TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Institut für Computertechnik
Institute of Computer Technology

Extracting Data from the Deep Web with Global-as-View Mediators Using Rule-Enriched Semantic Annotations

Benjamin Dönz
doenz[at]ict.tuwien.ac.at

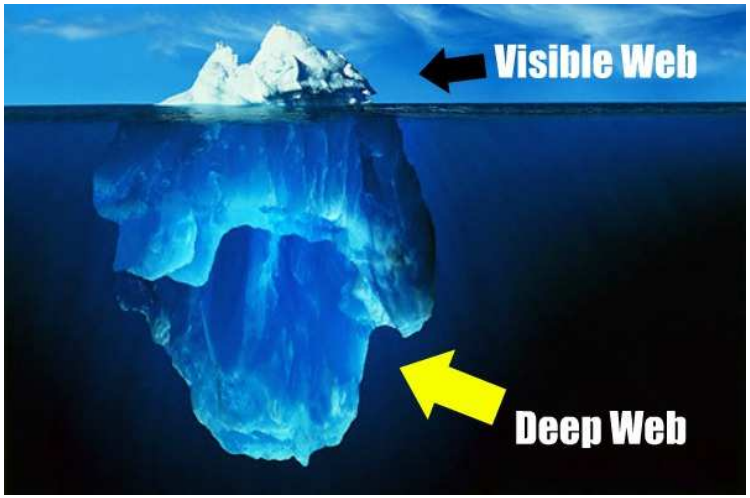
Vienna University of Technology
Institute of Computer Technology
Vienna, Austria

Harold Boley
harold.boley[at]unb.ca

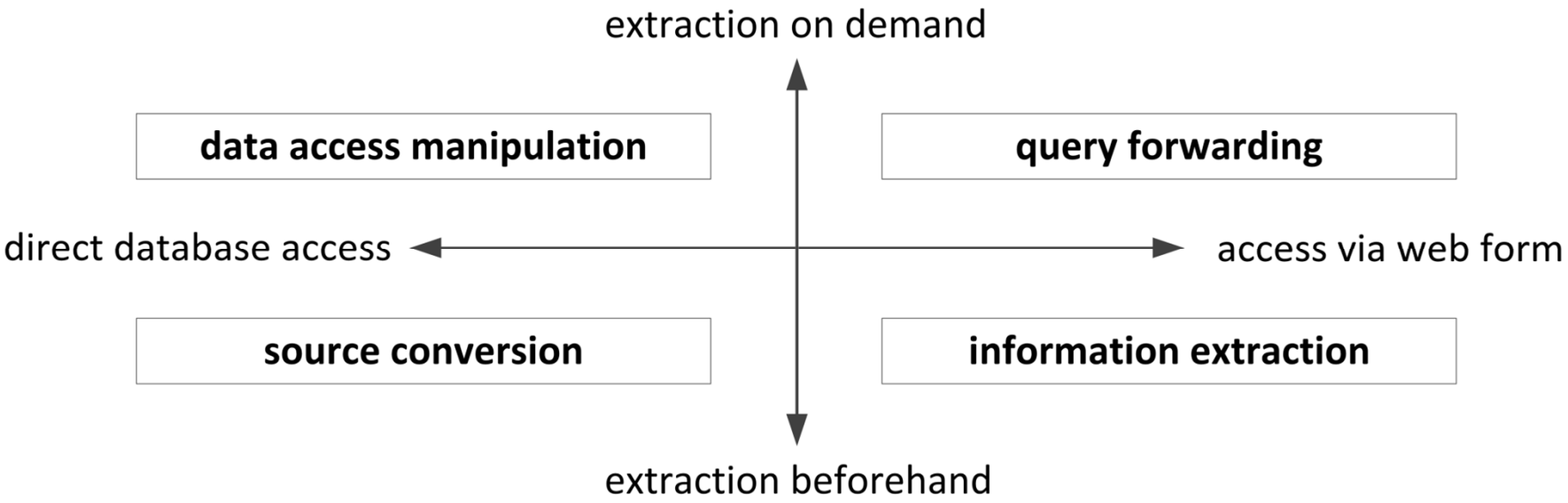
University of New Brunswick
Faculty of Computer Science
Fredericton, NB, Canada

The „Deep Web“ – What is it?

- Data hidden behind search forms and interfaces
- Estimated 400-500 times more information than the indexable World Wide Web
- 77% of the content classified as structured information
- Template based – so understanding how to extract one result allows to extract them all
- Examples:
  - Web shops
  - Classified advertising
  - Miscellaneous databases



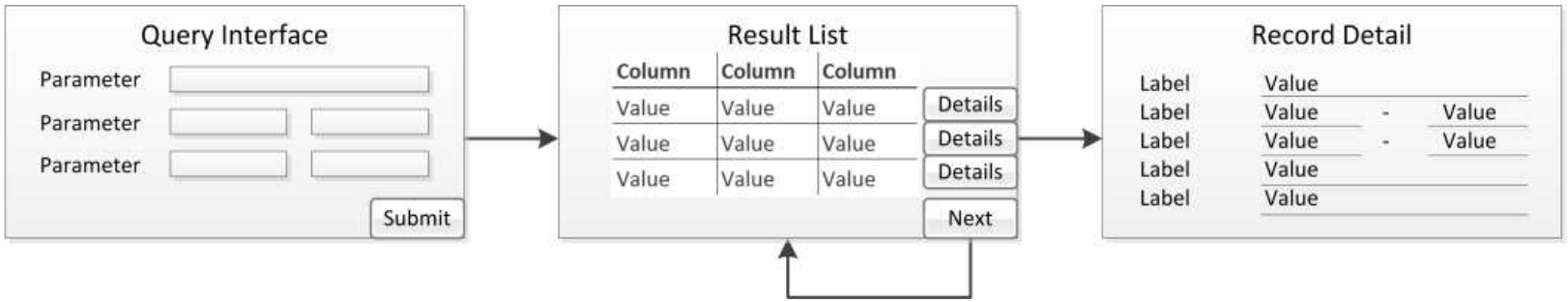
Accessing the Deep Web



Our Approach: Upper-Right Quadrant

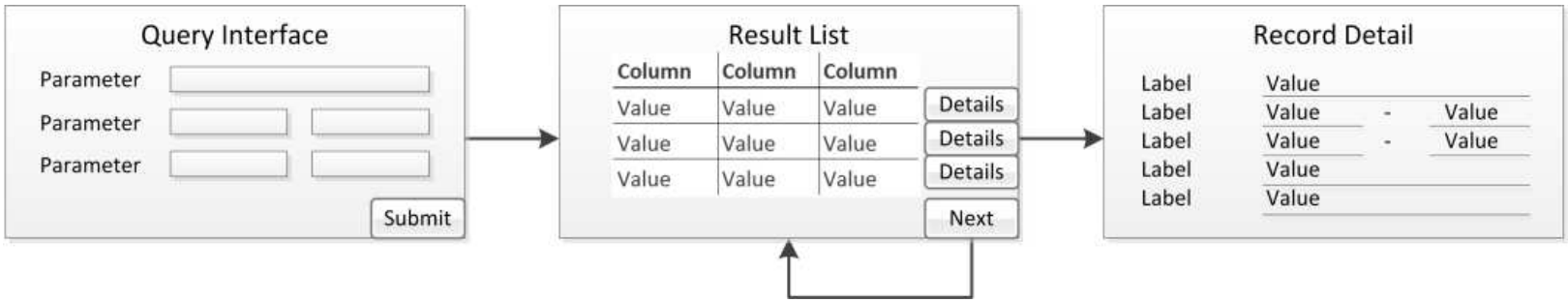
- Processing of queries using a query forwarding approach
  - SPARQL queries as input
  - Query transformation and forwarding via mediators
  - Global-as-View mapping of local sources
- Web form interaction and information extraction
  - Extraction process based on an extensible model
  - Semantic annotations for mapping real-world Web pages to the model
  - Feature-based rules for creating annotations

Model Overview



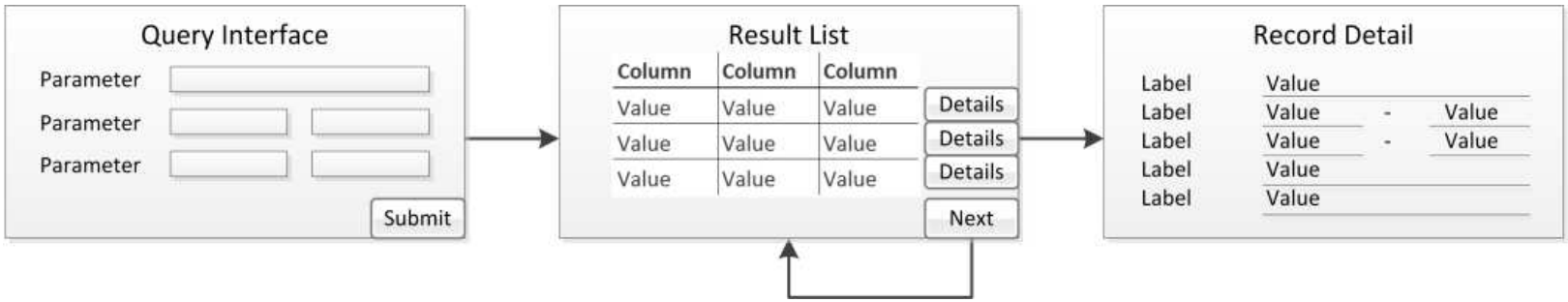
- Query interface for submitting conjunctive queries

Model Overview



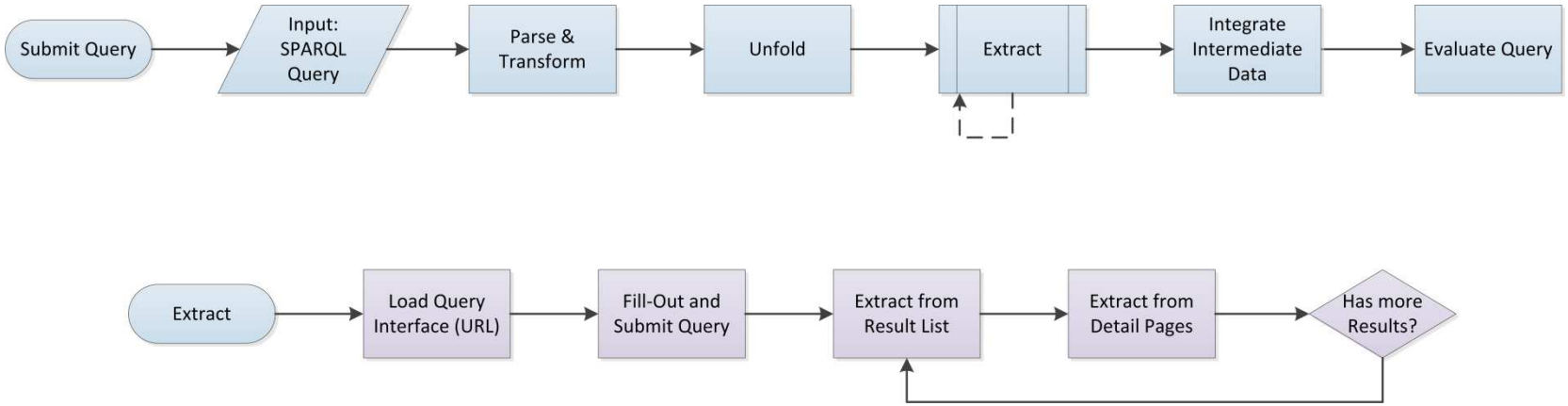
- Query interface for submitting conjunctive queries
- Result list: all valid records, but only key attribute/value pairs

Model Overview

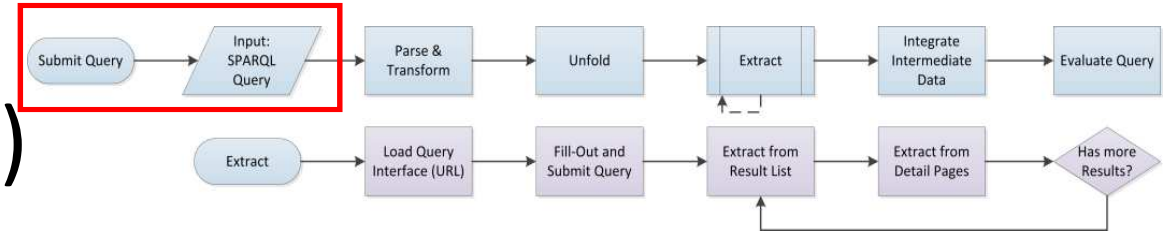


- Query interface for submitting conjunctive queries
- Result list: all valid records, but only key attribute/value pairs
- Result detail: all attribute/value pairs, but only one record

Query Process



Walkthrough (1)



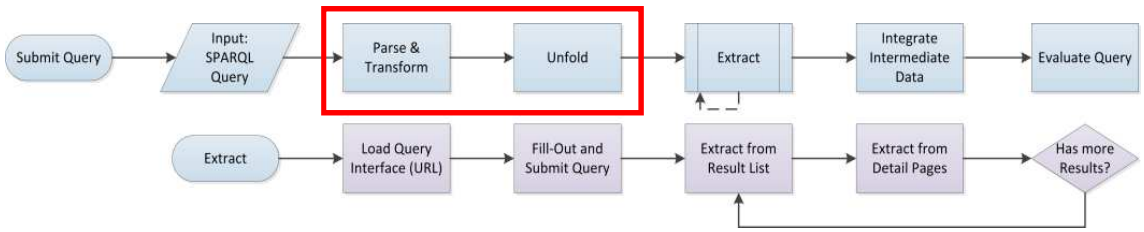
- Query:

  "Return details of real estate offers with a rent between 800€ and 1200€ and either at least 3 rooms or 80m²"

- SPARQL:

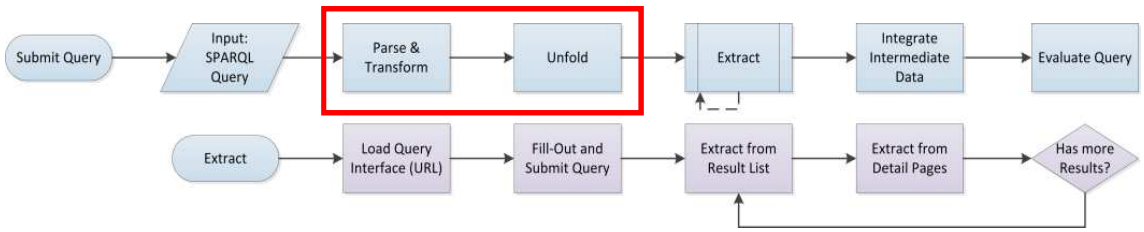
```
SELECT ? townname ?offername ?description ?realestaterent ?realestaterooms ?realestatefloorSpace
WHERE {?object rdf:type realestate:RealEstateOffer;
               realestate:townname ?townname;
               realestate:offername ?offername;
               realestate:description ?description;
               realestate:rent ?realestaterent;
               realestate:rooms ?realestaterooms;
               realestate:floorSpace ?realestatefloorSpace}.
      FILTER (?realestaterent>=800 && ?realestaterent<=1200 &&
             (?realestaterooms>=3 ||  ?realestatefloorSpace >= 80)) }
```

Walkthrough (2)



- Transformation:
  - Parse Query
  - Transform filter to Disjunctive Normal Form and split into subqueries
  - Unfold to include relevant sources via Global-as-View mappings

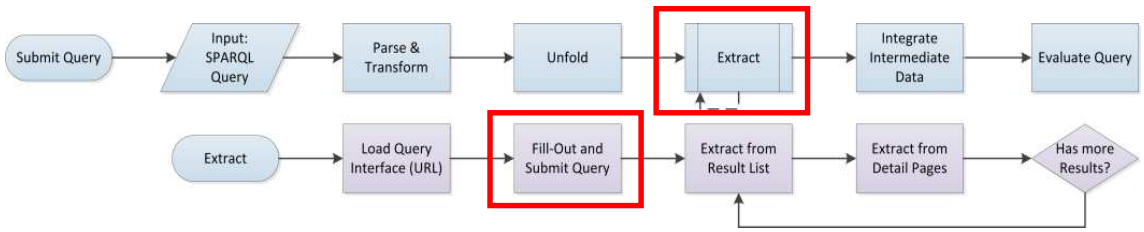
Walkthrough (2)



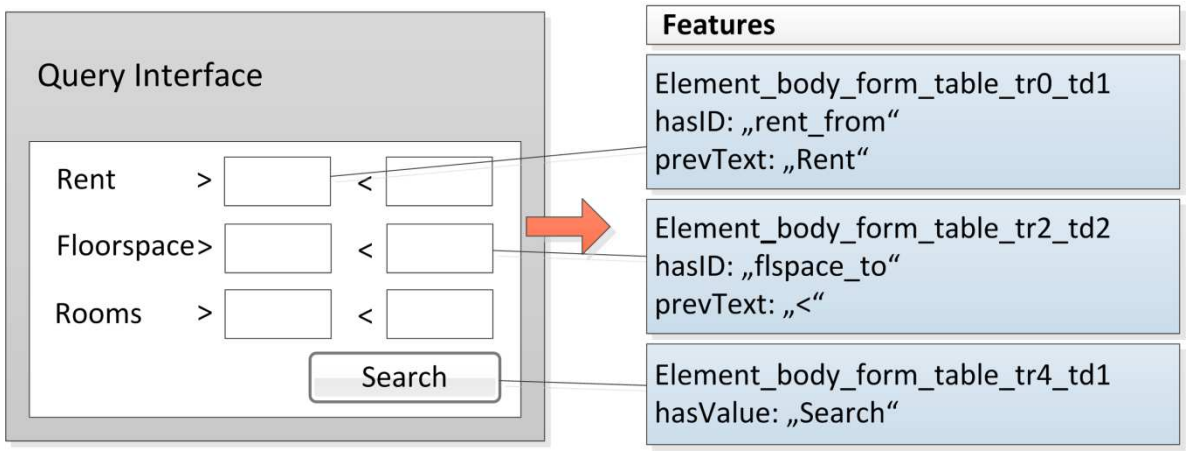
- Transformation:
  - Parse Query
  - Transform filter to Disjunctive Normal Form and split into subqueries
  - Unfold to include relevant sources via Global-as-View mappings
- Result

```
SELECT FROM <http://derStandard.at>: greaterOrEqual(realestate:rent,800) ∧
lessOrEqual(realestate:rent,1200) ∧ greaterOrEqual(realestate:rooms,3)
UNION
SELECT FROM <http://at.immolive24.com>: greaterOrEqual(realestate:rent,800) ∧
lessOrEqual(realestate:rent,1200) ∧ greaterOrEqual(realestate:rooms,3)
UNION
SELECT FROM <http://derStandard.at>: greaterOrEqual(realestate:rent,800) ∧
lessOrEqual(realestate:rent,1200) ∧ greaterOrEqual(realestate:floorSpace,100)
UNION
SELECT FROM <http://at.immolive24.com>: greaterOrEqual(realestate:rent,800) ∧
lessOrEqual(realestate:rent,1200) ∧ greaterOrEqual(realestate:floorSpace,100)
```

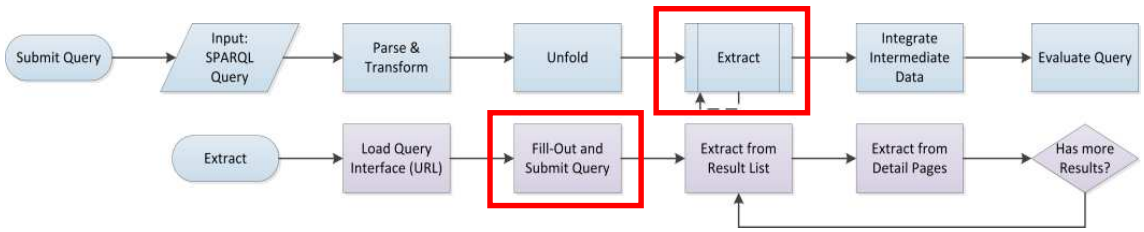
Walkthrough (3)



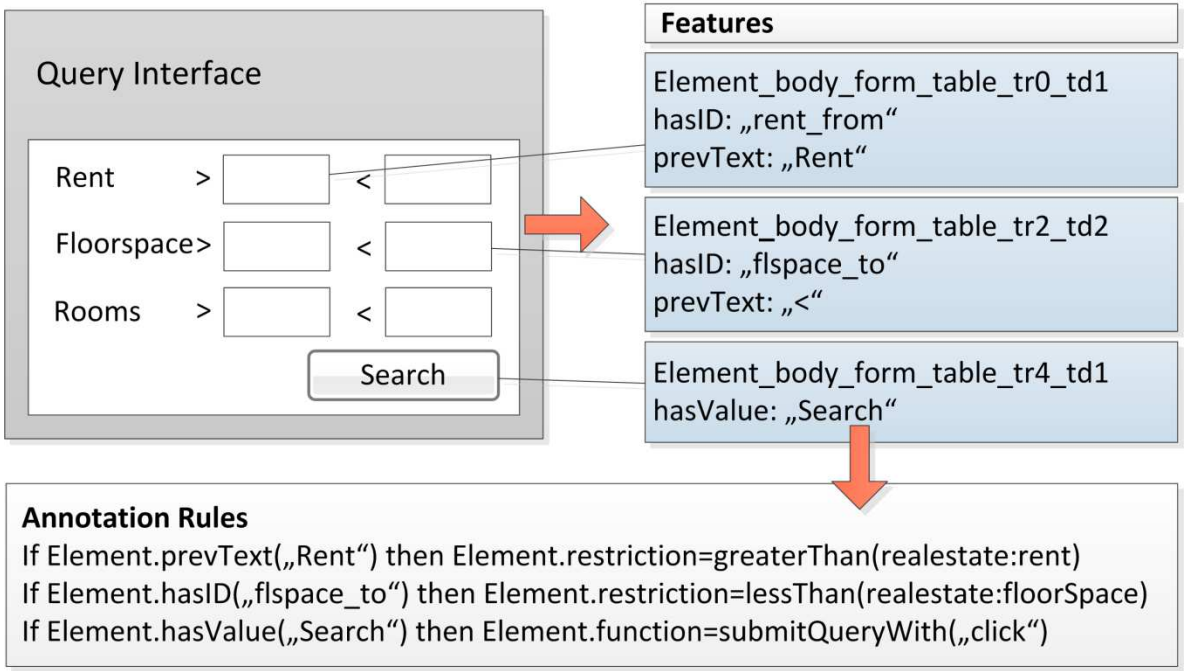
- Features used for identification elements
  - “Properties” of the HTML tags
  - For example id, class, value, tag path, associated text, …



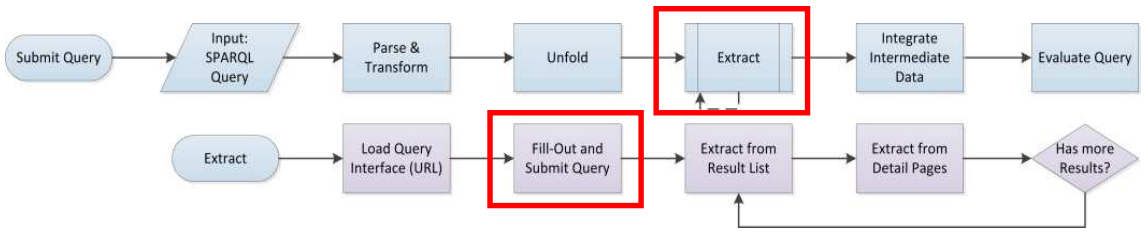
Walkthrough (4)

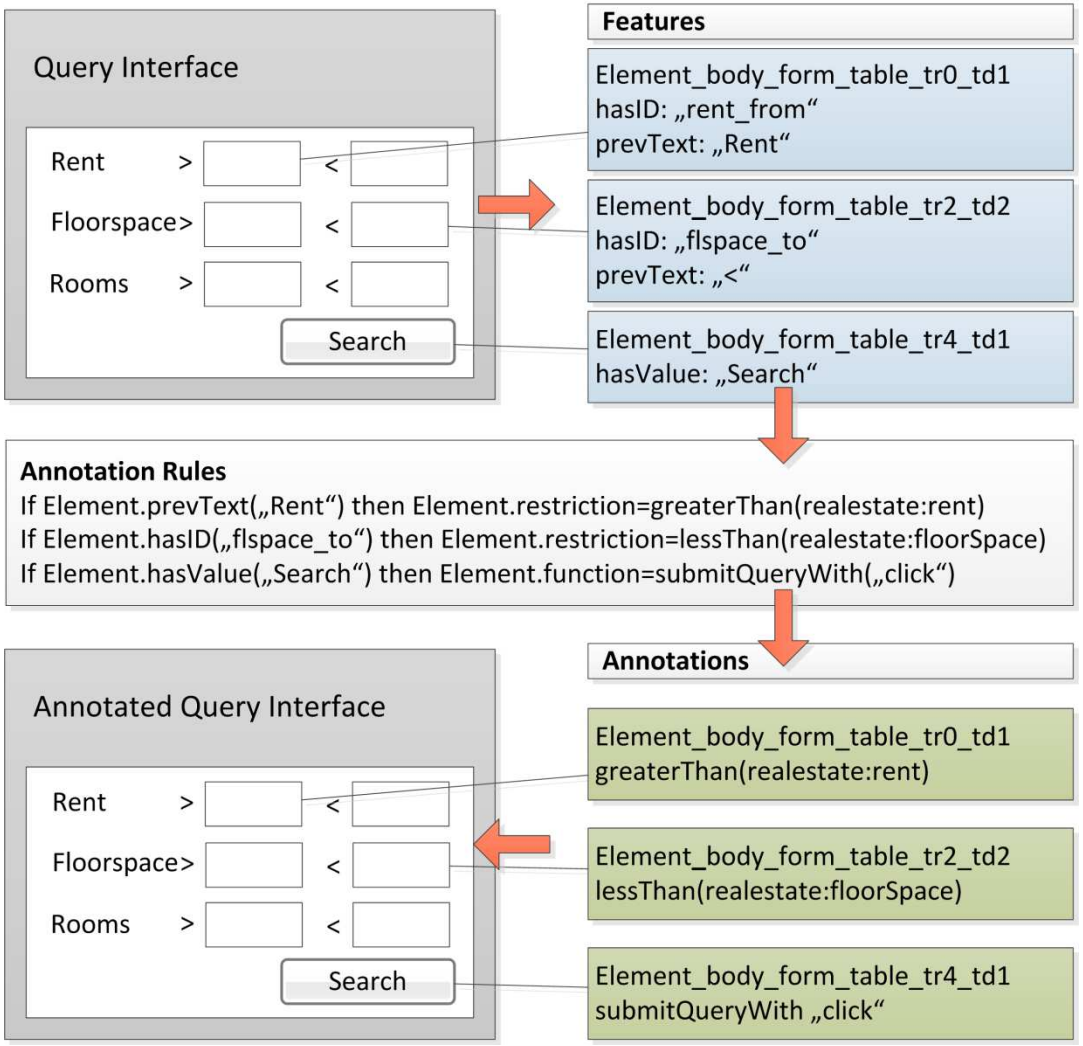


- Object-centered Datalog rules
  - Conditions: conjunction of req. features
  - Conclusions: concepts of the model
- Single efficient evaluation pass

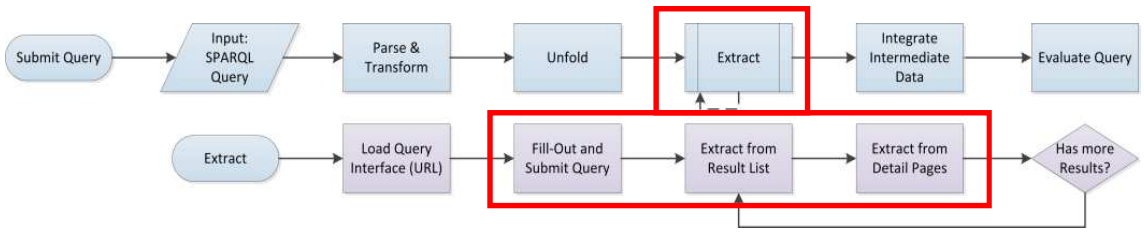


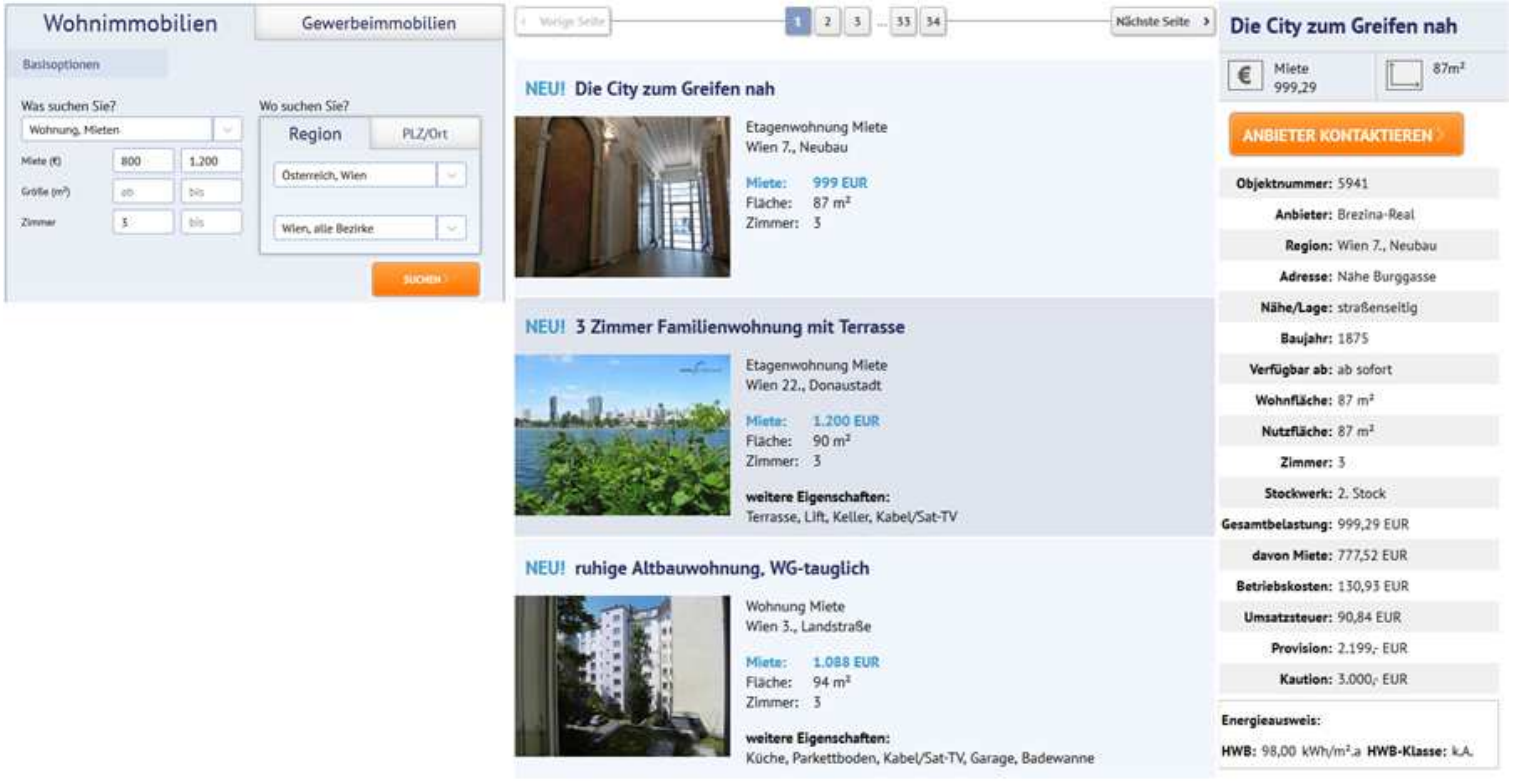
Walkthrough (5)

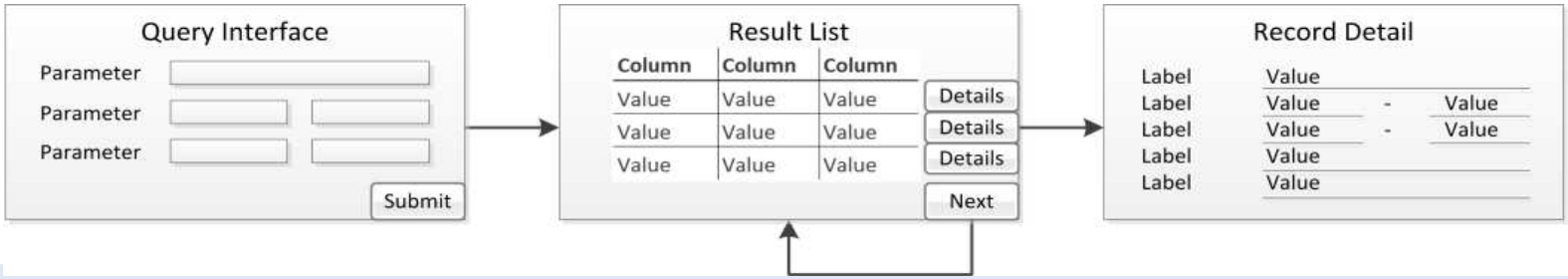




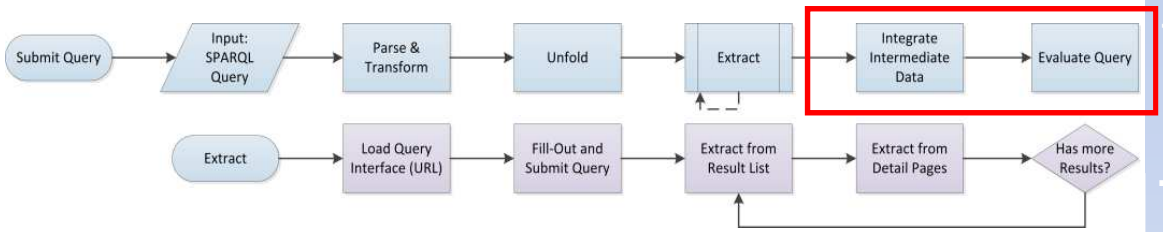
Walkthrough (6)







Walkthrough (7)



- SPARQL

```
SELECT ? townname ?offername ?description ?realestaterent ?realestaterooms ?realestatefloorSpace
WHERE {?object rdf:type realestate:RealEstateOffer;
               realestate:townname ?townname;
               realestate:offername ?offername;
               realestate:description ?description;
               realestate:rent ?realestaterent;
               realestate:rooms ?realestaterooms;
               realestate:floorSpace ?realestatefloorSpace}.
       FILTER (?realestaterent>=800 && ?realestaterent<=1200 &&
              (?realestaterooms>=3 ||  ?realestatefloorSpace >= 80)) }
```

- Output

realestate#townname	realestate#offername	realestate#description	realestate#rent	realestate#rooms	realestate#floorSpace
Naehe Burggasse	Die City zum Greifen nah	Die City zum Greifen nah Das Museumsquartier ums Eck, der 1.Bez...	999,29	3	87
Godlewskigasse	3 Zimmer Familienwohnung mit...	3 Zimmer Familienwohnung mit Terrasse Diese durchdachte Wohn...	1199,75	3	91
	ruhige Altbauwohnung, WG-ta...	ruhige Altbauwohnung, WG-tauglich Sehr schoene Altbauwohnung...	1088	3	95
	ERSTBEZUG: Schicke Neubauw...	ERSTBEZUG: Schicke Neubauwohnung mit Garten Zur Vermietung ...	1150	3	84
Lugeck	Einzigartige Designerwohnung ...	Einzigartige Designerwohnung mit Autoabstellplatz! Einzigartige De...	1184,24	3	107
Wertheimstein-Park	Absolute TOP-Lage - hauseige...	Absolute TOP-Lage - hauseigene Parkanlage! Sanierte 3-Zimmer- ...	990	3	69
Murlingengasse	Exklusiver Erstbezug im Herzen...	Exklusiver Erstbezug im Herzen des 12. Bezirks Exklusiver Erstbezu...	984,16	3	77
Murlingengasse	Exklusives Wohnen im Herzen ...	Exklusives Wohnen im Herzen des 12. Bezirks Exklusives Wohnen i...	945	3	66

Live Presentation of Example Use Cases

Deep Web Mediator and examples available online: http://semann.bdoenz.com/default.aspx

#1 Plain list: Extract the average rent per town from a single site.

#2 Search and result list: Extract test results on cars of the brand Audi from a single site and return brand, model and the test conclusion.

#3 Search, list and detail page: Extract real estate offers from a single site and return details for offers with 3 or more rooms and a rent of 800€ to 1200€.

#4 Disjunctive query: Extract used car offers from a single site and return details of all offers for cars of the brand “Audi” that are priced under 12.500€ if the construction year is after 2011 or under 15.000€ if the construction year is after 2012.

#5 Union: Extract used car offers from all available sites and return details of offers for cars of the brand “Audi” that are priced under 12.500€ and have a construction year after 2011.

#6 Disjunctive union: Extract used car offers from all available sites and return details of all offers for cars of the brand “Audi” that are priced under 12.500€ if the construction year is after 2011 or under 15.000€ if the construction year is after 2012.

#7 Relations between sources: Extract average rents and real estate offers from all available sites and return those that are located in a specific town and have a lower rent/m² than the average for that town.

#8 Deep Web and local databases: Extract real estate offers from all available sites and add the type of town and population from a local dataset.

#9 Deep Web and external databases: Extract real estate offers from all available sites and add a description of the town and the population from an external SPARQL endpoint (dbPedia).

Conclusion

- Processing of queries using a query forwarding approach
  - SPARQL queries as input
  - Query transformation and forwarding via mediators
  - Global-as-View mapping of local sources
- Web form interaction and information extraction
  - Extraction process based on an extensible model
  - Semantic annotations for mapping real-world Web pages to the model
  - Feature-based rules for creating annotations

TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

Institut für
Computertechnik
Institute of
Computer Technology

Extracting Data from the Deep Web with Global-as-View Mediators Using Rule-Enriched Semantic Annotations

Benjamin Dönz
Doenz[at]ict.tuwien.ac.at

Vienna University of Technology
Institute of Computer Technology
Vienna, Austria

Harold Boley
harold.boley[at]unb.ca

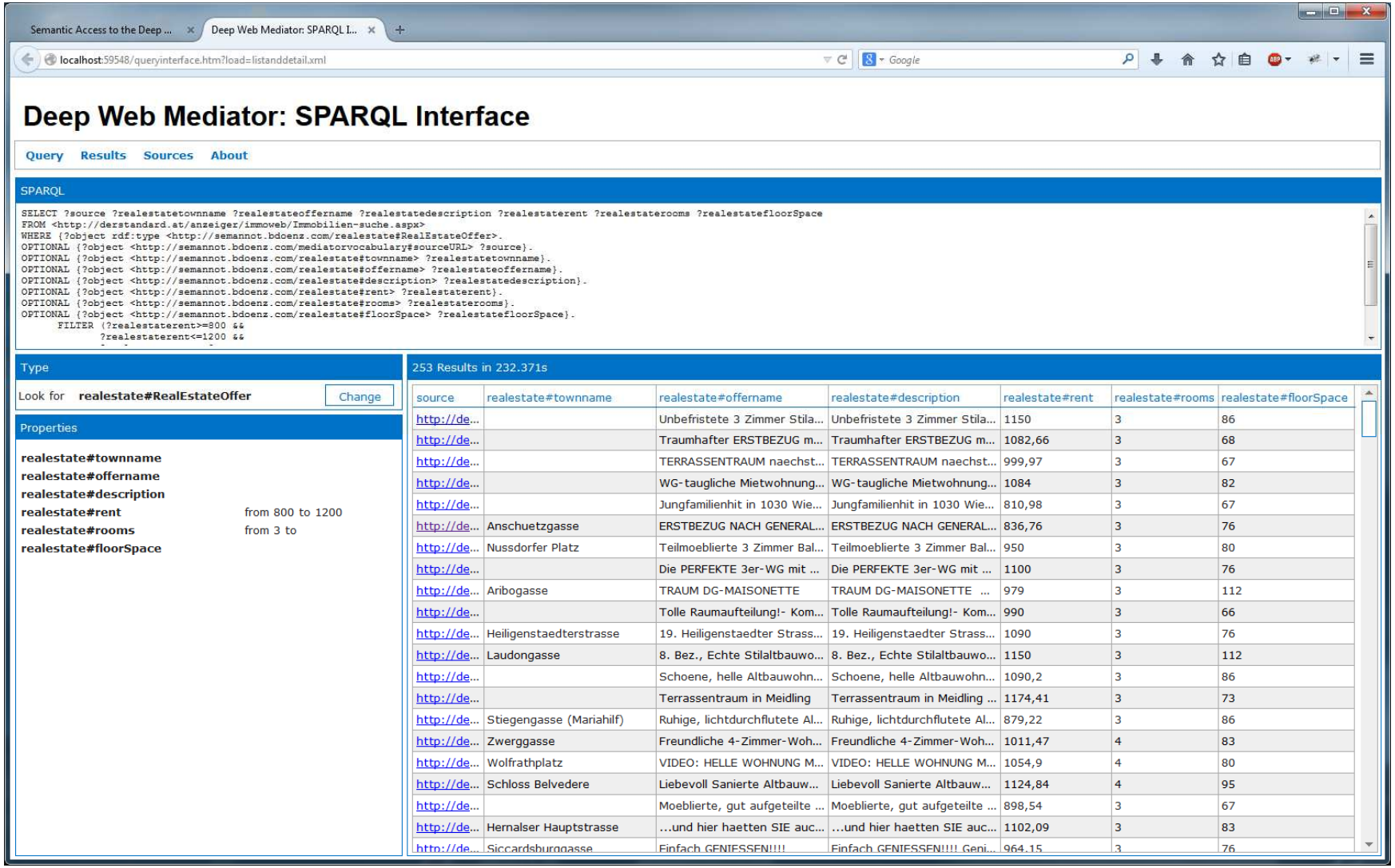
University of New Brunswick
Faculty of Computer Science
Fredericton, NB, Canada

Use case #3

This example is situated in the domain of real estate, and asks for the name of the offer, a description, the number of rooms, the floor space and the rent of all offers with 3 or more rooms and a rent in the range of 800€ and 1200€ from a specific real estate site. To process this query, the mediator accesses the query interface of the site, sets the parameters in the fields of the web form and triggers the search function to submit the query. The returned result lists are iterated to extract the values from the list itself, but also from subpages by following the the corresponding link for each record. All extracted facts are collected in a database and presented to the user in a tabular style including a link to the page where the offer was found.

```
SELECT ?source ?realestatetownname ?realestateoffername ?realestatedescription ?realestaterent
?realestaterooms ?realestatefloorSpace
FROM <http://derstandard.at/anzeiger/immoweb/Immobilien-suche.aspx>
WHERE {?object rdf:type <http://semannot.bdoenz.com/realestate#RealEstateOffer>.
OPTIONAL {?object <http://semannot.bdoenz.com/mediatorvocabulary#sourceURL> ?source}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#townname> ?realestatetownname}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#offername> ?realestateoffername}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#description> ?realestatedescription}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#rent> ?realestaterent}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#rooms> ?realestaterooms}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#floorSpace> ?realestatefloorSpace}.
      FILTER (?realestaterent>=800 &&
             ?realestaterent<=1200 &&
             ?realestaterooms>=3
      )
}
```

Use case #3



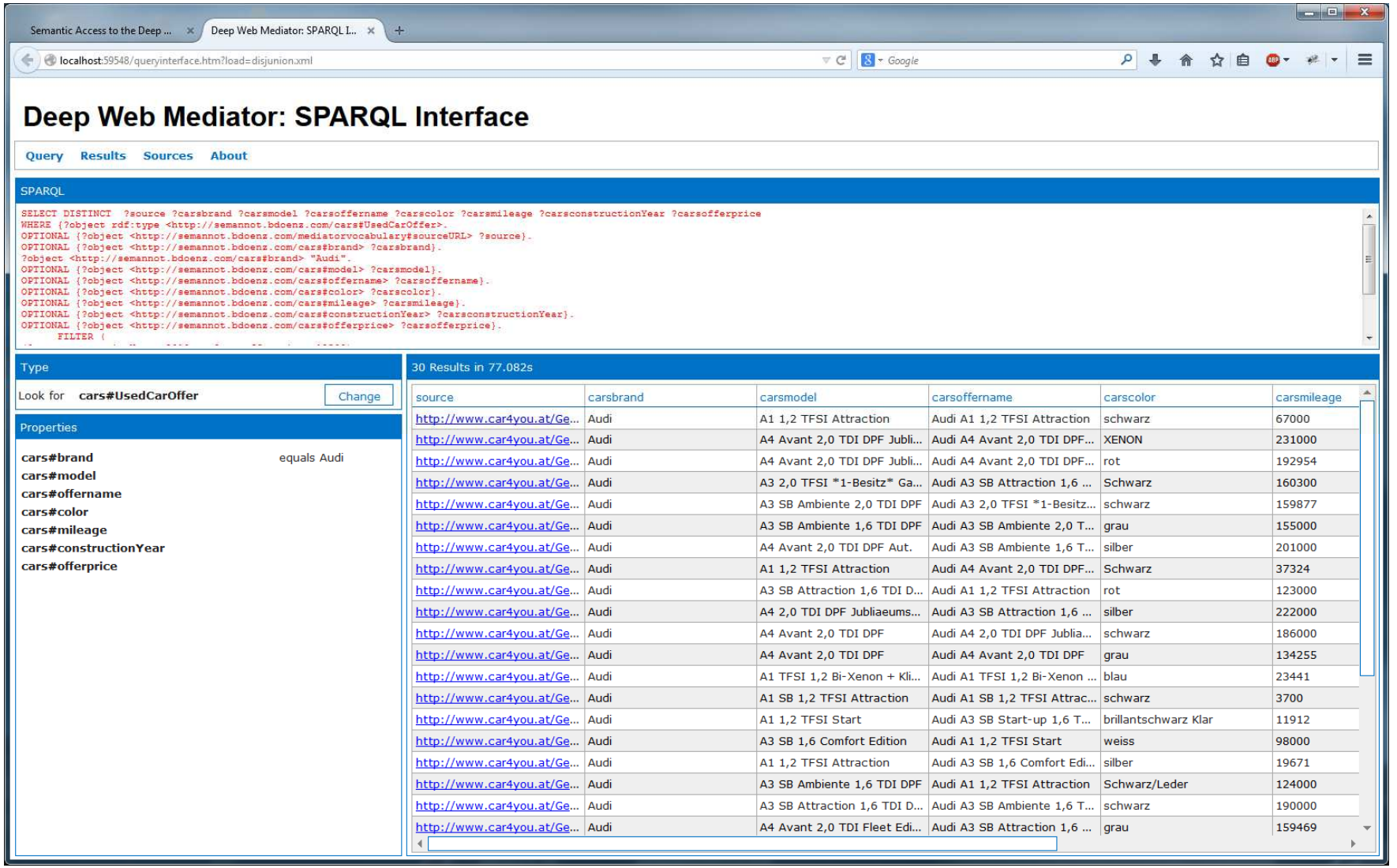
source	realestate#townname	realestate#offername	realestate#description	realestate#rent	realestate#rooms	realestate#floorSpace
http://de...		Unbefristete 3 Zimmer Stila...	Unbefristete 3 Zimmer Stila...	1150	3	86
http://de...		Traumhafter ERSTBEZUG m...	Traumhafter ERSTBEZUG m...	1082,66	3	68
http://de...		TERRASSENTRAUM naechst...	TERRASSENTRAUM naechst...	999,97	3	67
http://de...		WG-taugliche Mietwohnung...	WG-taugliche Mietwohnung...	1084	3	82
http://de...		Jungfamilienhit in 1030 Wie...	Jungfamilienhit in 1030 Wie...	810,98	3	67
http://de...	Anschuetzgasse	ERSTBEZUG NACH GENERAL...	ERSTBEZUG NACH GENERAL...	836,76	3	76
http://de...	Nussdorfer Platz	Teilmoeblierte 3 Zimmer Bal...	Teilmoeblierte 3 Zimmer Bal...	950	3	80
http://de...		Die PERFEKTE 3er-WG mit ...	Die PERFEKTE 3er-WG mit ...	1100	3	76
http://de...	Aribogasse	TRAUM DG-MAISONETTE	TRAUM DG-MAISONETTE ...	979	3	112
http://de...		Tolle Raumaufteilung!- Kom...	Tolle Raumaufteilung!- Kom...	990	3	66
http://de...	Heiligenstaedterstrasse	19. Heiligenstaedter Strass...	19. Heiligenstaedter Strass...	1090	3	76
http://de...	Laudongasse	8. Bez., Echte Stilaltbauwo...	8. Bez., Echte Stilaltbauwo...	1150	3	112
http://de...		Schoene, helle Altbauwohn...	Schoene, helle Altbauwohn...	1090,2	3	86
http://de...		Terrassentraum in Meidling	Terrassentraum in Meidling ...	1174,41	3	73
http://de...	Stiegengasse (Mariahilf)	Ruhige, lichtdurchflutete Al...	Ruhige, lichtdurchflutete Al...	879,22	3	86
http://de...	Zwerggasse	Freundliche 4-Zimmer-Woh...	Freundliche 4-Zimmer-Woh...	1011,47	4	83
http://de...	Wolfrathplatz	VIDEO: HELLE WOHNUNG M...	VIDEO: HELLE WOHNUNG M...	1054,9	4	80
http://de...	Schloss Belvedere	Liebevoll Sanierte Altbauw...	Liebevoll Sanierte Altbauw...	1124,84	4	95
http://de...		Moeblierte, gut aufgeteilte ...	Moeblierte, gut aufgeteilte ...	898,54	3	67
http://de...	Hernalser Hauptstrasse	...und hier haetten SIE auc...	...und hier haetten SIE auc...	1102,09	3	83
http://de...	Siccardsburggasse	Einfach GENIESSEN!!!!	Einfach GENIESSEN!!!! Geni...	964,15	3	76

Use case #6

This example is situated in the domain of used cars and is the combination of the previous two examples: The query requests the name, model, color, construction year, mileage and price of offers from a single site, where the brand of the car is "Audi" and that are priced under 12.500€ if the construction year is after 2010, or under 15.000€ if the construction year is after 2011. This query is split into two conjunctive subqueries and submit to all three available sites returning the union of a total of 6 queries.

```
SELECT DISTINCT  ?source ?carsbrand ?carsmodel ?carsoffername ?carscolor ?carsmileage
?carsconstructionYear ?carsofferprice
WHERE {?object rdf:type <http://semannot.bdoenz.com/cars#UsedCarOffer>.
OPTIONAL {?object <http://semannot.bdoenz.com/mediatorvocabulary#sourceURL> ?source}.
OPTIONAL {?object <http://semannot.bdoenz.com/cars#brand> ?carsbrand}.
?object <http://semannot.bdoenz.com/cars#brand> "Audi".
OPTIONAL {?object <http://semannot.bdoenz.com/cars#model> ?carsmodel}.
OPTIONAL {?object <http://semannot.bdoenz.com/cars#offername> ?carsoffername}.
OPTIONAL {?object <http://semannot.bdoenz.com/cars#color> ?carscolor}.
OPTIONAL {?object <http://semannot.bdoenz.com/cars#mileage> ?carsmileage}.
OPTIONAL {?object <http://semannot.bdoenz.com/cars#constructionYear> ?carsconstructionYear}.
OPTIONAL {?object <http://semannot.bdoenz.com/cars#offerprice> ?carsofferprice}.
      FILTER (
(?carsconstructionYear>=2010 && ?carsofferprice<=12500)
||
(?carsconstructionYear>=2011 && ?carsofferprice<=15000)
)
}
```

Use case #6

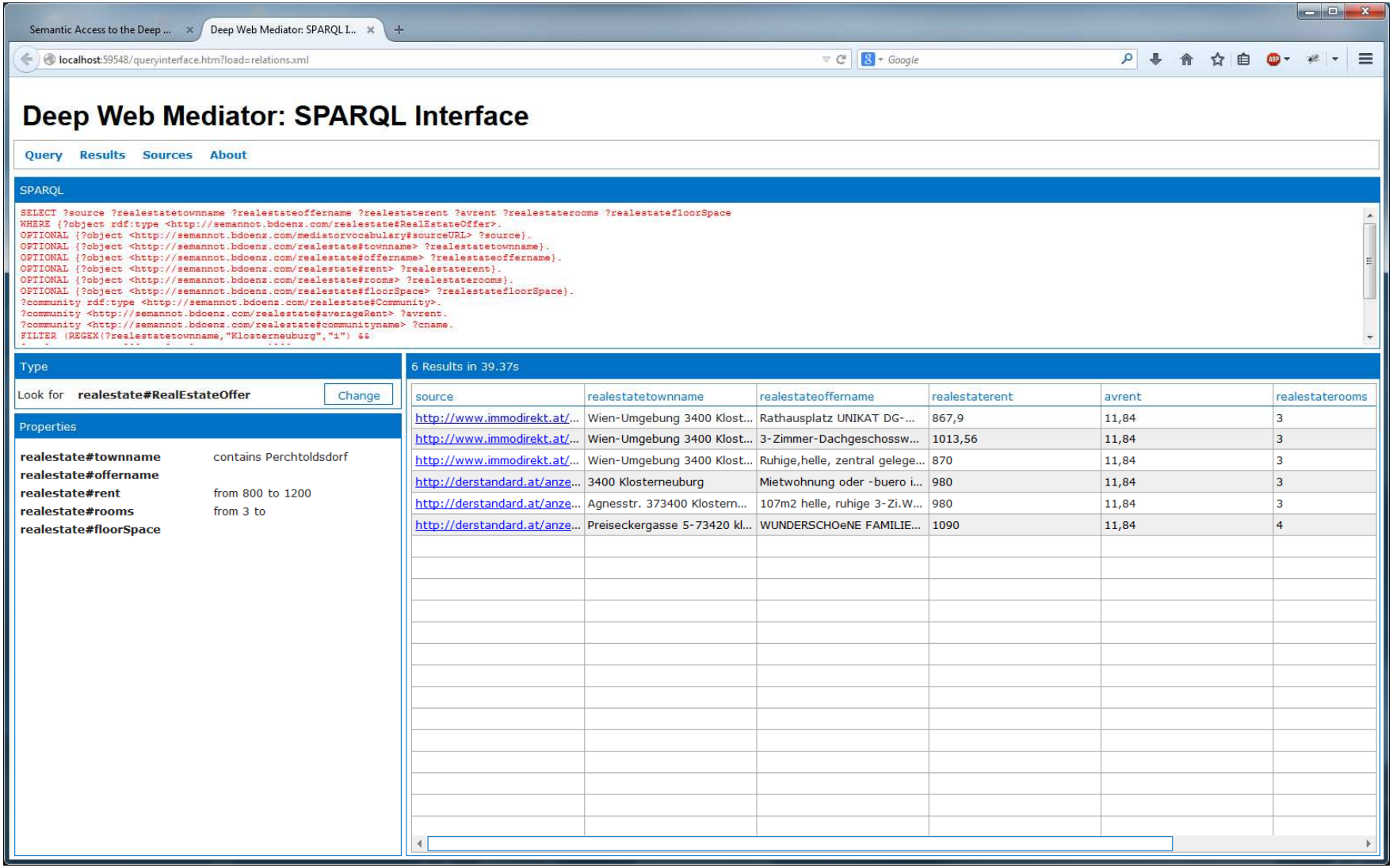


Use case #7

This example is situated in the domain of real estate, and asks for the name of the offer, a description, the number of rooms, the floor space and the rent of all offers with 3 or more rooms in the town of "Klosterneuburg" where the rent is lower than the average rent per square meter for that town. No specific site is referenced in the query, the mediator therefore includes all sites with real estate offers and also a s site containig average rents. Each of these are accessed in turn collecting the intermediate results in a database before applying the filter and returning the results. Intermediate results are only available after the average rents have been extracted and can be compared to the offers in the defined manner. Note that this type of query cannot be generated by the query wizard, but is entered directly as SPARQL

```
SELECT ?source ?realestatetownname ?realestateoffername ?realestaterent ?realestaterooms
?realestatefloorSpace
WHERE {?object rdf:type <http://semannot.bdoenz.com/realestate#RealEstateOffer>.
OPTIONAL {?object <http://semannot.bdoenz.com/mediatorvocabulary#sourceURL> ?source}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#townname> ?realestatetownname}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#offername> ?realestateoffername}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#rent> ?realestaterent}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#rooms> ?realestaterooms}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#floorSpace> ?realestatefloorSpace}.
?community rdf:type <http://semannot.bdoenz.com/realestate#Community>.
?community <http://semannot.bdoenz.com/realestate#averageRent> ?avrent.
?community <http://semannot.bdoenz.com/realestate#communityname> ?cname.
FILTER (REGEX(?realestatetownname,"Klosterneuburg","i") &&
?realestaterent>=800 && ?realestaterent<=1200 &&
?realestaterooms>=3 && REGEX(?cname,"Klosterneuburg","i") &&
?realestaterent/?realestatefloorSpace <(?avrent))}
```

Use case #7


Semantic Access to the Deep ...
Deep Web Mediator: SPARQL I...
localhost:59548/queryinterface.htm?load=relations.xml
Deep Web Mediator: SPARQL Interface
Query Results Sources About
SPARQL
SELECT ?source ?realestatetownname ?realestateoffername ?realestaterent ?avrent ?realestaterooms ?realestatefloorSpace
WHERE {?object rdf:type <http://semannot.bdoenz.com/realestate#RealEstateOffer>.
OPTIONAL {?object <http://semannot.bdoenz.com/mediatorvocabulary#sourceURL> ?source}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#townname> ?realestatetownname}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#offername> ?realestateoffername}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#rent> ?realestaterent}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#rooms> ?realestaterooms}.
OPTIONAL {?object <http://semannot.bdoenz.com/realestate#floorSpace> ?realestatefloorSpace}.
?community rdf:type <http://semannot.bdoenz.com/realestate#Community>.
?community <http://semannot.bdoenz.com/realestate#averageRent> ?avrent.
?community <http://semannot.bdoenz.com/realestate#communityname> ?cname.
FILTER (REGEX(?realestatetownname,"Klosterneuburg","i") &&
Type
Look for realestate#RealEstateOffer
Change
Properties
realestate#townname contains Perchtoldsdorf
realestate#offername
realestate#rent from 800 to 1200
realestate#rooms from 3 to
realestate#floorSpace
6 Results in 39.37s
source
realestatetownname
realestateoffername
realestaterent
avrent
realestaterooms
http://www.immodirekt.at/... Wien-Umgebung 3400 Klost... Rathausplatz UNIKAT DG-... 867,9 11,84 3
http://www.immodirekt.at/... Wien-Umgebung 3400 Klost... 3-Zimmer-Dachgeschossw... 1013,56 11,84 3
http://www.immodirekt.at/... Wien-Umgebung 3400 Klost... Ruhige,helle, zentral gelege... 870 11,84 3
http://derstandard.at/anze... 3400 Klosterneuburg Mietwohnung oder -buero i... 980 11,84 3
http://derstandard.at/anze... Agnesstr. 373400 Klostern... 107m2 helle, ruhige 3-Zi.W... 980 11,84 3
http://derstandard.at/anze... Preiseckergasse 5-73420 kl... WUNDERSCHOeNE FAMILIE... 1090 11,84 4