SDM Project 2 - Knowledge Graphs

Shofiyyah Nadhiroh, Hieu Nguyen Minh (BDMA)

June 8, 2024

B.1. TBOX Definition

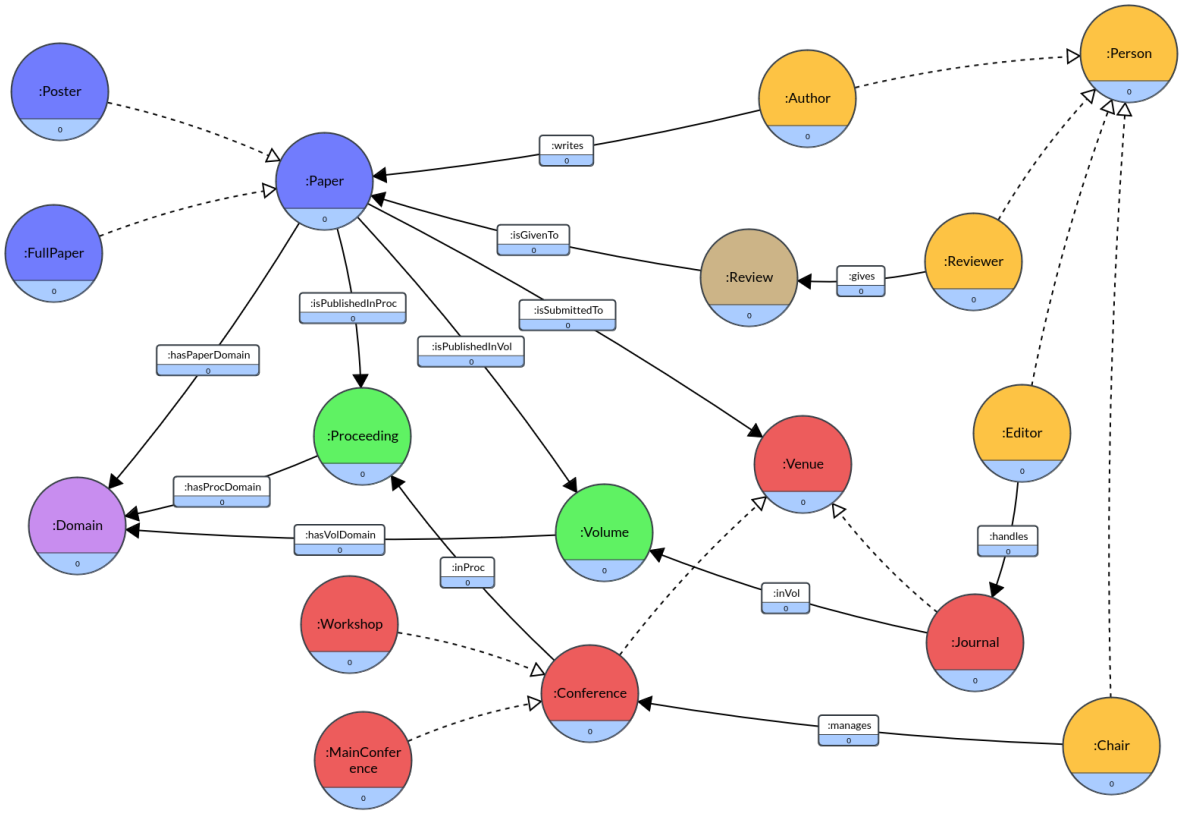

Figure 1: Graph for creating TBOX.

We model our graph as shown Figure 1. We make the following assumptions for this graph:

- A `Paper` can be a `FullPaper` or a `Poster`.
- There are two types of `Venue`: `Conference` and `Journal`. A `Conference` can be a `Workshop` or a `MaincCnference`.
- Papers in a conference and a journal are published in a `Proceeding` and a `Volume`, respectively. Every `Paper`, `Proceeding` and `Volume` have a specific `Domain`.
- A `Person` can be an `Author` who writes `Papers`, a `Reviewer` who gives `Reviews` to `Papers`, an Editor who handles `Journals`, or a `Chair` who manages `Conferences`.

Below is a code excerpt in Python for creating classes with RDFLib[1]. The class hierarchy is shown in Figure 2.

[1]https://rdflib.readthedocs.io/en/stable/

```python
from rdflib import Graph, Namespace, Literal
from rdflib.namespace import RDFS, RDF, XSD

graph = Graph()
lab2 = Namespace("http://sdmlab2.org/")

# Paper Superclass
graph.add((lab2.Paper, RDF.type, RDFS.Class))
graph.add((lab2.Paper, RDFS.label, Literal("Paper")))
```

Listing 1: TBOX creation.

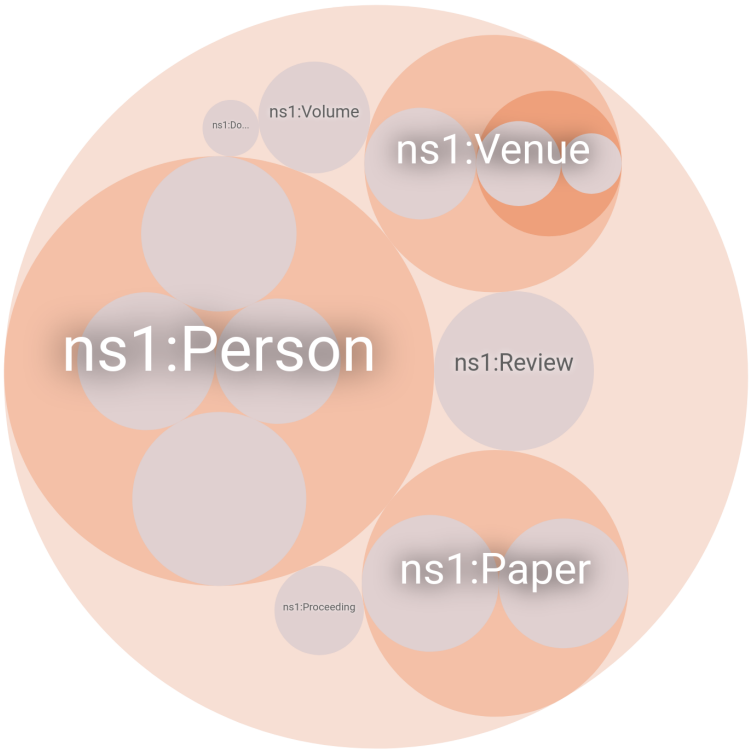

Figure 2: Class hierarchy.

B.2. ABOX Definition

For the dataset, we crawl the data about academic papers from Semantic Scholar[2]. We make the following assumptions during the creation of ABox:

- Conferences and journals are distinguished by `conferenceJournalId`, which begins with `c` for conferences and `j` for journals.
- Decisions given by reviewers for submitted papers are randomly generated as Boolean values with a probability of 0.8 for accepted papers.

[2]https://www.semanticscholar.org/

We use the `graph.add()` function to create concrete instances for class attributes and relationships. The instance values are taken from DataFrames loaded from `csv` files in the dataset. Below is an excerpt of a function creating `paperTitle` ABox from the DataFrame `papers_df`.

```
def paperTitle_ABox():
    graph.add((lab2.Paper, lab2.paperTitle, XSD.string))
    for k in range(len(papers_df['paperId'])):
        graph.add((URIRef(lab2+papers_df['paperId'][k]), lab2.paperTitle, Literal(
    papers_df['paperTitle'][k])))
```

Listing 2: ABOX creation.

B.2. Create the Final Ontology

We use the `graph.add()` function to connect ABox and TBox by specifying that an instance has `RDF.type` of a created class. Below is an excerpt of a function connecting relationship `hasPaperDomain` to class `Domain`.

```
def connect_hasPaperDomain():
    for k in range(len(domainsPapers_df['paperId'])):
        graph.add((URIRef(domainsPapers_df['domainId'][k]), RDF.type, lab2.Domain)
    )
```

Listing 3: ABOX creation.

Then we import our RDF files (in `ttl` format) to GraphDB under Base IRI `http://sdmlab2.org/` and Target graph `http://localhost:7200/sdmlab2/`, resulting in a repository in Figure 3.

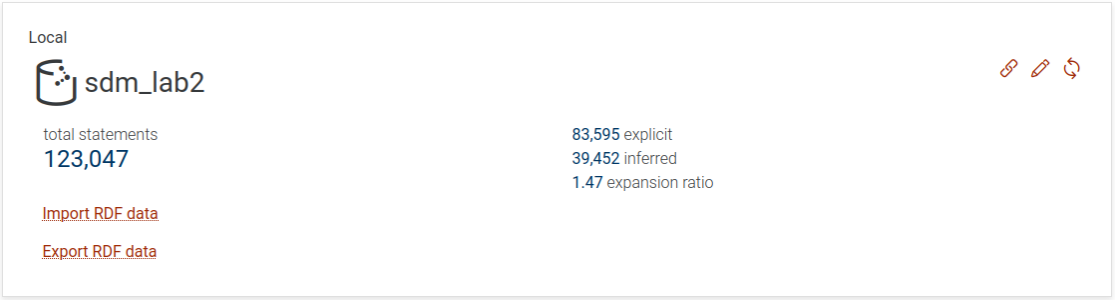

Figure 3: Overview of GraphDB local repository.

In total, we have 17 classes, 31 properties, 28,116 instances, and 83,595 triples. Table 1 shows statistics about number of instances by class.

Table 1: Number of instances by class.

Paper	Poster	FullPaper	Person	Author	Reviewer	Review	Editor	Chair
2500	1468	1016	10684	10684	3966	5000	798	1552

Domain	Venue	Conference	MainConference	Workshop	Journal	Proceeding	Volume
19	499	115	91	24	384	115	384

B.4. Querying the ontology

1. Find all authors

We search for instances whose `rdf:type` is `Author`. Part of the result is shown in Figure 4.

```
PREFIX lab2: <http://sdmlab2.org/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?authorName
WHERE
{
    ?author rdf:type lab2:Author .
    ?author lab2:authorName ?authorName .
}
```

Listing 4: Query 1.

	authorName
1	"E. Bolyen"
2	"J. Rideout"
3	"Yang Bai"
4	"Yunhu Wan"
5	"Min Li"
6	"Emily Stull"
7	"Jennifer Williams"
8	"A. McCormack"
9	"D. Schieltz"
10	"V. Wakelam"

Figure 4: Result of query 1.

2. Find all properties whose domain is Author

We search for instances whose `rdfs:domain` is `Author`. The result is shown in Figure 5.

```
PREFIX lab2: <http://sdmlab2.org/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?propertyName
WHERE
{
    ?propertyName rdfs:domain lab2:Author .
}
```

Listing 5: Query 2.

	propertyName
1	ns1:authorName
2	ns1:writes

Figure 5: Result of query 2.

3. Find all properties whose domain is either Conference or Journal

We search for instances whose `rdfs:domain` is either `Conference` or `Journal` and merge them together. The result is shown in Figure 6.

```
PREFIX lab2: <http://sdmlab2.org/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?propertyName
WHERE
{
    {?propertyName rdfs:domain lab2:Conference}
    UNION
    {?propertyName rdfs:domain lab2:Journal}
}
```

Listing 6: Query 3.

	propertyName
1	ns1:confTitle
2	ns1:inProc
3	ns1:journalTitle
4	ns1:inVol

Figure 6: Result of query 3.

4. Find all the papers written by a given author that where published in database conferences

We choose an author "J. Tate" and find all papers written by him. From that, we find papers which are published in proceedings whose domain is "Database". The result is shown in Figure 7.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lab2: <http://sdmlab2.org/>
SELECT (?pTitle as ?paper_title) (?pName as ?proc_name)
WHERE
{
    ?paper      rdf:type                lab2:Paper ;
                lab2:paperTitle         ?pTitle ;
                lab2:isPublishedInProc  ?proceeding .

    ?author     rdf:type                lab2:Author ;
                lab2:authorName         "J. Tate" ;
                lab2:writes             ?paper .

    ?proceeding lab2:procName           ?pName ;
                lab2:hasProcDomain      ?domain.

    ?domain     rdf:type                lab2:Domain ;
                lab2:domainName         "Database" .
}
```

Listing 7: Query 4.

	paper_title	proc_name
1	"The Pfam protein families database: towards a more sustainable future"	"proceeding16"
2	"Pfam: the protein families database"	"proceeding92"
3	"The Pfam protein families database"	"proceeding28"
4	"Rfam 12.0: updates to the RNA families database"	"proceeding110"
5	"InterPro in 2011: new developments in the family and domain prediction database"	"proceeding22"
6	"Rfam: updates to the RNA families database"	"proceeding34"

Figure 7: Result of query 4.

Additional queries

5. Find all editors who handle journals whose domain is "Engineering"

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lab2: <http://sdmlab2.org/>
SELECT (?eName as ?reviewer_name) (?jTitle as ?journal_title)
WHERE
{
    ?editor   rdf:type            lab2:Editor ;
              lab2:authorName     ?eName ;
              lab2:handles        ?journal.

    ?journal  rdf:type            lab2:Journal ;
              lab2:journalTitle   ?jTitle ;
              lab2:inVol          ?volume .

    ?volume   rdf:type            lab2:Volume ;
              lab2:hasVolDomain   ?domain.

    ?domain   rdf:type            lab2:Domain ;
              lab2:domainName     "Engineering" .
}
```

Listing 8: Query 5.

	editor_name	journal_title
1	"S. Herrmann"	"Nature Biotechnology"
2	"A. Allocca"	"Nature Biotechnology"
3	"Fangfang Xia"	"Nature Biotechnology"
4	"A. Al-Chalabi"	"Nature Biotechnology"
5	"D. Mishmar"	"Nature Biotechnology"
6	"B. Yao"	"Nature Biotechnology"
7	"C. Raetz"	"BMC Bioinformatics"
8	"N. Gregor"	"BMC Bioinformatics"
9	"K. Kumaran"	"BMC Bioinformatics"
10	"R. Schwacke"	"BMC Bioinformatics"

Figure 8: Part of result of query 5.

6. Find all reviewers who give reviews to papers whose domain is "Business"

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lab2: <http://sdmlab2.org/>
SELECT (?eName as ?editor_name) (?jTitle as ?journal_title)
WHERE
{
    ?editor     rdf:type        lab2:Editor ;
                lab2:authorName ?eName ;
                lab2:handles  ?journal.

    ?journal    rdf:type        lab2:Journal ;
                lab2:journalTitle ?jTitle ;
                lab2:inVol   ?volume .

    ?volume     rdf:type        lab2:Volume ;
                lab2:hasVolDomain   ?domain.

    ?domain     rdf:type lab2:Domain ;
                lab2:domainName "Engineering" .
}
```

Listing 9: Query 6.

	reviewer_name	paper_title
1	"Y. Matsubara"	"Role of bioinformatics and pharmacogenomics in drug discovery and development process"
2	"Jinna Choi"	"Big Data Science: Opportunities and Challenges to Address Minority Health and Health Disparities in the 21st Century."
3	"Charles Larson"	"Data Science, Predictive Analytics, and Big Data: A Revolution that Will Transform Supply Chain Design and Management"
4	"Benjamin Müller"	"Assessing the vulnerability of supply chains using graph theory"
5	"C. Gill"	"Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential"
6	"Xiaofang Wu"	"A New Database on Financial Development and Structure"
7	"A. Fuchs"	"Systemic Banking Crises: A New Database"
8	"S. Staehli"	"Data science ethics in government"
9	"H. Le"	"Measuring Financial Inclusion: The Global Findex Database"
10	"John Millar Carroll"	"Systemic Banking Crises Database: An Update"

Figure 9: Part of result of query 6.