universität innsbruck

Department of Computer Science

STI INNSBRUCK

SE Research Seminar: Knowledge Graphs Knowledge Assessment

Midterm Presentation

By Penz Manuel, Rasmusen Sven

Agenda

- Recap
  - Hosting
  - Finding external sources
  - Metrics & Dimensions definition
  - External Source Conclusion
- Web Scraper
- Mapping
- Assessment
- Loading into GTKG
- Data source comparison

Recap

What have we done so far?

Hosting

- **Add new repository**

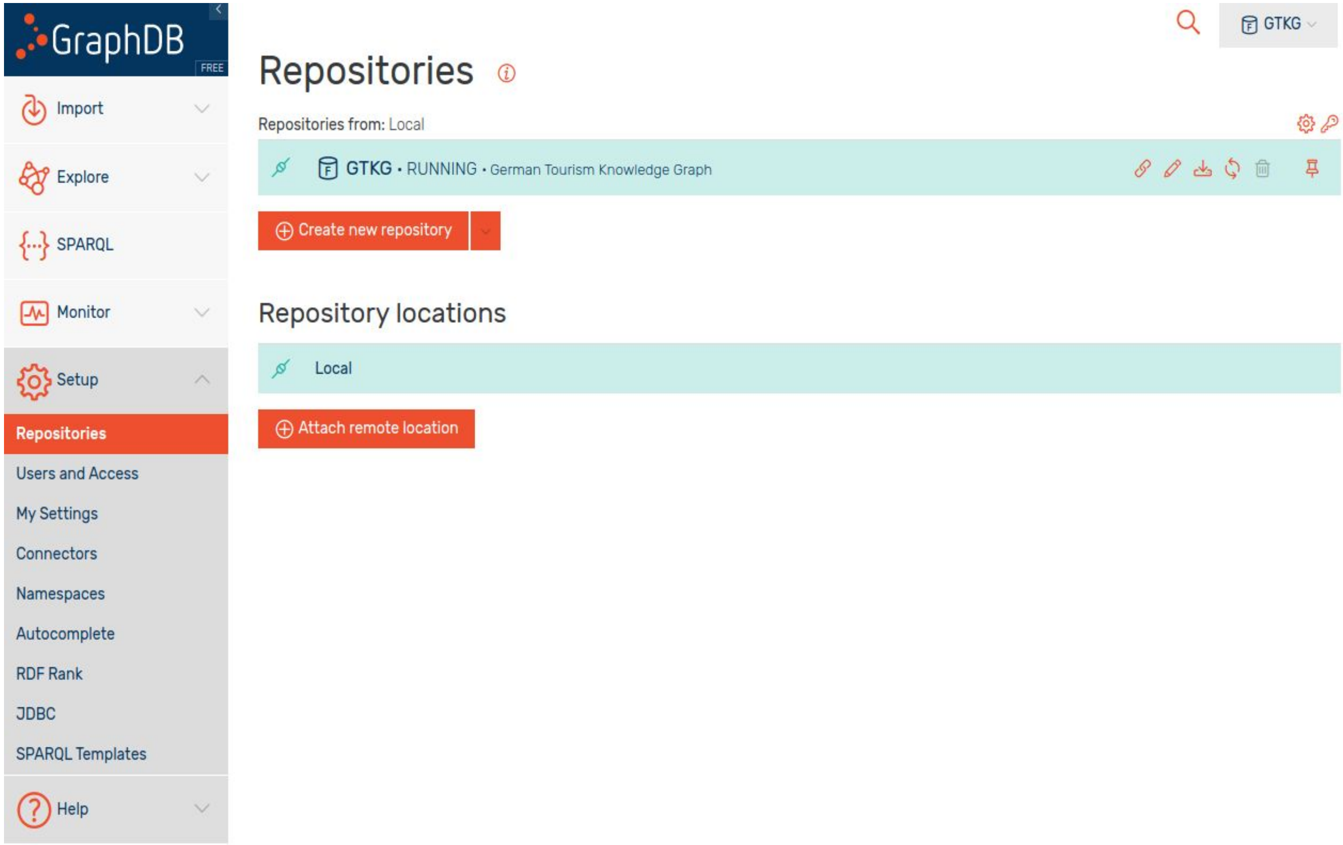


Figure 1: The GraphDB view of the GTKG repository.

Hosting

- **Import RDF as „Server files“**
  - **User data limited to 200MB**

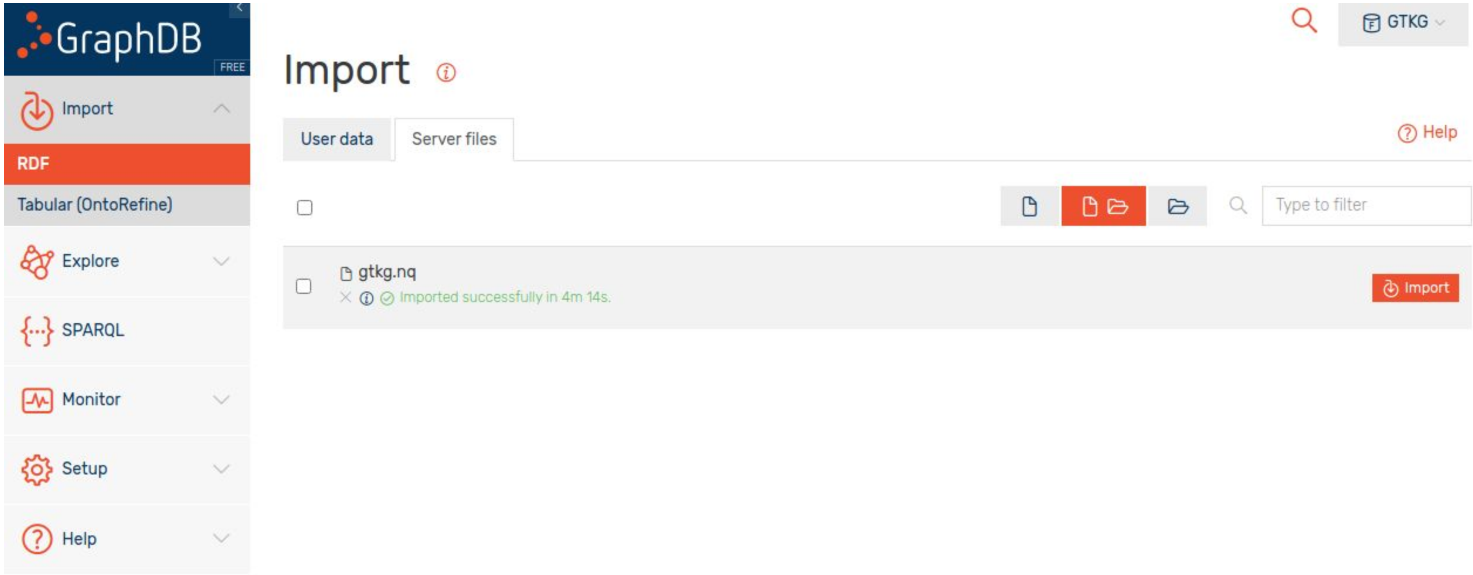
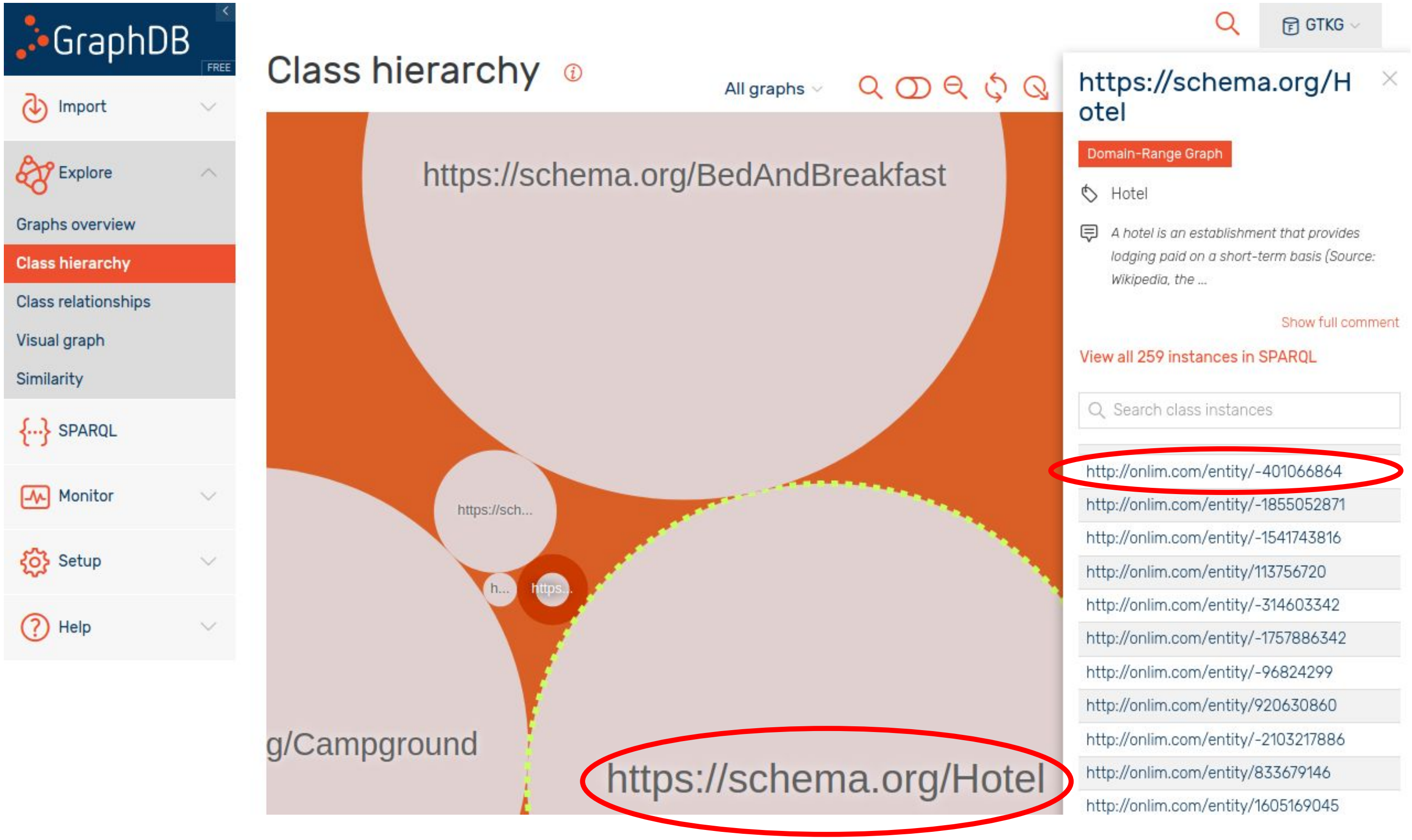


Figure 2: Importing the GTKG data into GraphDB.

Hosting



https://schema.org/Hotel

Domain-Range Graph

Hotel

A hotel is an establishment that provides lodging paid on a short-term basis (Source: Wikipedia, the ...

Show full comment

View all 259 instances in SPARQL

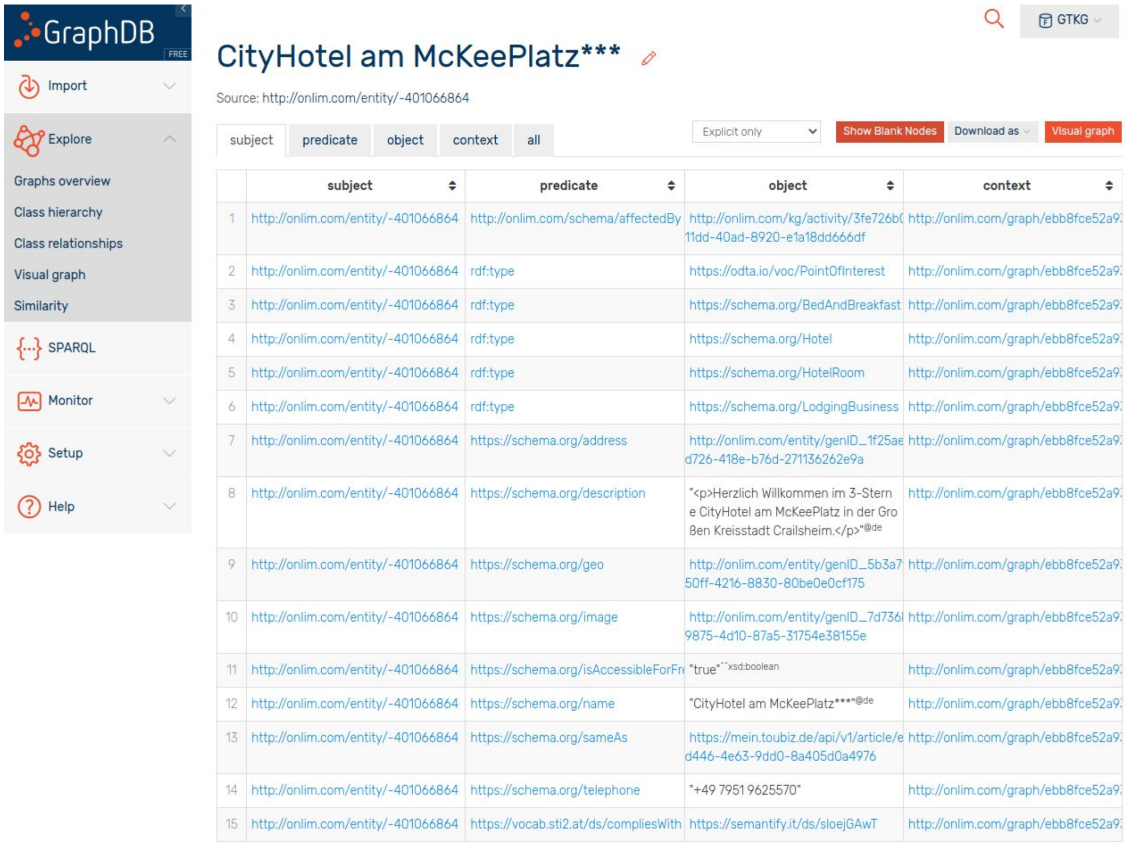
Search class instances

http://onlim.com/entity/-401066864
http://onlim.com/entity/-1855052871
http://onlim.com/entity/-1541743816
http://onlim.com/entity/113756720
http://onlim.com/entity/-314603342
http://onlim.com/entity/-1757886342
http://onlim.com/entity/-96824299
http://onlim.com/entity/920630860
http://onlim.com/entity/-2103217886
http://onlim.com/entity/833679146
http://onlim.com/entity/1605169045

Figure 3: The class hierarchy visualisation of GTKG displaying the existing hotel instances.

Hosting

- **Properties:**
  - **address**
  - **description**
  - **geo**
  - **image**
  - **isAccessibleForFree**
  - **name**
  - **sameAs**
  - **telephone**
  - **compliesWith**
- **Missing Properties:**
  - **rating**
  - **reviews**



CityHotel am McKeePlatz***

Source: http://onlim.com/entity/-401066864

	subject	predicate	object	context
1	http://onlim.com/entity/-401066864	http://onlim.com/schema/affectedBy	http://onlim.com/kg/activity/3fe726b11dd-40ad-8920-e1a18dd666df	http://onlim.com/graph/ebb8fce52a9
2	http://onlim.com/entity/-401066864	rdf:type	https://odta.io/voc/PointOfInterest	http://onlim.com/graph/ebb8fce52a9
3	http://onlim.com/entity/-401066864	rdf:type	https://schema.org/BedAndBreakfast	http://onlim.com/graph/ebb8fce52a9
4	http://onlim.com/entity/-401066864	rdf:type	https://schema.org/Hotel	http://onlim.com/graph/ebb8fce52a9
5	http://onlim.com/entity/-401066864	rdf:type	https://schema.org/HotelRoom	http://onlim.com/graph/ebb8fce52a9
6	http://onlim.com/entity/-401066864	rdf:type	https://schema.org/LodgingBusiness	http://onlim.com/graph/ebb8fce52a9
7	http://onlim.com/entity/-401066864	https://schema.org/address	http://onlim.com/entity/genID_1f25aed726-418e-b76d-271136262e9a	http://onlim.com/graph/ebb8fce52a9
8	http://onlim.com/entity/-401066864	https://schema.org/description	"<p>Herzlich Willkommen im 3-Sterne CityHotel am McKeePlatz in der Großen Kreisstadt Crailsheim.</p>"@de	http://onlim.com/graph/ebb8fce52a9
9	http://onlim.com/entity/-401066864	https://schema.org/geo	http://onlim.com/entity/genID_5b3a750ff-4216-8830-80be0e0cf175	http://onlim.com/graph/ebb8fce52a9
10	http://onlim.com/entity/-401066864	https://schema.org/image	http://onlim.com/entity/genID_7d7369875-4d10-87a5-31754e38155e	http://onlim.com/graph/ebb8fce52a9
11	http://onlim.com/entity/-401066864	https://schema.org/isAccessibleForFr	"true"^^xsd:boolean	http://onlim.com/graph/ebb8fce52a9
12	http://onlim.com/entity/-401066864	https://schema.org/name	"CityHotel am McKeePlatz***"@de	http://onlim.com/graph/ebb8fce52a9
13	http://onlim.com/entity/-401066864	https://schema.org/sameAs	https://mein.toubiz.de/api/v1/article/ed446-4e63-9dd0-8a405d0a4976	http://onlim.com/graph/ebb8fce52a9
14	http://onlim.com/entity/-401066864	https://schema.org/telephone	"+49 7951 9625570"	http://onlim.com/graph/ebb8fce52a9
15	http://onlim.com/entity/-401066864	https://vocab.sti2.at/ds/compliesWith	https://semantify.it/ds/sloejGAwT	http://onlim.com/graph/ebb8fce52a9

Figure 4: A detailed view of a specific hotel instance within the GTKG.

External Sources

- **www.booking.com:**
  - **API:**
    - **Guest Review API -** https://connect.booking.com/user_guide/site/en-US/review-api/
  - **Offers:**
    - **Guest Reviews**
- **www.tripadvisor.com**
  - **API:**
    - ***Content API -*** *https://developer-tripadvisor.com/content-api/*
  - **Offers:**
    - **Location ID, name, address, latitude & longitude**
    - **Read reviews link, write-a-review link**
    - **Overall rating, ranking, subratings, awards, the number of reviews the rating is based on, rating bubbles image**
    - **Price level symbol, accommodation category/subcategory, attraction type, restaurant cuisine(s)**

External Sources

- **www.govdata.de:**
  - **SPARQL endpoint available!**
  - **Offers:**
    - **Various statistical data collections regarding hotels / tourism**
- **www.google.at/maps:**
  - **API:**
    - ***Maps API - https://developers.google.com/maps***
  - **Offers:**
    - **Places API (Nearby businesses, customer reviews, etc.)**
    - **Geocoding API (longitude, latitude)**
    - **JSON data**

External Sources

- **www.wikidata.org:**
  - **SPARQL endpoint available!**
  - **Offers:**
    - **JSON, CSV and TSV data**
- **www.dbpedia.org:**
  - **SPARQL endpoint available!**
  - **Offers:**
    - **JSON, RDF, CSV and many more**

Dimensions & Metrics

- **Accessibility**
  - **Provisioning of public endpoint**
    - Weight = 0.45
    - 1 If SPARQL and REST API
    - 0.75 If SPARQL or REST API endpoint is publicly available
    - 0.5 any form of offline data
    - 0 Otherwise
  - **Retrievable format**
    - Weight = 0.45
    - 1 If RDF export available
    - 0.75 If JSON export available
    - 0.5 If semi-structured data available
    - 0 Otherwise
  - **Content negotiation**
    - Weight = 0.1
    - 1 If content negotiation is supported
    - 0 Otherwise

Dimensions & Metrics

- **Completeness**
  - **Instance completeness**
    - Weight = 0.5
    - $$m = \frac{1}{N} \Sigma \frac{number\ of\ values\ from\ classes\ \&\ properties\ in\ instance\ in\ subset}{number\ of\ total\ values\ from\ \ classes\ \&\ properties\ according\ to\ DS};\ N \ldots\ subset\ size$$
  - **Population completeness**
    - Weight = 0.5
    - $$m = \frac{number\ of\ objects\ per\ domain\ represented\ in\ the\ data\ source}{total\ number\ objects\ per\ domain}$$

Dimensions & Metrics

- **Accuracy**
  - **Formal Semantic validity**
    - Weight = 0.5
    - $$m = \frac{|\{o \mid (s,p,o) \in r \wedge o \in L \wedge semValid(o)\}|}{|\{o \mid (s,p,o) \in r \wedge o \in L\}|}$$
  - *semValid()* rule examples:
    - Postal Code:
      - length of 5
      - starting from 01 to 99
    - Phone number:
      - start with +49 or 0049 followed by a valid area code
        - starting from 02 to 09
        - total length between 3 and 5

Dimensions & Metrics

- **Accuracy**
  - **Formal Syntactic validity**
    - Weight = 0.5
    - $$m = \frac{|\{o \mid (s,p,o) \in r \land o \in L \land synValid(o)\}|}{|\{o \mid (s,p,o) \in r \land o \in L\}|}$$
  - *synValid()* rule examples:
    - Postal Code:
      - given: https://schema.org/Text
      - needed: https://schema.org/Integer
    - Address:
      - If a lodging business is present in the GTKG and in the data source then we use the Levenshtein distance to calculate the difference between two strings/numbers in order to detect syntactic errors.

Conclusion to External Sources

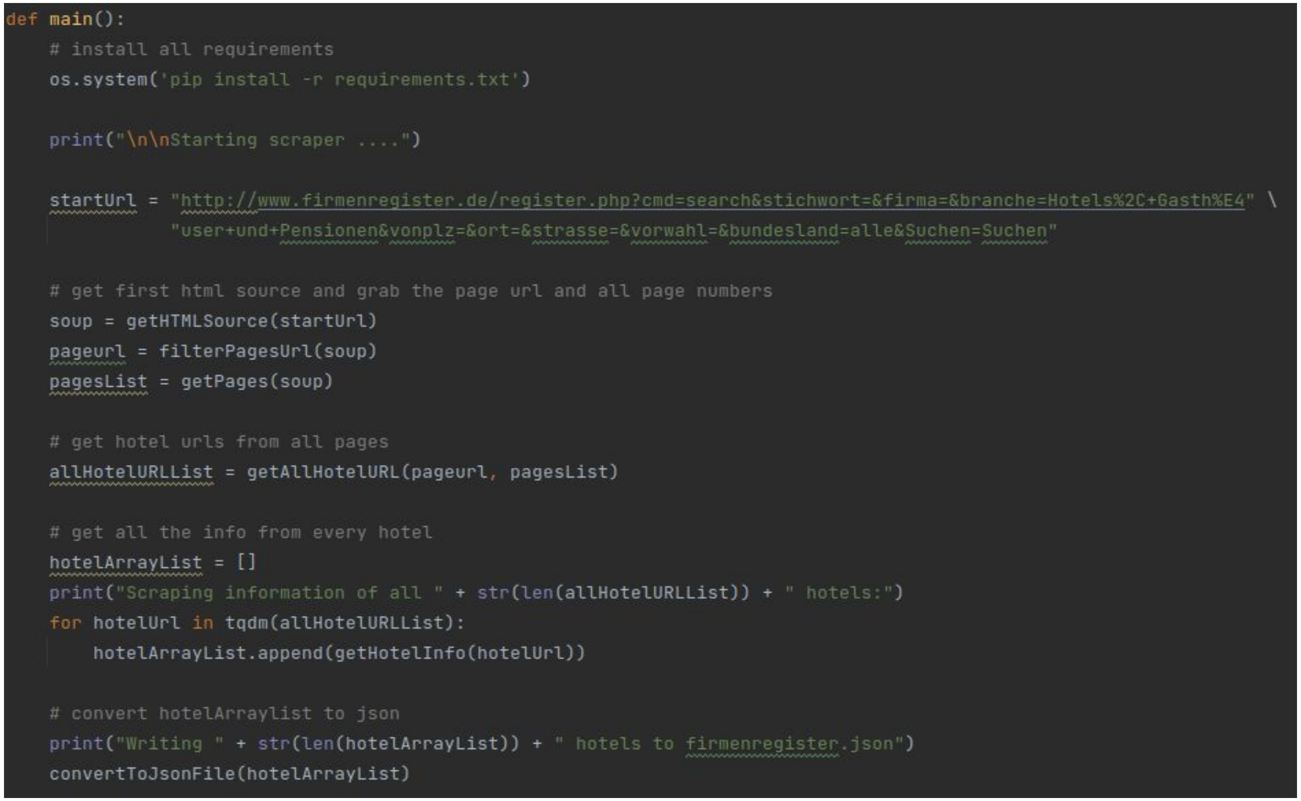
- **www.wikidata.org:**
  - **SPARQL endpoint available!**
- **www.dbpedia.org:**
  - **SPARQL endpoint available!**
- **www.firmenregister.de**
  - **No endpoint available!**
  - **Solution:**
    - **Scrape website**

Data source: www.firmenregister.de

Building a web scraper, schema alignment, mapping, and assessment.

Scraper

- **Python**
  - **BeautifulSoup**
  - **Proxy Server**

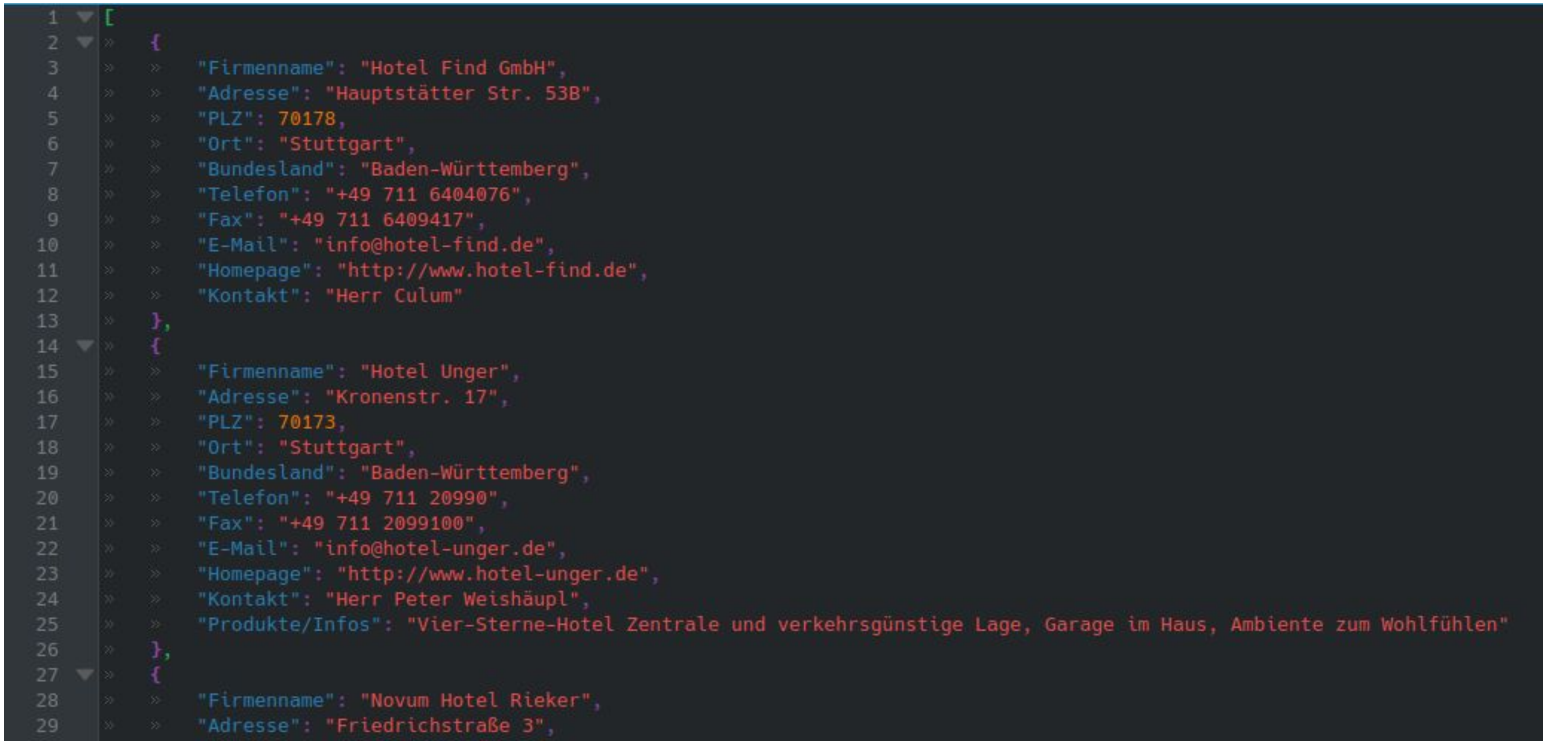


```python
def main():
    # install all requirements
    os.system('pip install -r requirements.txt')

    print("\n\nStarting scraper ....")

    startUrl = "http://www.firmenregister.de/register.php?cmd=search&stichwort=&firma=&branche=Hotels%2C+Gasth%E4" \
               "user+und+Pensionen&vonplz=&ort=&strasse=&vorwahl=&bundesland=alle&Suchen=Suchen"

    # get first html source and grab the page url and all page numbers
    soup = getHTMLSource(startUrl)
    pageurl = filterPagesUrl(soup)
    pagesList = getPages(soup)

    # get hotel urls from all pages
    allHotelURLList = getAllHotelURL(pageurl, pagesList)

    # get all the info from every hotel
    hotelArrayList = []
    print("Scraping information of all " + str(len(allHotelURLList)) + " hotels:")
    for hotelUrl in tqdm(allHotelURLList):
        hotelArrayList.append(getHotelInfo(hotelUrl))

    # convert hotelArraylist to json
    print("Writing " + str(len(hotelArrayList)) + " hotels to firmenregister.json")
    convertToJsonFile(hotelArrayList)
```

Figure 5: A python code snippet of the web scraper to gather information from www.firmenregister.de .

Scraper

- **Exports data of almost 9k german lodging businesses into JSON file**
    - **Advantage:**
        - **pyRML can handle .json file mapping into .n3 file**

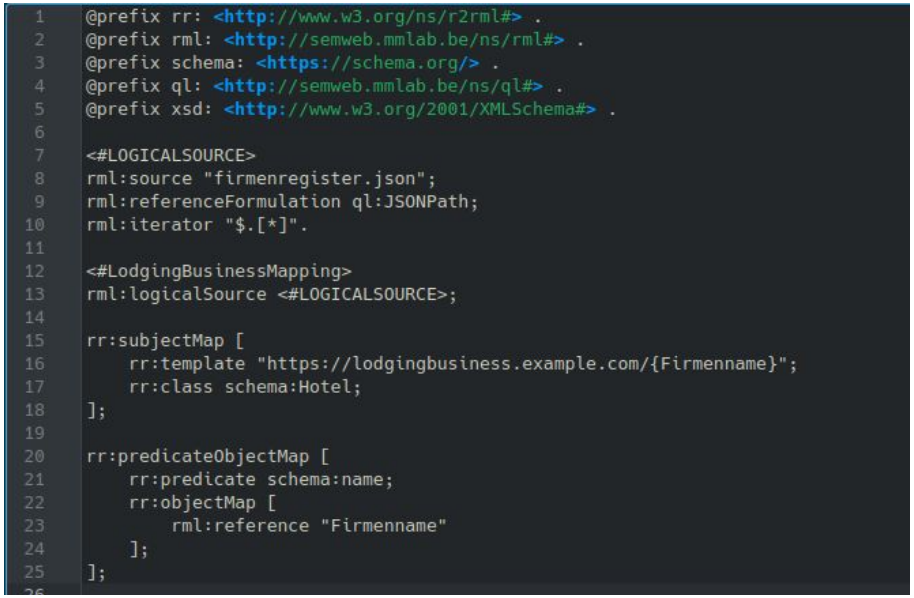


```
[
    {
        "Firmenname": "Hotel Find GmbH",
        "Adresse": "Hauptstätter Str. 53B",
        "PLZ": 70178,
        "Ort": "Stuttgart",
        "Bundesland": "Baden-Württemberg",
        "Telefon": "+49 711 6404076",
        "Fax": "+49 711 6409417",
        "E-Mail": "info@hotel-find.de",
        "Homepage": "http://www.hotel-find.de",
        "Kontakt": "Herr Culum"
    },
    {
        "Firmenname": "Hotel Unger",
        "Adresse": "Kronenstr. 17",
        "PLZ": 70173,
        "Ort": "Stuttgart",
        "Bundesland": "Baden-Württemberg",
        "Telefon": "+49 711 20990",
        "Fax": "+49 711 2099100",
        "E-Mail": "info@hotel-unger.de",
        "Homepage": "http://www.hotel-unger.de",
        "Kontakt": "Herr Peter Weishäupl",
        "Produkte/Infos": "Vier-Sterne-Hotel Zentrale und verkehrsgünstige Lage, Garage im Haus, Ambiente zum Wohlfühlen"
    },
    {
        "Firmenname": "Novum Hotel Rieker",
        "Adresse": "Friedrichstraße 3",
```

Figure 6: The JSON data received after scraping www.firmenregister.de .

Mapping - firmenregister.de

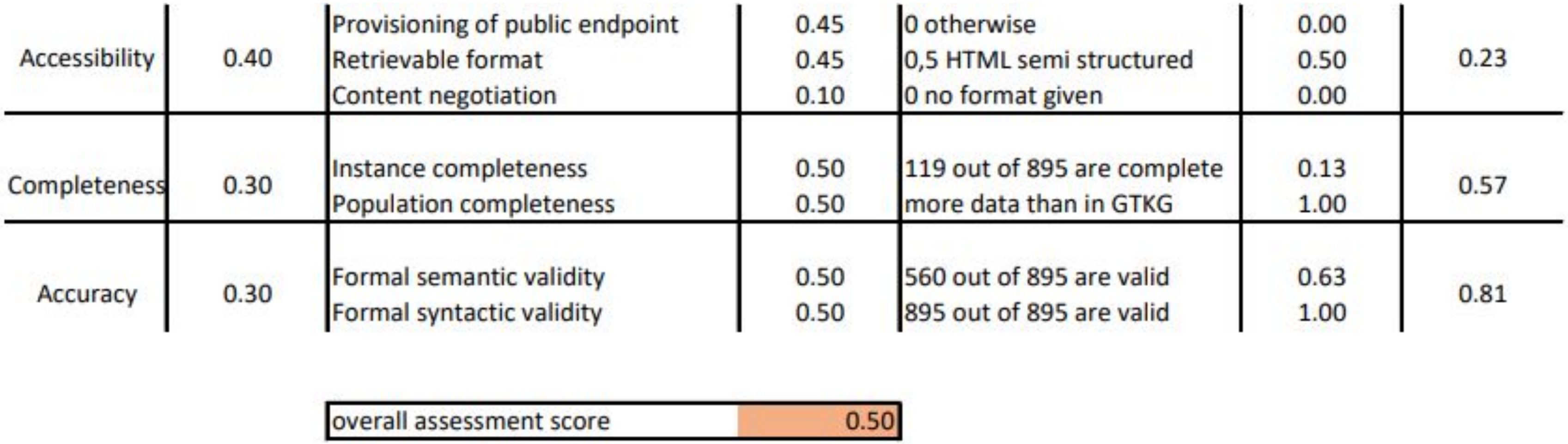
- **Properties:**
  - **name - “Firmenname”**
  - **telephone - “Telefon”**
  - **faxNumber - “Fax”**
  - **email - “E-Mail”**
  - **url - “Homepage”**
  - **description - “Produkte/Infos”**
  - **#AddressMapping_JSON -> PostalAddress**
    - **streetAddress - “Adresse”**
    - **addressLocality - “Ort”**
    - **postalCode - “PLZ”**
    - **addressRegion - “Bundesland”**
  - **#ContactMapping_JSON -> ContactPoint**
    - **name - “Kontakt”**



```
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix schema: <https://schema.org/> .
@prefix ql: <http://semweb.mmlab.be/ns/ql#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<#LOGICALSOURCE>
rml:source "firmenregister.json";
rml:referenceFormulation ql:JSONPath;
rml:iterator "$.[*]".

<#LodgingBusinessMapping>
rml:logicalSource <#LOGICALSOURCE>;

rr:subjectMap [
    rr:template "https://lodgingbusiness.example.com/{Firmenname}";
    rr:class schema:Hotel;
];

rr:predicateObjectMap [
    rr:predicate schema:name;
    rr:objectMap [
        rml:reference "Firmenname"
    ];
];
```

Figure 7: A code snippet used to map the www.firmenregister.de data.

Assessment - firmenregister.de



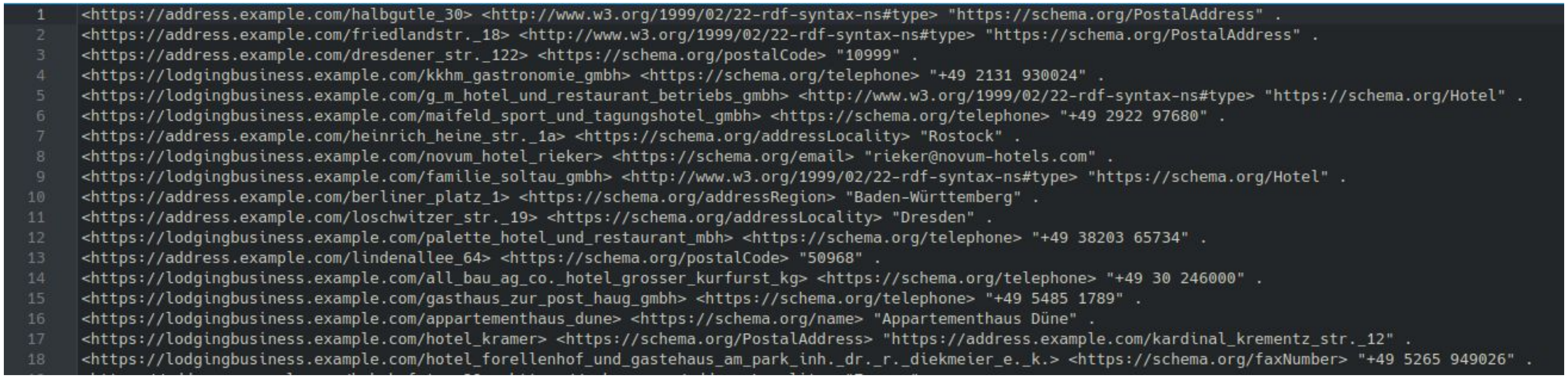
Accessibility	0.40	Provisioning of public endpoint	0.45	0 otherwise	0.00	0.23
		Retrievable format	0.45	0,5 HTML semi structured	0.50	
		Content negotiation	0.10	0 no format given	0.00	
Completeness	0.30	Instance completeness	0.50	119 out of 895 are complete	0.13	0.57
		Population completeness	0.50	more data than in GTKG	1.00	
Accuracy	0.30	Formal semantic validity	0.50	560 out of 895 are valid	0.63	0.81
		Formal syntactic validity	0.50	895 out of 895 are valid	1.00	

overall assessment score	0.50

Figure 8: The assessment results for www.firmenregister.de .

Importing - firmenregister.de

- **Problems:**
  - **data inside the .n3 is not sorted using pyRML**

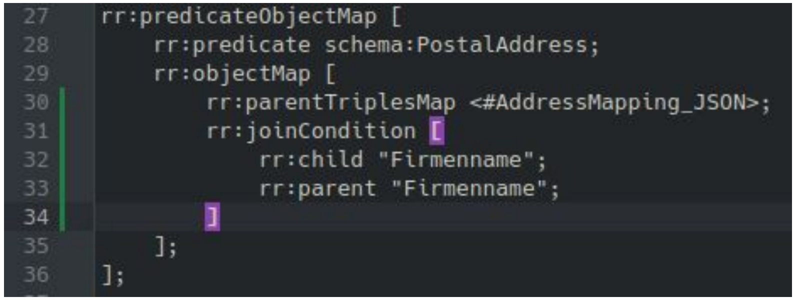


```
<https://address.example.com/halbgutle_30> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> "https://schema.org/PostalAddress" .
<https://address.example.com/friedlandstr._18> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> "https://schema.org/PostalAddress" .
<https://address.example.com/dresdener_str._122> <https://schema.org/postalCode> "10999" .
<https://lodgingbusiness.example.com/kkhm_gastronomie_gmbh> <https://schema.org/telephone> "+49 2131 930024" .
<https://lodgingbusiness.example.com/g_m_hotel_und_restaurant_betriebs_gmbh> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> "https://schema.org/Hotel" .
<https://lodgingbusiness.example.com/maifeld_sport_und_tagungshotel_gmbh> <https://schema.org/telephone> "+49 2922 97680" .
<https://address.example.com/heinrich_heine_str._1a> <https://schema.org/addressLocality> "Rostock" .
<https://lodgingbusiness.example.com/novum_hotel_rieker> <https://schema.org/email> "rieker@novum-hotels.com" .
<https://lodgingbusiness.example.com/familie_soltau_gmbh> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> "https://schema.org/Hotel" .
<https://address.example.com/berliner_platz_1> <https://schema.org/addressRegion> "Baden-Württemberg" .
<https://address.example.com/loschwitzer_str._19> <https://schema.org/addressLocality> "Dresden" .
<https://lodgingbusiness.example.com/palette_hotel_und_restaurant_mbh> <https://schema.org/telephone> "+49 38203 65734" .
<https://address.example.com/lindenallee_64> <https://schema.org/postalCode> "50968" .
<https://lodgingbusiness.example.com/all_bau_ag_co._hotel_grosser_kurfurst_kg> <https://schema.org/telephone> "+49 30 246000" .
<https://lodgingbusiness.example.com/gasthaus_zur_post_haug_gmbh> <https://schema.org/telephone> "+49 5485 1789" .
<https://lodgingbusiness.example.com/appartementhaus_dune> <https://schema.org/name> "Appartementhaus Düne" .
<https://lodgingbusiness.example.com/hotel_kramer> <https://schema.org/PostalAddress> "https://address.example.com/kardinal_krementz_str._12" .
<https://lodgingbusiness.example.com/hotel_forellenhof_und_gastehaus_am_park_inh._dr._r._diekmeier_e._k.> <https://schema.org/faxNumber> "+49 5265 949026" .
```

Figure 9: A sample result of the created N3 data.

  - **after successful import the data is still not present inside graphDB due to this sorting problem**
  - **rocketRML does export the data sorted but only if the number of properties is low and no *rr:parentTriplesMap* are used -> heap error**
    - **<u>Solution</u>: using *joinCondition***

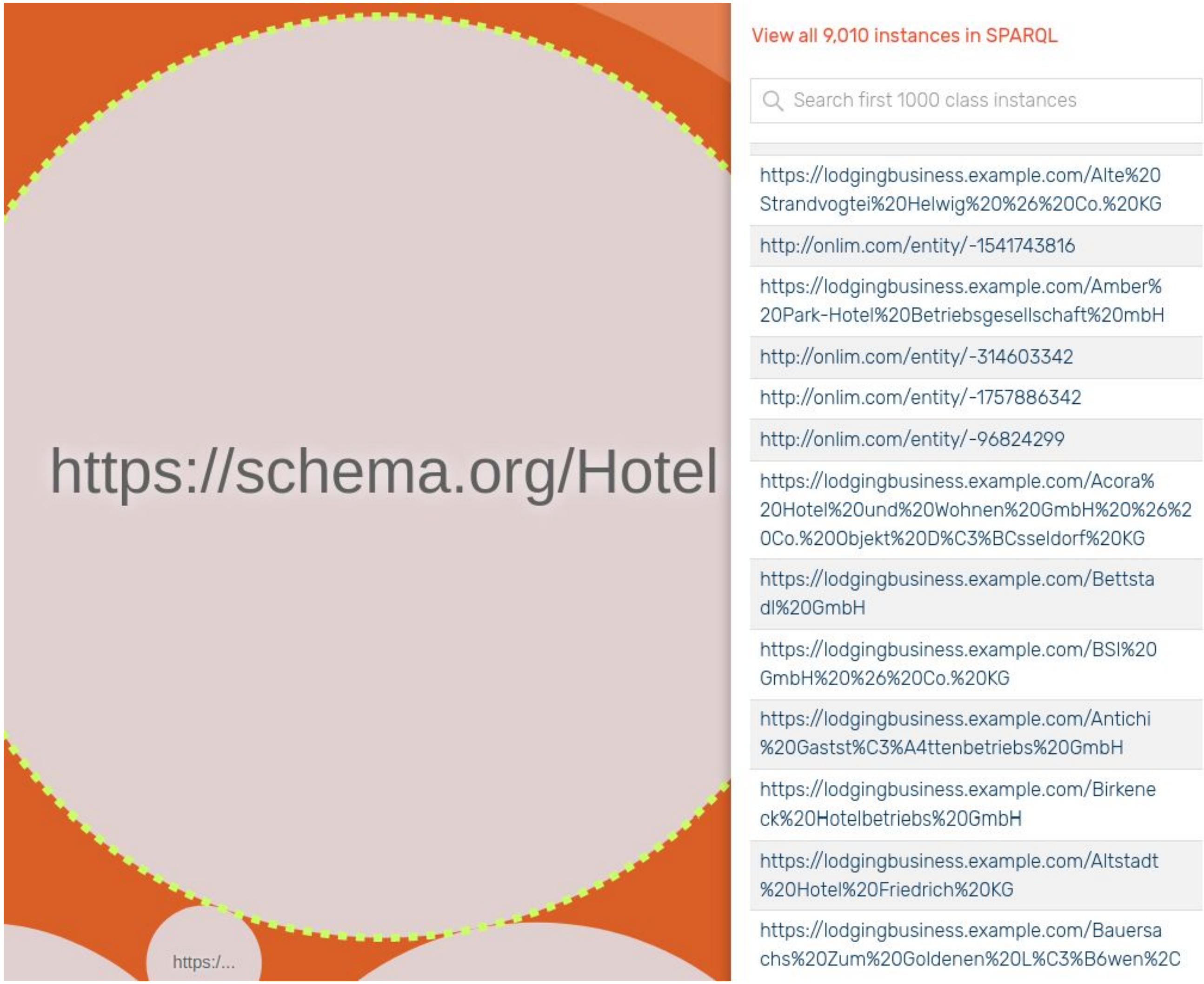
      and rocketRML



```
rr:predicateObjectMap [
    rr:predicate schema:PostalAddress;
    rr:objectMap [
        rr:parentTriplesMap <#AddressMapping_JSON>;
        rr:joinCondition [
            rr:child "Firmenname";
            rr:parent "Firmenname";
        ]
    ];
];
```

Figure 10: The joinCondition used to fix the heap error.

Importing contd. - firmenregister.de



View all 9,010 instances in SPARQL

Search first 1000 class instances

https://lodgingbusiness.example.com/Alte%20Strandvogtei%20Helwig%20%26%20Co.%20KG

http://onlim.com/entity/-1541743816

https://lodgingbusiness.example.com/Amber%20Park-Hotel%20Betriebsgesellschaft%20mbH

http://onlim.com/entity/-314603342

http://onlim.com/entity/-1757886342

http://onlim.com/entity/-96824299

https://lodgingbusiness.example.com/Acora%20Hotel%20und%20Wohnen%20GmbH%20%26%20Co.%20Objekt%20D%C3%BCsseldorf%20KG

https://lodgingbusiness.example.com/Bettstadl%20GmbH

https://lodgingbusiness.example.com/BSI%20GmbH%20%26%20Co.%20KG

https://lodgingbusiness.example.com/Antichi%20Gastst%C3%A4ttenbetriebs%20GmbH

https://lodgingbusiness.example.com/Birkeneck%20Hotelbetriebs%20GmbH

https://lodgingbusiness.example.com/Altstadt%20Hotel%20Friedrich%20KG

https://lodgingbusiness.example.com/Bauersachs%20Zum%20Goldenen%20L%C3%B6wen%2C

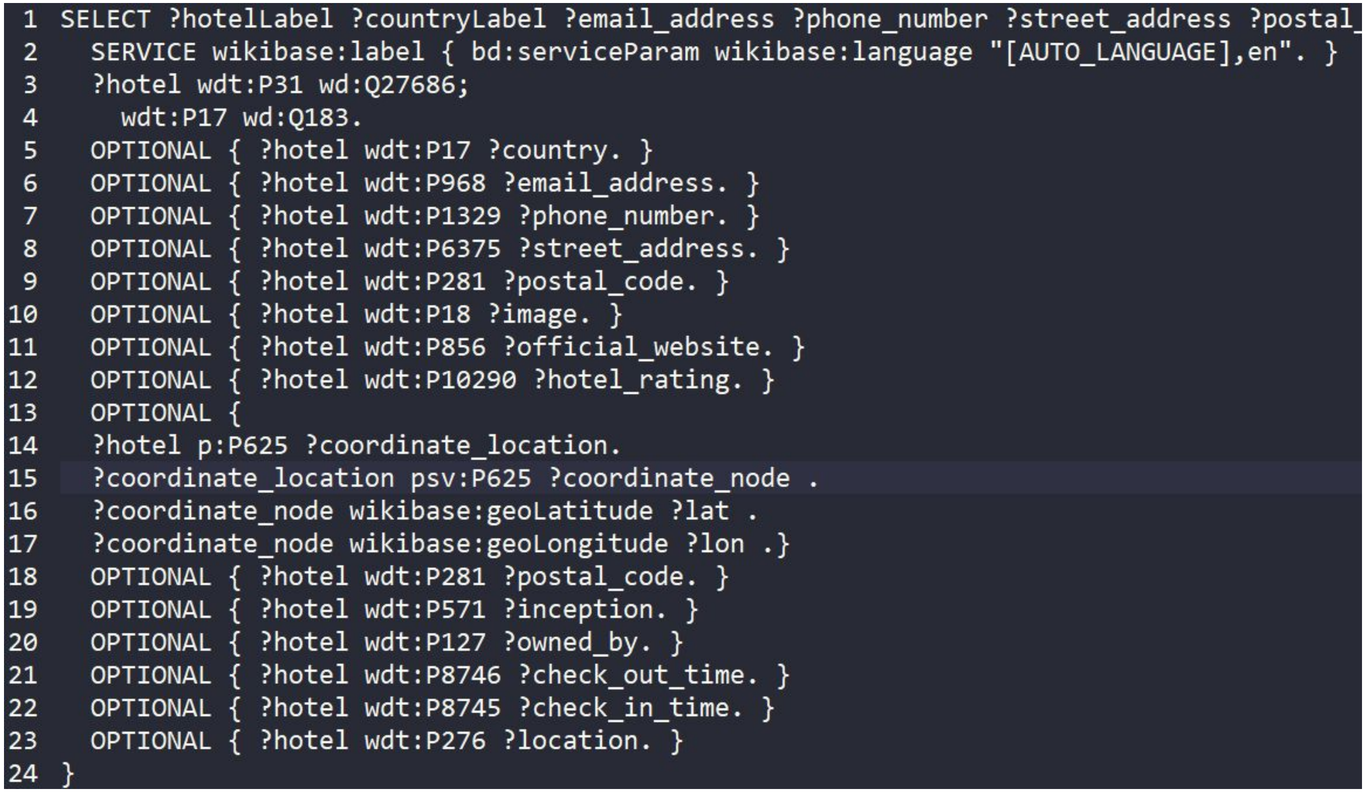
Figure 11: A view of the newly created data inserted into the GTKG.

Data source: www.wikidata.org

SPARQL query, schema alignment, mapping, and assessment.

SPARQL - Query

- **Focus on mandatory data first**

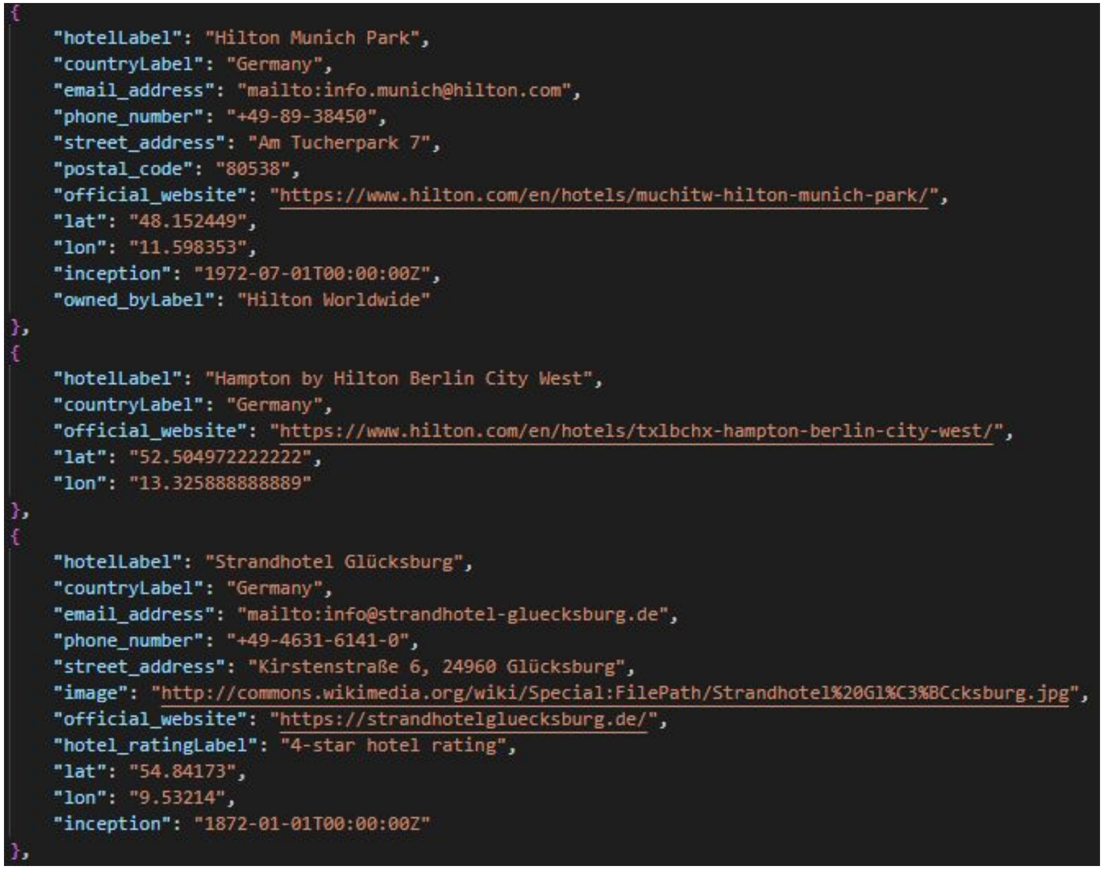


```
SELECT ?hotelLabel ?countryLabel ?email_address ?phone_number ?street_address ?postal_
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
  ?hotel wdt:P31 wd:Q27686;
    wdt:P17 wd:Q183.
  OPTIONAL { ?hotel wdt:P17 ?country. }
  OPTIONAL { ?hotel wdt:P968 ?email_address. }
  OPTIONAL { ?hotel wdt:P1329 ?phone_number. }
  OPTIONAL { ?hotel wdt:P6375 ?street_address. }
  OPTIONAL { ?hotel wdt:P281 ?postal_code. }
  OPTIONAL { ?hotel wdt:P18 ?image. }
  OPTIONAL { ?hotel wdt:P856 ?official_website. }
  OPTIONAL { ?hotel wdt:P10290 ?hotel_rating. }
  OPTIONAL {
  ?hotel p:P625 ?coordinate_location.
  ?coordinate_location psv:P625 ?coordinate_node .
  ?coordinate_node wikibase:geoLatitude ?lat .
  ?coordinate_node wikibase:geoLongitude ?lon .}
  OPTIONAL { ?hotel wdt:P281 ?postal_code. }
  OPTIONAL { ?hotel wdt:P571 ?inception. }
  OPTIONAL { ?hotel wdt:P127 ?owned_by. }
  OPTIONAL { ?hotel wdt:P8746 ?check_out_time. }
  OPTIONAL { ?hotel wdt:P8745 ?check_in_time. }
  OPTIONAL { ?hotel wdt:P276 ?location. }
}
```

Figure 12: The SPARQL query used on www.wikidata.org .

SPARQL - Query Result

- **Focus on mandatory data first**

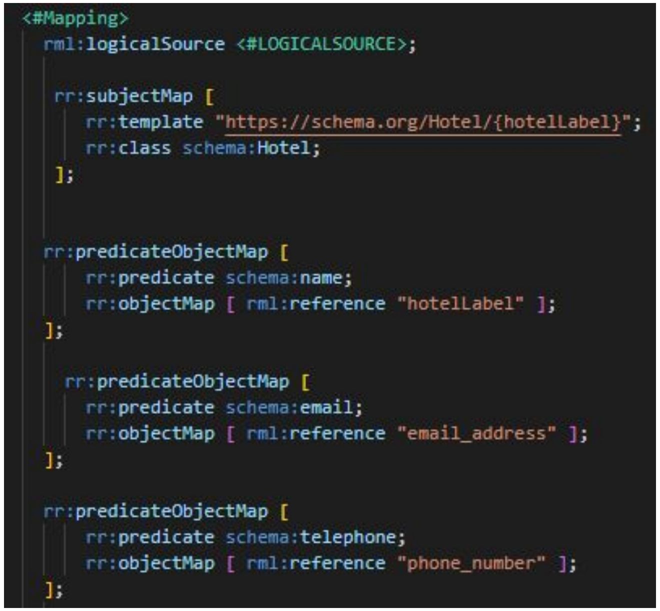


```
{
    "hotelLabel": "Hilton Munich Park",
    "countryLabel": "Germany",
    "email_address": "mailto:info.munich@hilton.com",
    "phone_number": "+49-89-38450",
    "street_address": "Am Tucherpark 7",
    "postal_code": "80538",
    "official_website": "https://www.hilton.com/en/hotels/muchitw-hilton-munich-park/",
    "lat": "48.152449",
    "lon": "11.598353",
    "inception": "1972-07-01T00:00:00Z",
    "owned_byLabel": "Hilton Worldwide"
},
{
    "hotelLabel": "Hampton by Hilton Berlin City West",
    "countryLabel": "Germany",
    "official_website": "https://www.hilton.com/en/hotels/txlbchx-hampton-berlin-city-west/",
    "lat": "52.504972222222",
    "lon": "13.325888888889"
},
{
    "hotelLabel": "Strandhotel Glücksburg",
    "countryLabel": "Germany",
    "email_address": "mailto:info@strandhotel-gluecksburg.de",
    "phone_number": "+49-4631-6141-0",
    "street_address": "Kirstenstraße 6, 24960 Glücksburg",
    "image": "http://commons.wikimedia.org/wiki/Special:FilePath/Strandhotel%20Gl%C3%BCcksburg.jpg",
    "official_website": "https://strandhotelgluecksburg.de/",
    "hotel_ratingLabel": "4-star hotel rating",
    "lat": "54.84173",
    "lon": "9.53214",
    "inception": "1872-01-01T00:00:00Z"
},
```

Figure 13: The JSON data received after using the SPARQL query on www.wikidata.org .

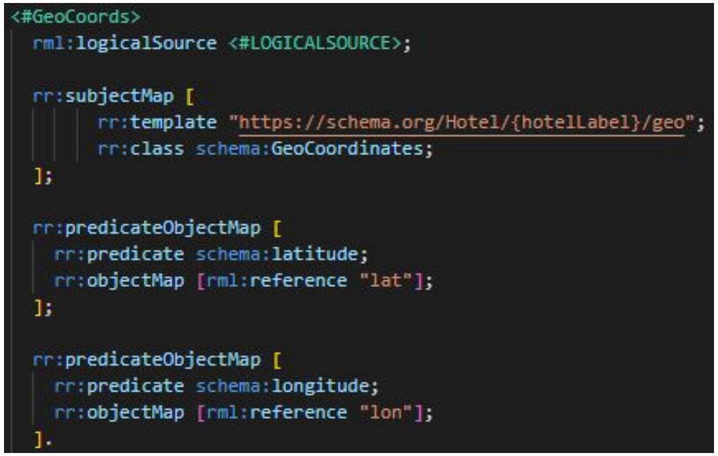
Mapping - wikidata.org

- **Properties:**
  - **name - “hotelLabel”**
  - **telephone - “phone_number”**
  - **faxNumber - “fax_number”**
  - **email - “email_address”**
  - **url - “official_website”**
  - **foundingDate- “inception”**
  - **checkinTime- “check_in_timeLabel”**
  - **checkoutTime- “check_out_timeLabel”**
  - **#Address -> PostalAddress**
  - **#GeoCoords -> GeoCoordinates**
  - **#Image -> ImageObject**
  - **#Owner -> Person**



```
<#Mapping>
  rml:logicalSource <#LOGICALSOURCE>;

   rr:subjectMap [
      rr:template "https://schema.org/Hotel/{hotelLabel}";
      rr:class schema:Hotel;
   ];


  rr:predicateObjectMap [
      rr:predicate schema:name;
      rr:objectMap [ rml:reference "hotelLabel" ];
  ];

    rr:predicateObjectMap [
      rr:predicate schema:email;
      rr:objectMap [ rml:reference "email_address" ];
  ];

  rr:predicateObjectMap [
      rr:predicate schema:telephone;
      rr:objectMap [ rml:reference "phone_number" ];
  ];
```

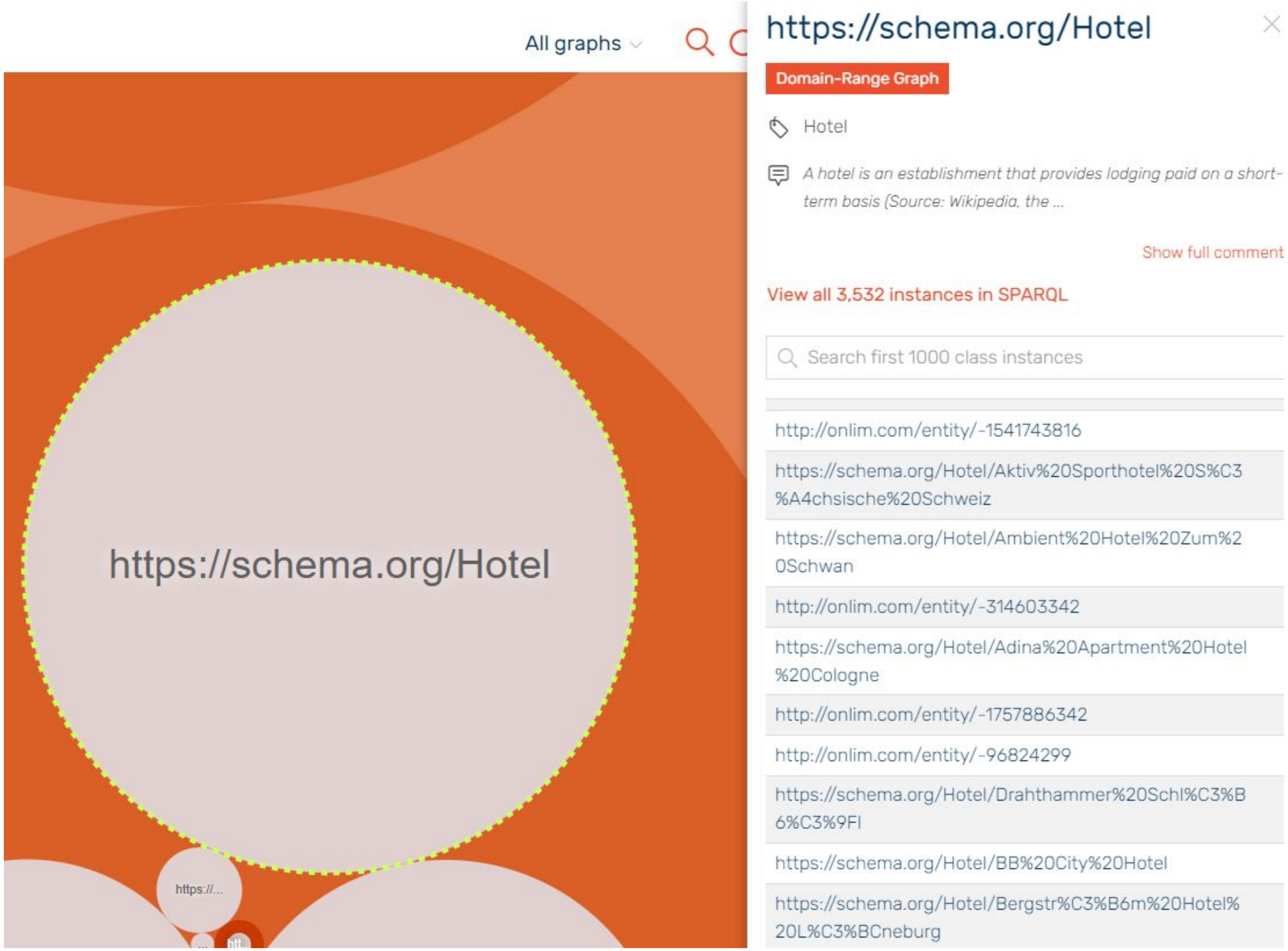
Figure 14: A code snippet used for mapping a hotel.



```
<#GeoCoords>
  rml:logicalSource <#LOGICALSOURCE>;

  rr:subjectMap [
        rr:template "https://schema.org/Hotel/{hotelLabel}/geo";
        rr:class schema:GeoCoordinates;
  ];

  rr:predicateObjectMap [
    rr:predicate schema:latitude;
    rr:objectMap [rml:reference "lat"];
  ];

  rr:predicateObjectMap [
    rr:predicate schema:longitude;
    rr:objectMap [rml:reference "lon"];
  ].
```

Figure 15: A code snippet used for mapping GeoCoordinates.

Importing contd. - wikidata.org



https://schema.org/Hotel

Domain-Range Graph

Hotel

A hotel is an establishment that provides lodging paid on a short-term basis (Source: Wikipedia, the ...

Show full comment

View all 3,532 instances in SPARQL

Search first 1000 class instances

http://onlim.com/entity/-1541743816

https://schema.org/Hotel/Aktiv%20Sporthotel%20S%C3%A4chsische%20Schweiz

https://schema.org/Hotel/Ambient%20Hotel%20Zum%20Schwan

http://onlim.com/entity/-314603342

https://schema.org/Hotel/Adina%20Apartment%20Hotel%20Cologne

http://onlim.com/entity/-1757886342

http://onlim.com/entity/-96824299

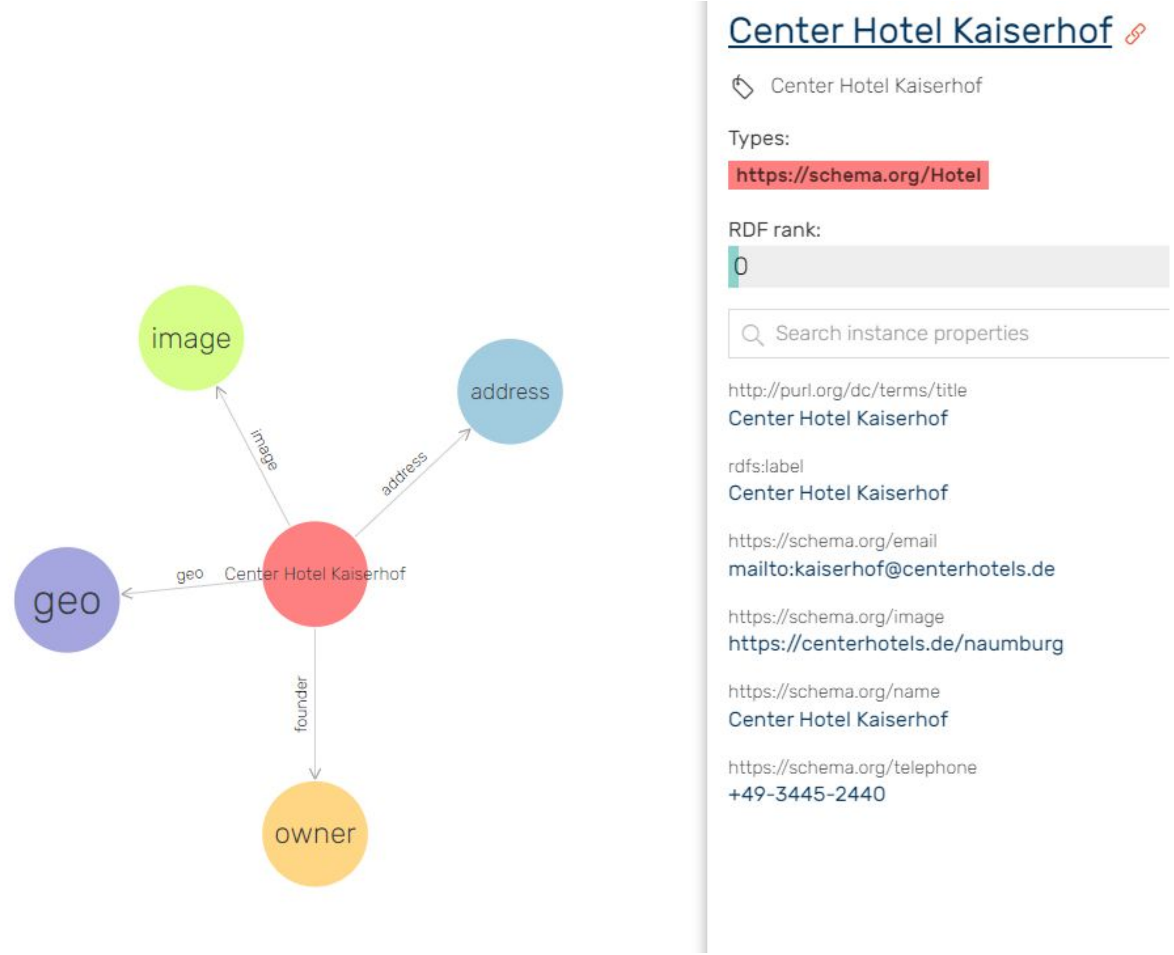
https://schema.org/Hotel/Drahthammer%20Schl%C3%B6%C3%9Fl

https://schema.org/Hotel/BB%20City%20Hotel

https://schema.org/Hotel/Bergstr%C3%B6m%20Hotel%20L%C3%BCneburg

Figure 16: A view of the newly created data inserted into the GTKG.

Importing contd. - wikidata.org



Center Hotel Kaiserhof

Center Hotel Kaiserhof

Types:

https://schema.org/Hotel

RDF rank:

0

Search instance properties

http://purl.org/dc/terms/title
Center Hotel Kaiserhof

rdfs:label
Center Hotel Kaiserhof

https://schema.org/email
mailto:kaiserhof@centerhotels.de

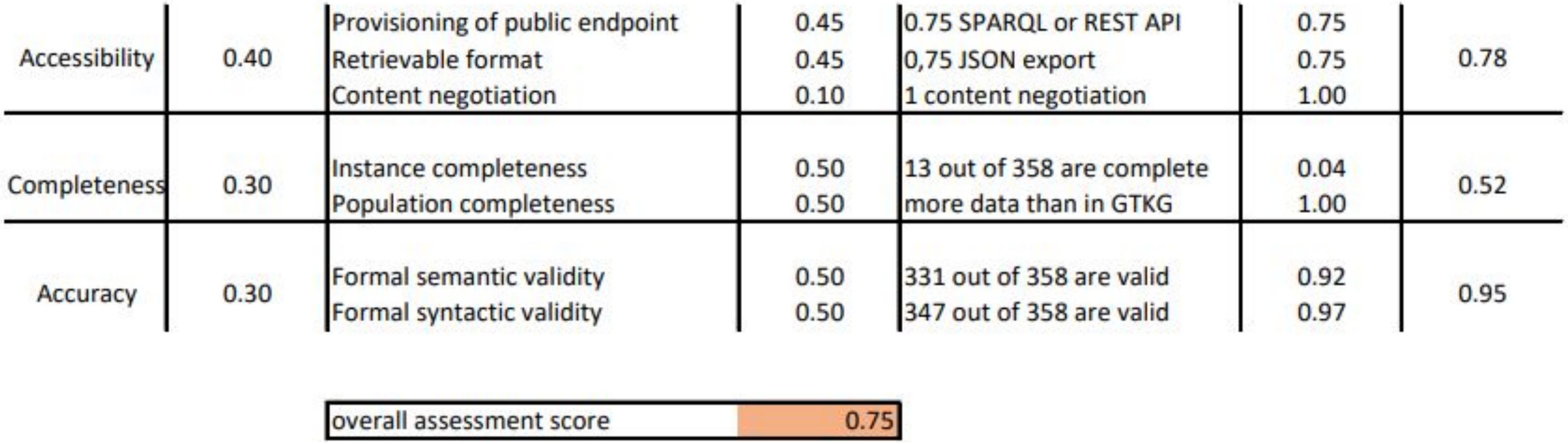
https://schema.org/image
https://centerhotels.de/naumburg

https://schema.org/name
Center Hotel Kaiserhof

https://schema.org/telephone
+49-3445-2440

Figure 17: A detailed view of a newly inserted hotel instance.

Assessment - wikidata.org



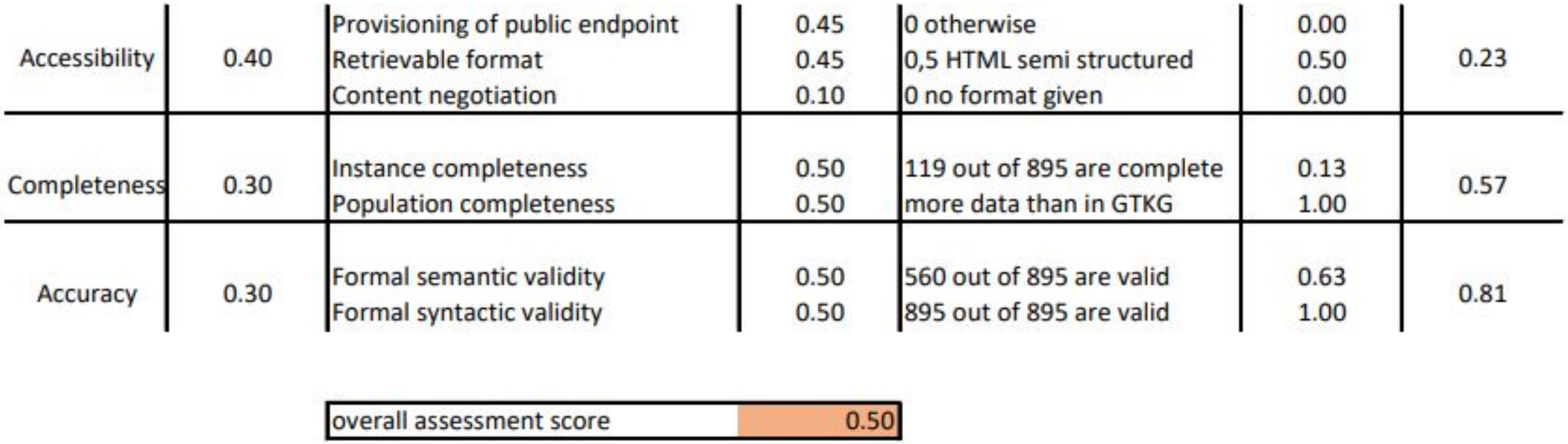
Accessibility	0.40	Provisioning of public endpoint	0.45	0.75 SPARQL or REST API	0.75	0.78
		Retrievable format	0.45	0,75 JSON export	0.75	
		Content negotiation	0.10	1 content negotiation	1.00	
Completeness	0.30	Instance completeness	0.50	13 out of 358 are complete	0.04	0.52
		Population completeness	0.50	more data than in GTKG	1.00	
Accuracy	0.30	Formal semantic validity	0.50	331 out of 358 are valid	0.92	0.95
		Formal syntactic validity	0.50	347 out of 358 are valid	0.97	

overall assessment score	0.75

Figure 18: The assessment results for www.wikidata.org .

Assessment - Comparison

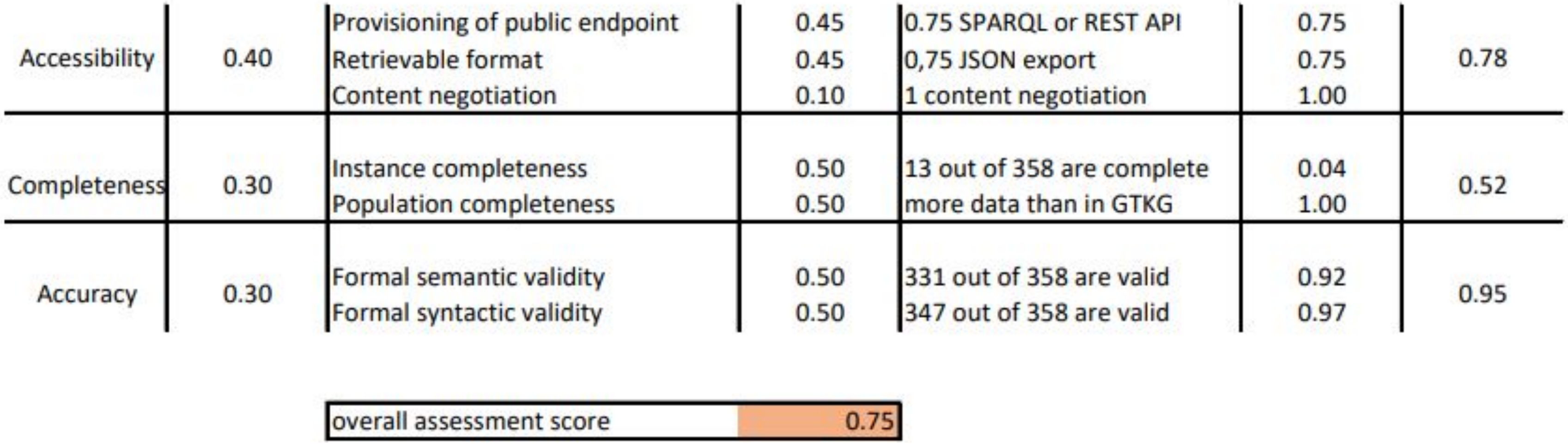
Data source: www.firmenregister.de



Accessibility	0.40	Provisioning of public endpoint	0.45	0 otherwise	0.00	0.23
		Retrievable format	0.45	0,5 HTML semi structured	0.50	
		Content negotiation	0.10	0 no format given	0.00	
Completeness	0.30	Instance completeness	0.50	119 out of 895 are complete	0.13	0.57
		Population completeness	0.50	more data than in GTKG	1.00	
Accuracy	0.30	Formal semantic validity	0.50	560 out of 895 are valid	0.63	0.81
		Formal syntactic validity	0.50	895 out of 895 are valid	1.00	

overall assessment score	0.50

Data source: www.wikidata.org



Accessibility	0.40	Provisioning of public endpoint	0.45	0.75 SPARQL or REST API	0.75	0.78
		Retrievable format	0.45	0,75 JSON export	0.75	
		Content negotiation	0.10	1 content negotiation	1.00	
Completeness	0.30	Instance completeness	0.50	13 out of 358 are complete	0.04	0.52
		Population completeness	0.50	more data than in GTKG	1.00	
Accuracy	0.30	Formal semantic validity	0.50	331 out of 358 are valid	0.92	0.95
		Formal syntactic validity	0.50	347 out of 358 are valid	0.97	

overall assessment score	0.75

Figure 19: The assessment results for both data sources (www.firmenregister.de and www.wikidata.org) .

Thank You!

We are open for feedback and questions.