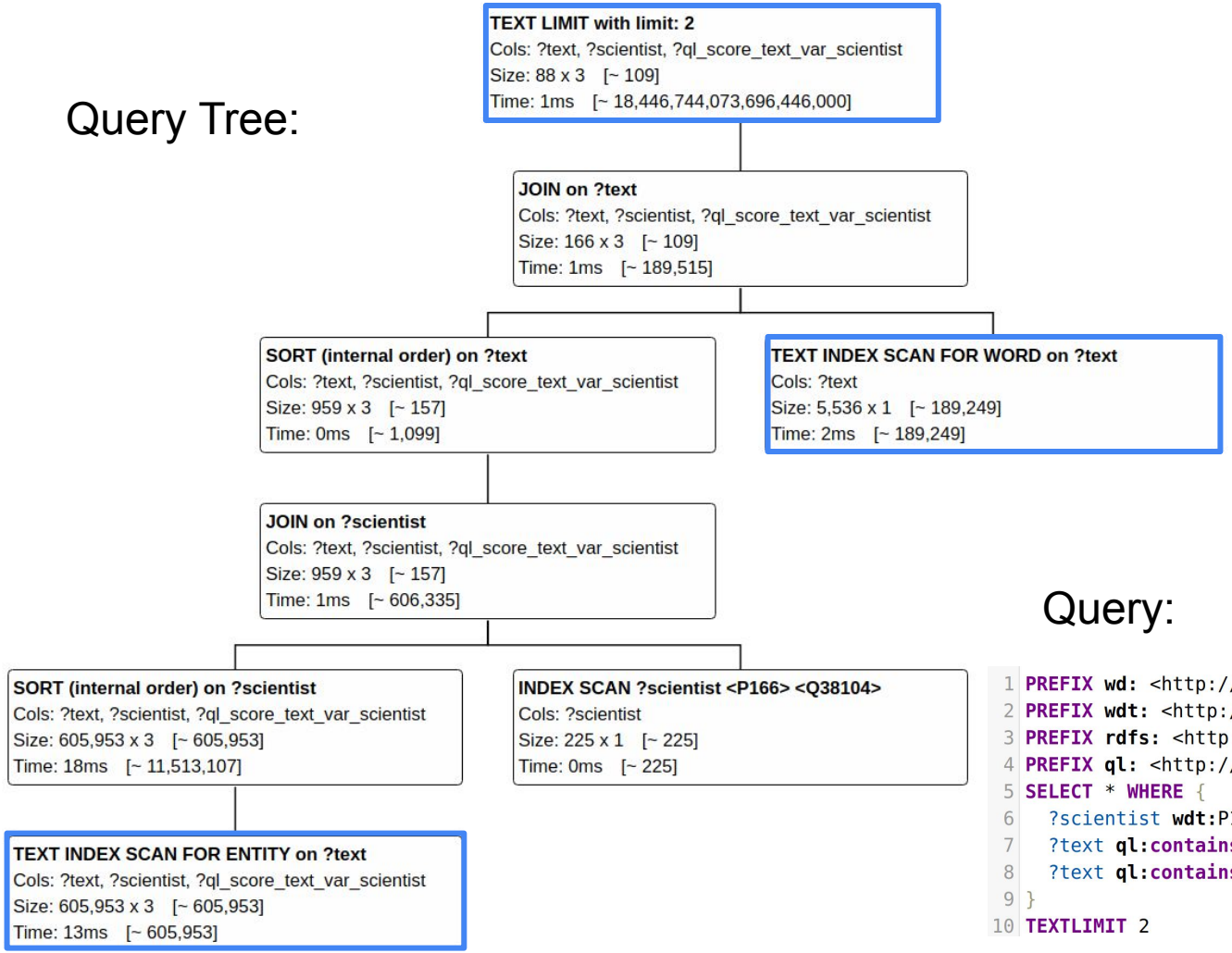
Query:

```
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ql: <http://qlever.cs.uni-freiburg.de/builtin-functions/>
SELECT * WHERE {
  ?scientist wdt:P166 wd:Q38104 .
  ?text ql:contains-entity ?scientist .
  ?text ql:contains-word "astrophysics" .
}
TEXTLIMIT 2
```

Why is this structure bad?

- hard to read
- hard to maintain
- code duplication (e.g. join operation)
- hard to implement new features

Query Tree:



NEW structure

Query:

```
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ql: <http://qlever.cs.uni-freiburg.de/builtin-functions/>
SELECT * WHERE {
  ?scientist wdt:P166 wd:Q38104 .
  ?text ql:contains-entity ?scientist .
  ?text ql:contains-word "astrophysics" .
}
TEXTLIMIT 2
```

Implementing feature

- only need to change code behind ql:contains-word
  - needs to read extra information from the text index
- other operations can stay the same

⇒ new text search structure made implementation way easier

Evaluation

Feature Implementation:

- implemented with full functionality as imagined
- high code quality

Restructuring:

- increased readability
- increased maintainability
- decreased duplication
- high code quality

Questions?